

Bayesian methods for source attribution using HIV deep sequence data

Xiaoyue Xi

Department of Mathematics
Imperial College London

This dissertation is submitted for the degree of
Doctor of Philosophy

May 2022

I would like to dedicate this thesis to my loving parents and grandparents.

Declaration

I certify that this thesis, and the research to which it refers, are the product of my own work, and that any ideas or quotations from the work of other people, published or otherwise, are fully acknowledged in accordance with the standard referencing practices of the discipline.

- Section 1.1 - 1.2, Section 2, Figure 2.1, Section 5.1, 5.3, 5.4, Section 6.2 - 6.3 and Appendix A were adapted or copied from the paper entitled "Xi, X., Spencer, S., Hall, M., Grabowski, M. K., Kagaayi, J. , Ratmann, O., 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".
- Chapter 3 has been integrated with `phyloscanner`, and an automatic tool for implementing the bioinformatics pipeline is under development.
- Chapter 4 will be prepared for publication.
- Chapter 8 was adapted from the paper entitled "Monod, M., Blenkinsop, A., Xi, X., Hebert, D., Bershan, S., Tietze, S., et al., 2021. Age groups that sustain resurging COVID-19 epidemics in the United States. Science, 371(6536), p.eabe8372".

Xiaoyue Xi

May 2022

Copyright

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a Creative Commons Attribution-Non Commercial-No Derivatives 4.0 International Licence (CC BY-NC-ND).

Under this licence, you may copy and redistribute the material in any medium or format on the condition that; you credit the author, do not use it for commercial purposes and do not distribute modified versions of the work.

When reusing or sharing this work, ensure you make the licence terms clear to others by naming the licence and linking to the licence text.

Please seek permission from the copyright holder for uses of this work that are not included in this licence or permitted under UK Copyright Law.

Acknowledgements

First I would like to express my deepest appreciation to my supervisor, Dr. Oliver Ratmann, for his continuous academic and emotional support throughout the years. I am deeply grateful for his expertise, patience, guidance, and encouragement. The thesis would not be possible without his wisdom and support. I am also grateful to my second supervisor Prof. Nick Heard and all my progression, internal and external examiners for their time, efforts, and valuable suggestions for improving this work.

I would like to thank my colleagues at the Statistic Section, Imperial College London for providing a friendly and stimulating research environment, and to the members in the Ratmann group, for their help, comments and collaboration. My big thanks also go to all the members of PANGEA-HIV consortium for training and discussions. I appreciate the participants of the Rakai Community Cohort Study and staff and investigators who made this study possible. I gratefully acknowledge the funding support for my study.

I would like to give my special thanks to my parents and families for their love, and Yue for her understanding and support during my ups and downs and helping me with proofreading.

Abstract

The advent of pathogen deep-sequencing technology provides new opportunities for infectious disease surveillance, especially for fast-evolving viruses like human immunodeficiency virus (HIV). In particular, multiple reads per host contain detailed information on viral within-host diversity. This information allows the reconstruction of partial directed transmission networks, where estimates of who is source and who is recipient are directly available from the phylogenetic ordering of the viruses of any two individuals. This is a new approach for phylodynamics, and the topic of my thesis.

In this thesis, I present updates to the bioinformatics pipeline used by the Phylogenetics And Networks for Generalised Epidemics in Africa consortium for processing HIV deep sequence data and running the phyloscanner program. I then present a semi-parametric Bayesian Poisson model 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 a finer resolution than previously possible. In this sense, the methods that I developed enable us to overcome some problems which have been unable to be solved by conventional phylodynamic approaches.

We apply the approach to densely sampled, population-based HIV deep-sequence data from Rakai, Uganda. I focus on characterising age-specific transmission dynamics, and examining the sources of HIV infections in adolescent and young women in particular.

Contents

List of Figures	xi
List of Tables	xxix
1 Introduction	3
1.1 The global HIV epidemic	3
1.2 Using HIV sequence data to characterise transmission flows	5
1.3 Aim, contributions and structure	7
1.4 Other projects	9
2 Data	10
3 Bioinformatics pipeline for processing HIV deep sequence data	13
3.1 Introduction	13
3.2 Potential transmission networks	16
3.2.1 Similarity scores between consensus sequences across windows . .	16
3.2.2 Training data for classifying pairs of individuals into close and not close	18
3.2.3 Classifier: Gaussian mixture model on log similarity scores	20
3.2.4 Applying the classifier to similarity scores from all possible pairs of individuals	24
3.2.5 Considering classifications for all windows	26
3.2.6 Building potential transmission networks	29
3.2.7 Analysing large potential transmission networks	29
3.2.8 Adding controls to potential transmission networks	33
3.3 Deep sequence alignments and phylogeny reconstruction	33
3.3.1 Background sequences	33
3.3.2 Alignment of background sequences	34

3.3.3	Selection of background sequences	35
3.3.4	Generating deep-sequence alignments	36
3.3.5	Reconstructing deep-sequence phylogenies	39
3.4	Identifying source-recipient pairs	43
3.4.1	Rules for deriving pairwise phylogenetic relationships between individuals in deep-sequence phylogenetic trees	43
3.4.2	Down-sampling of reads in case of read imbalance	46
3.4.3	Handling ancestral subgraphs with no branch length differences	47
3.4.4	Identified transmission networks and source-recipient pairs	50
3.5	Summary	52
4	The non-parametric Gaussian process flow model	53
4.1	Introduction	53
4.2	The basic Gaussian process flow model	55
4.3	Numerical inference of the Gaussian process flow model	60
4.3.1	Identifiability	60
4.3.2	Reparameterisation	61
4.3.3	Prediction	61
4.3.4	Experiments	62
4.3.5	Numerical performance on Rakai data	63
4.4	Model performance measures	64
4.5	Preliminary results	66
4.6	Summary	69
5	The basic Bayesian Poisson flow model	73
5.1	Introduction	73
5.2	Approaches for quantifying and controlling for sampling bias	74
5.3	Development of the Bayesian Poisson flow model	77
5.3.1	Model framework	79
5.3.2	Numerical inference with Stan	82
5.3.3	Numerical inference with a custom Markov Chain Monte Carlo algorithm	83
5.4	Results	85
5.4.1	ODE based simulations	85
5.4.2	Sensitivity analysis of the overall sample size of observed transmissions	89
5.4.3	More complex simulation experiments	91

5.4.4	Inference method comparison	93
5.4.5	Modelling the sampling cascade of transmission events	94
5.4.6	Numerical performance on Rakai data	96
5.4.7	Application to estimate transmission flows between areas with high and low disease prevalence in Rakai	96
5.5	Comparison with flow inference from consensus sequences	98
5.5.1	Mugration model	99
5.5.2	Birth-Death Model	103
5.5.3	Discussion	106
5.6	Summary	107
6	The semi-parametric Poisson flow model	109
6.1	Introduction	109
6.2	Development of the semi-parametric Poisson model	110
6.2.1	Model framework	110
6.2.2	Numerical inference using 2D Hilbert Space Gaussian Process Ap- proximations in Stan	111
6.3	Results	114
6.3.1	Simulations	114
6.3.2	Numerical performance on Rakai data	119
6.3.3	Transmission flows between age groups defined by 1-year age bands	119
6.3.4	Comparison to flows estimated using broader age strata	120
6.3.5	Sources of transmission to young and adolescent women	123
6.3.6	Role of age discordance in transmissions	129
6.3.7	Gender differences in HIV transmission flow	129
6.3.8	Impact of adjustments for sampling heterogeneity	130
6.4	Summary	134
7	Statistical methods for comparing flow models	135
7.1	Introduction	135
7.2	Summary of four methods for numerically approximating the marginal like- lihood	136
7.2.1	Four methods to estimate the marginal likelihood	136
7.2.2	Uncertainty quantification	140
7.2.3	Wrap-III bridge sampling	143
7.3	Implementation of the bridge sampling algorithm	146

7.4	Comparison of models that describe age-specific transmission flows	148
7.5	Summary	152
8	COVID-19 flow model	154
8.1	Introduction	154
8.2	Development of a Bayesian age-specific renewal model	155
8.2.1	Model framework	155
8.2.2	Data source to inform the prior	156
8.2.3	Prior specification	158
8.2.4	Generated quantities under the flow model	160
8.3	Results	162
8.4	Discussion	163
8.5	Summary	167
9	Conclusion	168
	Bibliography	171
	Appendix A Modelling and estimation of the sampling cascade	186
A.1	Modelling and estimating participation probabilities of source cases	187
A.2	Modelling and estimating sequencing probabilities of source cases	189
A.3	Modelling and estimating participation probabilities of recipient cases . . .	192
A.4	Modelling and estimating sequencing probabilities of recipient cases	193
A.5	Pairwise sampling probabilities of sources and recipients	198
	Appendix B Appendix for Chapter 4	203
	Appendix C Appendix for Chapter 6	204
	Appendix D Appendix for Chapter 7	207
	Appendix E Codes	209
E.1	Bioinformatics pipeline for processing HIV deep sequence data	209
E.2	The basic Bayesian Poisson flow model	210
E.3	The semi-parametric Bayesian Poisson flow model	212
E.3.1	Gaussian process	212
E.3.2	Hilbert space Gaussian process approximation	215

List of Figures

2.1	Location of the Rakai Community Cohort Study. Left: Location of Rakai District (red) is in south-eastern Uganda at the shores of Lake Victoria. Right: HIV surveillance data were obtained from 2 survey rounds in 36 inland communities of the Rakai Community Cohort Study in Rakai district (green circles) and three survey rounds in the main 4 fishing communities within 3km of Lake Victoria (blue triangles) between August 10, 2011, and January 30, 2015.	11
3.1	Sketch of bioinformatics pipeline. The consensus sequences were summarised from deep sequences and used as an initial step to identify the potential transmission networks. To do this, we explored the couple pairs and set up the standard for classification. The potential networks were further processed to improve the scalability and the robustness. Next, the deep sequences were divided by the potential networks. We added the background sequences to each network, aligned the sequences and constructed the phylogenies. These phylogenies were analysed to obtain the pairwise relationship between the hosts and construct the most likely transmission chains. . . .	15
3.2	Sketch of making potential networks. Potential networks were identified through consensus sequences. These sequences were divided into overlapping windows. In each genome window, we evaluated similarity scores, characterised a Gaussian mixture model through the scores among couples, and extracted the threshold for identifying close pairs. Summary of pairwise relationships across the genome provides the closeness between two individuals. If we consider a graph composed of individuals as nodes and the close relationships as edges, individuals were divided into several connected components. Large components were further divided in order to be processed by phyloscanner and controls were added in case any close pairs were missing.	16

- 3.3 **Histograms of log similarity scores among couples.** The five panels show the histograms of log similarity scores among couples in five equidistant windows. The start and end of the windows are specified in the sub-titles. Each panel shows the bi-modal pattern of the similarity scores across the couples. The high peaks around 0 represent that a majority of individuals acquire infections from its partner, and the low peaks indicate some individuals are infected outside the couple relationship. The mixture model in Section 3.2.3 was used to fit the two-component mixture distributions, and the fitted distributions are shown in the black lines. 19
- 3.4 **Comparison of score distributions among couples using the sequences that were collected prior to January 2015 and all the sequences.** We evaluated the similarity scores from couples over the entire genome, either using sequences collected prior to January 2015 or all the sequences. The scores are shown in the histograms for the samples that were collected before January 2015 in the left and those between August 2011 to 2019 in the right. 21
- 3.5 **Decomposition of score distributions among the couples.** We first focus on the couples that were analysed by Ratmann et al. (2019) in the left, and compare the density of similarity scores using the sequences that were collected before January 2015 in blue and after January 2015 in red. We next analyse the new samples, and visualise the density of the similarity scores by whether the couple was involved before January 2015 (in blue) or not (in red) in the right. 21
- 3.6 **Trace plots for the three parameters with the lowest ESSs in the 68th window.** Stan's dynamic HMC was run for 10^4 iterations on the scores in each genome window. The minimum ESS across the windows was reached in the 68th window. Shown are trace plots for the three parameters with the lowest ESSs after the burn-ins, in order to present the chains clearer and mitigate the effect of initialisation. 22
- 3.7 **Comparison of the estimated mixture distribution parameters in the whole genome using the sequences that were collected before and after January 2015.** Estimated means (top panels) and standard deviations (bottom) of two components (the lower component in the left and the upper component in the right) are shown along the genome positions (in the x-axis). Estimates are described by the median estimates in the lines and the 95% CIs in the shaded areas. 23

- 3.8 **Cutoffs for viral linkage across the genome.** We considered the 5% quantile of the upper component in the mixture model to be the threshold, which separated distant pairs and close pairs. The thresholds are shown along with the genome positions in the line, and the 95% CIs are in the shaded area. 25
- 3.9 **Relationship between cutoffs and distances.** Average similarity scores (3.2) and tree-based distances in each window are shown against the cutoff values in the left and right panels respectively. Each point represents the distance measure and the threshold in a genome window. 25
- 3.10 **Histogram of similarity scores among all the pairwise combinations of individuals in the first window.** We calculated the similarity scores for all the pairwise combinations in the first window, and show the histogram in the left panel. We then focus on the region around the threshold in the right panel. The vertical black lines refer to the threshold on the first window that was depicted in Figure 3.8, and the dashed lines show the 95% CI of the estimated threshold. 25
- 3.11 **Proportion of pairs classified as close along the genome.** We evaluated the proportion of pairs classified as close in each genome window, using the thresholds in Figure 3.8. This figure shows the proportions by couples (in blue, bottom) and all the other pairwise combinations (in red, top). 27
- 3.12 **Histogram of the proportion of windows that support the close relationship.** We classified one pair of consensus sequences on each genome window by the proportion of windows supporting the close relationship. We show the histogram of the proportions for the couples in the left panel and all the pairwise combinations in the right panel. Note that loads of individuals (over 3×10^7) have no windows classified as close among all the samples, and we thus cut the y-axis to 10^4 for the clear visualisation of other parts of the diagram. 27

- 3.13 **The score distributions by the couple status and the classification for the close pairs along the genome windows.** We aggregated the shreds of evidence across the windows to classify an individual pair into close and distant. Here we show the score distributions by the classified relationship and the couple status, under the three standards. The classifications are presented in colour, distant in blue and close in red. The couple status is shown in the panels, couples on the left and other pairs on the right. The three standards are in the top, middle and bottom panels respectively. If at least (A) one (B) 50% (C) 80% windows indicate the close relationship for the two hosts, the pair is considered as close. The score distribution is specified by the median (the darkest line), the interquartile range (the dark shaded region) and the 95% CI (the light shaded region). The black line represents the estimated window-specific threshold in Figure 3.8. 28
- 3.14 **Visualisation of the largest connected component.** We identified close pairs by analysing the genetic distance between consensus sequences. The close relationship performs as links and breaks individuals into connected components. Here we show the largest connected component in Table 3.1. The yellow nodes represent individuals, and the black edges indicate the consensus sequences from two individuals are close. There are 291 individuals in the largest component and it is difficult to process them in one goal in the further analysis. 30
- 3.15 **Decomposition of the largest connected component.** We identified the structure of the largest cluster and decomposed it into 14 densely connected clusters. The clusters are displayed by the blocks of individuals that are represented by the nodes in the same colour. The grey edge represents the within-group link and the red edge shows the link between two clusters. The two ends of the red link are called bridging individuals which connect the clusters. The division of the largest component requires a separate cluster composed of all the bridging individuals, to avoid missing the strongly linked individuals between clusters. 32

- 3.16 **Sketch of deep sequence phylogenetic analysis.** We started from the deep sequences in one potential network and added the background sequences to each network. Background sequences included subtype representatives and east African whole genome sequences, and were aligned to HXB2. The hundreds of background sequences may slow down the phylogenetic analysis, and we thus made a selection while remaining the outcomes unaffected. Then we made deep sequence alignments together with the selected background sequences, and phylogenies were built upon the alignments. The software in all the steps is shown in grey shaded texts. 34
- 3.17 **Estimated TMRCAs using four subsets of background sequences.** We estimated the TMRCa from the reconstructed phylogeny in each genome window for the first potential transmission network. The estimates were smoothed along the genome. The smoothed estimates and the 95% confidence intervals are shown in the coloured lines and grey shaded areas. The colour represents the use of different subsets of background sequences in phylogeny reconstruction. The left panel shows the impact of different types of background sequences and the right panel shows the impact when the chunks of gaps were deleted. 37
- 3.18 **Estimated TMRCAs using the downsampled subsets of background sequences.** We estimated the TMRCa from the reconstructed phylogeny in each genome window for the first potential transmission network. The smoothed estimates and the 95% confidence intervals are shown in the coloured lines and grey shaded areas. The colour represents the subset of downsampled background sequences in phylogeny reconstruction. The left panel shows the impact of down-sampled background sequences, and the right panel shows the impact when the chunks of gaps were deleted. . . . 37
- 3.19 **Measures of alignment quality.** We aligned the sequences, built the phylogenetic tree, and inferred the time of common ancestry in the trees in one of the largest potential networks (404th). We focused on two measures of alignment quality, the length of the largest gap chunks in the alignment and the estimated time of common ancestry. The long branch length is a common indicator for the mistakes in the phylogeny reconstruction. We first show the branch length against one measure of inferior alignment quality on the left, and then display the relationship between the two measures. Each point represents quality measures of the alignment in one genome window. . . . 40

- 3.20 **Example of misalignment and the proposed solution.** We applied the first part of phyloscanner to generate alignment for the sequences in the 404th potential network. Here we took the genome window between 5850 and 6099 as an example. We visualise the subset of the aligned sequences in the left panel and observe too many insertions and blocks of sequences in the incorrect positions. We proposed to re-align the sequences in the left using MAFFT, and this produced the alignment in the right which has the enhanced alignment quality. 40
- 3.21 **Effect of gap removal on the estimated time of the most recent common ancestry.** We used phyloscanner and specified the MAFFT option as "iterative refinement methods using weighted sum-of-pairs and consistency scores" to align the sequences. The resulting alignments sometimes have a large chunk of gaps, especially when we consider the relatively large potential transmission networks. We removed the gap chunks, rebuilt the trees, and estimated the age of common ancestry in one of the largest networks, the 404th network. Then we compare the estimates in boxplots, when (1) all the positions with $\geq 95\%$ gaps were removed, abbreviated as "all" (2) no positions were deleted, abbreviated as "none" (3) only large gap chunks taking at least 6 positions were removed, abbreviated as "large gap chunks". 41
- 3.22 **The effect of realignment strategy on the length of the largest gap chunk.** We aligned the sequences from the first and 404th potential transmission networks in phyloscanner, and show the length of the largest chunk of gaps in the box plots (in blue). To mitigate the problem in Figure 3.20, we applied the realignment to the phyloscanner output and plot the same type of box plots in red. The left shows the comparison for the 1st potential network which includes $< 1,000$ taxa, and the right is for the 404th potential network which includes $> 2,000$ taxa. 41

- 3.23 **The effect of realignment strategy on the estimated time of common ancestry.** We aligned the sequences from the first and 404th potential transmission networks in phyloscanner, built the phylogeny and show the time of common ancestry in the box plots (in blue). To mitigate the problem in Figure 3.20, we applied the realignment to the phyloscanner output and plot the same type of box plots in red. The left shows the comparison for the 1st potential network which includes $< 1,000$ taxa, and the right is for the 404th potential network which includes $> 2,000$ taxa. The dashed black lines act as the reference and represent the known time of HIV passed to humans. 42
- 3.24 **Histograms of summarised quality measures across the potential networks.** We calculated the minimum (left), median (middle) and maximum (right) of the length of the largest gap chunks (on the top) and the estimated time of common ancestry (in the bottom) for each potential network, and plot the histograms. The reference lines are provided by the corresponding quantities in the first and 404th potential transmission networks, and depicted in the black dashed lines. 42
- 3.25 **Illustration of four types of relationships.** The top row illustrates i is ancestral to j , i is descendant to j , i and j are not adjacent. The bottom row shows the complex or no ancestry relationship. The subgraphs of host i are in red, j in blue, and any other hosts in grey. 44
- 3.26 **Histogram of the number of sequence pairs of which the differences in reads are greater than 100.** We gathered the pair of sequences with read differences > 100 across the phyloscanner runs, counted the number for each pair of individuals, and show it in the histogram. 47
- 3.27 **Classification of multiple ancestral relationships.** We display the parts of a phylogeny which contain AID3114 (in blue), AID6098 (in pink), AID1311 (in yellow), and AID0719 (in red) in the genome window 1,250bp-1,499bp. Tips indicate unique read fragments, and the number in the tip label reflects the copy number. 48

4.1	Observed transmission events. The map shows the reconstructed directed transmission events by the residence geo-locations that are randomised by $\pm 2\text{km}$. Arrow colours represent the genders of sources. There are substantial transmissions along the main road in the north-south direction and between the lake (in grey) and the area away from the coast. However, the general pattern stays unclear.	54
4.2	Posterior distributions under weak informative and uninformative priors. We imposed weakly informative (left) and uninformative (right) priors on the length-scales and plot pairwise posterior samples of the length-scales. The divergent transitions are highlighted in red.	63
4.3	Trace plots for parameter inference on the Rakai data set. Stan's dynamic Hamiltonian Monte Carlo was run for 3×10^3 iterations. Shown are trace plots for three parameters with the lowest ESSs.	64
4.4	Estimated directed transmission by gender on a regular grid. A non-parametric model was fitted to the observed directed transmissions. Given a location, we are able to estimate directed transmissions by gender, and decompose these into angles (top left) and distances (top right). Top left: arrows represent transmission directions at a location (points) by gender (in colour). Top right: the distance of transmissions from a location (points) by gender (in shape) is shown in colour. The standard deviations associated with angles and distance are in the bottom left and right panels respectively. The standard deviations of estimates at a spatial location (points) are shown in colour by gender (in shape). As the standard deviations of distance depend on the scale of distance, we considered the coefficient of variations as the uncertainty measure of distance.	68
4.5	Posterior distributions of angles of transmissions at two points with maximum gender differences in angle median estimates in Figure 4.4. The two subplots show distributions of transmission angles from men (blue) and women (red) at each of the two locations.	70
4.6	Estimated recipient locations of directed transmissions by gender on a grid of locations. The transmission destinations were predicted for a grid of points, and are shown on the map of Rakai District. Arrows represent the estimated directed transmissions, and the colour shows the transmissions from men (blue) and women (red).	71

4.7	Estimated recipient locations of directed transmissions by gender on a grid of locations, where we included only between household transmissions. The transmission destinations were predicted for a grid of points, and are shown on the map of Rakai District. Arrows represent the estimated directed transmissions, and the colour shows the transmissions from men (blue) and women (red).	72
5.1	Analysis aims. 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 disproportionately transmit and receive infections.	74
5.2	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 simulated trajectories. (B) Transmission flows were estimated under the Poisson likelihood model (5.12), without adjustments for sampling differences (dark grey), and with adjustments for sampling differences (light grey). Accuracy was measured with the WCE between posterior median estimates and the simulated true values, and shown are the average error and 95% range in 100 replicate simulations.	87
5.3	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 (5.3). From the simulated sampling status data, posterior estimates of the sampling proportions in a and b were obtained with Equation (5.29), and then used in flow inference. We show posterior densities of the sampling probabilities in a and b in black, compared to a Beta(25.5,25.5) prior density in grey. Vertical lines indicate the true sampling fractions, $\xi_a = 0.6$ and $\xi_b = 0.35$	88

- 5.4 **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 (5.3) in Section 5.4.2. Columns show results for four simulated data sets, and rows show results for increasing samples size. The true transmission flows are shown 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 (5.29) (orange). 90
- 5.5 **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 (5.3) as described in Section 5.4.1. Transmission flows were re-estimated with the hierarchical Poisson model (5.12) 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 WCE 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. 91
- 5.6 **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 5.4.3. 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 (5.29) (orange). 92

- 5.7 **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 5.4.3. 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, WCE was reported, after aggregating flow estimates to the 4 flow combinations 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 WCE were similar to those found in the more simple ODE-based simulation experiments. 93
- 5.8 **Sampling cascade of transmission events.** 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 RCCS 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. 95
- 5.9 **Trace plots for parameter inference on the Rakai data set, Poisson flow model with Gamma prior on transmission intensities.** Stan's dynamic HMC algorithm was run using the code for 10^4 iterations. Shown are trace plots for the first 3×10^3 iterations of the three parameters with the lowest ESSs. 96

- 5.10 **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 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. 97
- 5.11 **Transition rates.** We selected high coverage nucleotide sequences of subtypes A1 and D to build phylogenies. We used BEAST to reconstruct discrete ancestral states for each phylogeny and inferred transition rates between inland and fishing communities. The error bar plot shows the transition rates from fishing to inland and from inland to fishing, based on the subtype A1 phylogeny (left panel) and the subtype D phylogeny (right panel). 102
- 5.12 **Indicator for non-zero transition rates.** We used BEAST to reconstruct discrete ancestral states for each phylogeny and inferred transition rates between inland and fishing communities. Meanwhile, we applied BSSVS to select the nonzero rates. The error bar plot shows the indicator for non-zero transition rates from fishing to inland and from inland to fishing, based on the subtype A1 phylogeny (left panel) and the subtype D phylogeny (right panel). 102
- 5.13 **Trace plots of the birth-death model.** Three thinned chains (in colour) of sampling fractions without the burn-in are combined and plotted in lines, showing these chains converge to different stationary distributions. 105

- 6.1 **Age-specific transmission data.** 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 the 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. 110
- 6.2 **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 (6.2), 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 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. 116
- 6.3 **Estimated sources of infection in women under the final HSGP approximation.** We simulated age-specific transmission flows under the kernel parameters shown in Table 6.1, and re-estimated transmission flows along with all other parameters under HSGP approximations to (6.2). 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 (CI), showing 40% CIs in dark red to 80% CIs in light red. Posterior medians are shown as red lines. True values are shown in black. We found no systematic differences in source estimates when using the final HSGP approximation over the GP prior. 117

- 6.4 **Estimated covariance kernel parameters under the final HSGP approximation.** We simulated age-specific transmission flows under the kernel parameters shown in Table 6.1, and re-estimated transmission flows along with all other parameters under HSGP approximations to (6.2). The sub-figures illustrate the marginal posterior densities of the kernel parameters under the GP model (in red) and the HSGP approximation (in blue) based on $B = 1.25, m = 30$ (right panels). 118
- 6.5 **Trace plots for parameter inference on the Rakai data set, Poisson flow model with HSGP prior on the transmission intensities.** Stan's dynamic HMC algorithm was run for 10^4 iterations. Shown are trace plots for the first 3×10^3 iterations of the three parameters with the lowest effective sample sizes. 120
- 6.6 **Impact of using regularising prior densities on the 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 (5.16), to estimates obtained under the regularising HSGP prior on the log transmission intensities (6.7). 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. 121
- 6.7 **Unregularised flow estimates by age bands and genders.** We compared estimates of the sources of HIV infection in women without regularisation, using the Gamma prior densities (5.16), when 2-year age bands and 5-year age bands were applied. Each colour corresponds to infection recipients of a particular age group. **(A)** Each line shows the posterior median estimates of the probability of infections attributed to the 2-year increment age group of individuals of the opposite gender, i.e. the sources of infection, obtained without regularisation. **(B)** Same as A, using 5-year increment age groups. 122

- 6.8 **Aggregated male-to-female transmission counts.** We compare 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 that are the same age as recipients (top) and 5 years older than recipients. 124
- 6.9 **Estimated male-to-female transmission flows.** Using the aggregated counts in Figure 6.8, we modelled the transmission intensities using the GP prior densities (6.2), and predicted flows by 1-year age bands. Panels A-D show the estimated transmission flows (in colour) from men (x-axis) to women (y-axis) using corresponding data in Figure 6.8. The two black dashed lines are reference lines, representing sources that are the same age as recipients (top) and 5 years older than recipients. 125
- 6.10 **Source of infections in women under 4 types of aggregation.** Using the aggregated counts in Figure 6.8, we modelled the transmission intensities using the GP prior densities (6.2), predicted flows by 1-year age bands and evaluated 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) are shown against ages of female recipients in the bar plots by four types of aggregations (in colour). 126
- 6.11 **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 CIs. 127
- 6.12 **Estimated sources of infection in men in Rakai, Uganda, 2009-2015.** We estimated the sources of HIV infection in men of increasing age. Sources were defined as women that were younger or the same age as the men (grey), 1-5 years older (orange), and over 5 years older (blue), and sum to 100% for each age of infected men on the x-axis. Posterior median source estimates are shown along with 95% marginal CIs. 127

- 6.13 Comparison of male to female and female to male transmission flows.** We visualise posterior median flow estimates between age groups for male-to-female (left) and female-to-male (right) transmissions through a heat map. Red dots show the observed phylogenetically likely source-recipient pairs and contours are 2D kernel density estimates. The white dashed lines are diagonal and provide a reference for transmission between individuals of the same ages. 131
- 6.14 Estimated age profile of male and female sources.** Shown is the contribution of age groups to the infections in women of all ages (left) and infections in men of all ages (right). Posterior median estimates are shown as points, and associated 95% CIs are plotted by error bars. 131
- 6.15 Estimated age profile of infections by age of the infected partner.** We estimated the flows between one-year age bands, and normalised the flows to one age group (in the x-axis) among women (in the left) and men (in the right). Posterior median estimates are plotted by lines, and the colours of the lines indicate the given age of the recipients. 132
- 6.16 Estimated age profile of infections by age of infected partner, second version.** The left panel shows posterior interquartile ranges of male sources to infection in women, and the right panel shows posterior interquartile ranges of female sources to infection in men. Posterior medians are shown as lines, and the 50% highest posterior density interval (HPI) and 90% HPI are shown in darker and lighter shaded areas. The diagonal dashed lines act as the reference and indicate the same source and recipient ages. 132
- 6.17 Comparison of sampling adjusted and sampling unadjusted transmission flows.** The two figures on the top show the relative differences (abbreviated by rd) of flow estimates (by posterior medians) with and without sampling adjustment, for male-to-female transmissions (left) and female-to-male transmissions (right). The two figures on the bottom show sampling fractions of transmissions by one-year age band and gender. 133

7.1	Log marginal likelihood estimates for the candidate models over iterations, using the standard bridge sampling. We estimated the log marginal likelihood using the standard bridge sampling scheme in (7.13). The estimates were updated along with the iterative scheme, and are shown against the iterations (in the x-axis). The legend specifies the model features, including the number of Gaussian process components (in the line types) and the number of baseline terms (in the colours).	151
7.2	Log marginal likelihood estimates for the candidate models over iterations, using the warp-III bridge sampling. We estimated the log marginal likelihood using the warp-III bridge sampling in (7.34). The estimates were updated along with the iterative scheme, and are shown against the iterations (in the x-axis). The legend specifies the model features, including the number of Gaussian process components (in the line types) and the number of baseline terms (in the colours).	151
7.3	Comparison of log marginal likelihood for the candidate models. We evaluated the log marginal likelihoods using the warp-III bridge sampler for seven candidate models in (7.41)-(7.47), and show the median and range of the estimates in the points and error bars respectively. The colours represent models and are comparable to those in Figure 7.2.	152
8.1	Time evolution of age-specific reproduction numbers across states. Each panel shows the estimated posterior probability (in colour) that the daily effective reproduction number in each age group (in the y-axis) is < 1 over time (in the x-axis).	164
8.2	New cases by age. Each panel shows the proportion of new daily COVID-19 infections (in the y-axis) for seven age bands (in colour) along the time (in the x-axis).	165
8.3	Estimated accumulated contribution of age groups to COVID-19 cases for October, versus the proportion of the population in the same age group. Posterior median estimates of the contributions from each age group (blue fill bars with 95% credible intervals) are compared against the population age composition of each state (black contour bars).	166
A.1	Marginal posterior densities of regression coefficients of model P4 to estimate population participation probabilities. Circle: median; bar: interquartile range; line: 95% credibility interval.	189

A.2	Marginal posterior densities of population participation probabilities under model P4. Circle: median; line: 95% credibility interval.	190
A.3	Marginal posterior densities of regression coefficients of model S5 to estimate conditional deep-sequencing probabilities. Circle: median; bar: interquartile range; line: 95% credibility interval.	191
A.4	Marginal posterior densities of source sequencing probabilities under model S5. Circle: median; line: 95% credibility interval.	193
A.5	Marginal posterior densities of regression coefficients of model N3 to estimate conditional deep-sequencing probabilities. Circle: median; bar: interquartile range; line: 95% credibility interval.	195
A.6	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.	199
A.7	Marginal posterior densities of recipient sequencing probabilities under model S'5. Circle: median; line: 95% credibility interval.	200
A.8	Comparison of marginal posterior densities of regression coefficients between model S5 and model S'5. Circle: median; bar: interquartile range; line 95% credibility interval.	201
A.9	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.	202
A.10	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.	202

List of Tables

3.1	Size distributions of identified networks.	29
3.2	Size distributions of networks after application of the multi-level modularity optimisation algorithm.	31
3.3	Inference accuracy on the seroconversion pairs when down-sampling is applied or not.	47
3.4	Inference accuracy on the seroconversion pairs when zero branch lengths between subgraphs are taken into account or not in determining the pairwise relationship.	50
3.5	Linked and directed pairs when zero branch lengths between subgraphs are taken into account in determining the pairwise relationship.	50
3.6	Inference accuracy on the seroconversion pairs when down-sampling is applied or not, taking the zero branch lengths between ancestral subgraphs into account.	51
3.7	Cluster size distributions of transmission networks.	51
3.8	Reconstructed pairs in the most likely transmission chain when zero lengths are adjusted.	52
4.1	Numerical performance of the Gaussian process flow model.	64
4.2	Comparison of prediction accuracy in the non-parametric Gaussian Process flow model.	67
4.3	Summary of posterior median of flow directions in the regions stratified by the latitude -0.7485917 and longitude 31.70931. The summary is provided by median (minimum - maximum).	67
5.1	Parameter values used for the ODE-based simulations.	86
5.2	Additional model performance measures in the ODE based simulation. . .	89
5.3	Performance of MCMC and HMC under the Gamma prior.	94

5.4	Comparison of conditional flow ratios in "mugration" model and deep sequence based method.	101
5.5	Prior choices for the multi-type birth and death model.	104
6.1	Parameter values used to simulate transmission flows in populations stratified by 1-year age bands under a variation of GP model (6.2).	115
A.1	Prediction accuracy of models P1-P4 on hold-out population participation data in 10-fold cross-validation.	189
A.2	Prediction accuracy of models S1-S8 on hold-out deep-sequencing data in 10-fold cross-validation.	191
A.3	Prediction accuracy of models N1-N3 on hold-out infection time data in 10-fold cross-validation.	194
A.4	Demographic characteristics of infected participants by estimated infection time.	197
A.5	Prediction accuracy of models S'1-S'8 on hold-out deep-sequencing data in 10-fold cross-validation.	198

Acronyms

AIDS Acquired Immune Deficiency Syndrome

ART Antiretroviral Therapy

BF Bayes Factor

BSSVS Bayesian Stochastic Search Variable Selection

COVID-19 Coronavirus Disease 2019

CRF Circulating Recombinant Form

CRPS Continuous Ranked Probabilistic Score

CTMC Continuous-Time Markov Chain

ELPD Expected Log pointwise Predictive Density

ESS Effective Sample Size

GP Gaussian Process

HIV Human Immunodeficiency Virus

HMC Hamiltonian Monte Carlo

HSGP Hilbert Space Gaussian process approximation

HXB2 Human immunodeficiency virus type 1 group M subtype B

LANL Los Alamos National Laboratory

LKJ Lewandowski-Kurowicka-Joe

LOO Leave-One-Out Cross-Validation

MAE Mean Absolute Error

MAFFT Multiple Alignment using Fast Fourier Transform

MCMC Markov Chain Monte Carlo

NGS Next-Generation Sequencing

NUTS No-U-Turn Sampler

ODE Ordinary Differential Equation

PANGEA Phylogenetics And Networks for Generalised Epidemics in Africa consortium

PPrCI Posterior Predictive Credibility Interval

PSIS Pareto Smoothed Importance Sampling

RCCS Rakai Community Cohort Study

RHSP Rakai Health Science Programme

RMSE Relative Mean-Square Error

SARWS Sampled At Random Within Strata

SIR Susceptible-Infected-Removed

STD Sexually Transmitted Diseases

TMRCA Time to Most Recent Common Ancestor

VMC Voluntary Medical Male Circumcision

WCE Worst Case Error

Chapter 1

Introduction

1.1 The global HIV epidemic

Since its emergence in 1981, the human immunodeficiency virus (HIV) has caused ongoing epidemics across the globe. The latest estimates indicate that in 2020, 37.7 million [30.2 million–45.1 million] individuals were globally infected with HIV, among which 1.5 million [1.0 million–2.0 million] were newly infected, and this resulted in 680 000 [480 000–1 million] deaths (UNAIDS, 2021). In 2013, the United Nations set the goal to reach the 90-90-90 target by 2020, and end HIV epidemics by 2030 (UNAIDS, 2017). The 90-90-90 goal means that 90% of people who are HIV infected will be diagnosed, 90% of people who are diagnosed will be on antiretroviral treatment and 90% of those who receive antiretrovirals will be virally suppressed. Though HIV treatment can effectively reduce infections and acquired immune deficiency syndrome (AIDS) related deaths, the 90-90-90 target was missed by 2020 (UNAIDS, 2020). All world regions are significantly affected by HIV, but Sub-Saharan Africa remains the most affected area in terms of the number of new infections and HIV prevalence (CDC, 2021).

In Sub-Saharan Africa, which is the home to 12% of the world population, about 60% of all new infections occur. A large number of effective HIV interventions are available to reduce viral spread, including 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 the onward transmission (Grabowski et al., 2017; Kagaayi et al., 2019). Since the widespread adoption of such prevention measures in Africa, rates of incident cases have overall dropped considerably, although they remain well above the Joint United Nations Programme on HIV/AIDS (UNAIDS) threshold for HIV elimination (UNAIDS, 2018b). One particular challenge is the marked heterogeneity in HIV

epidemics. For example, Dwyer-Lindgren et al. (2019) observed considerable differences in HIV prevalence within countries. In particular, HIV prevalence varies between 9% and 43% between communities within the Rakai District. In addition, three-quarters of worldwide new infections among 15-19-year-old adolescents are among girls in sub-Saharan Africa (UNAIDS, 2018a). Therefore, the risk of acquiring and spreading HIV infection varies substantially across locations, genders and ages. It is urgent to understand the population groups that drive transmission, optimise the resource allocation, and maximise the effect of the intervention.

A great deal of effort has been devoted to identifying high-risk populations. The standard approach is based on longitudinal follow-up of uninfected individuals, who are later observed HIV positive. The data are often subject to reporting or social desirability biases, especially for sexual diseases that are associated with stigma or remain criminalised in many countries (Barré-Sinoussi et al., 2018). Zuma et al. (2003) randomly sampled women in Khutsong, distributed questionnaires on the background and sexual behaviour and provided screening for HIV infections. The study found 37.1% higher prevalence among migrant women than the residents, and HIV is most prevalent among women aged 26-35 (50.9%). Wawer et al. (2005) focused on monogamous, HIV-discordant couples from a randomised, community trial of intensive sexually transmitted diseases control for AIDS prevention in Rakai, and estimated HIV transmission rate per coital act by the stage of infection in the index partner. The study suggested higher transmission risks associated with younger index partners and early-stage infections. Mermin et al. (2008) analysed Uganda HIV/AIDS Sero-Behavioral Survey (UHSBS), identified recent infections in the past 155 days through laboratory testing and studied the association between risk factors and recent infections. It was shown that recent HIV infections are associated with female, marital status, geographical locations, and other factors. Bunnell et al. (2008) assessed factors associated with HIV risk behaviours among HIV infected adults in UHSBS. They found an association between awareness of HIV status and condom use and highlighted the need to test HIV infected people and their partners. These classic studies either investigated the risk factors associated with HIV prevalence and incidence, or transmission risks across couples. Despite clear methodological approaches to identify risk factors associated with HIV status, these analyses do not describe or identify the population groups that drive the HIV transmission.

1.2 Using HIV sequence data to characterise transmission flows

HIV sequence data can provide new insights to understand transmission dynamics. Pathogen sequences are the result of an evolutionary process that occurs in infected individuals. Thus, for fast-evolving pathogens like HIV, the phylogenetic relationship of pathogen sequences can be used to evaluate transmission dynamics. These methods have, for example, been used to determine the origins of HIV (Faria et al., 2014), to quantify the contribution of different disease phases to onward spread (Ratmann et al., 2016; Volz et al., 2013), to improve outbreak detection (Poon et al., 2016), and they have been applied to a large range of other infectious diseases (Dellicour et al., 2018; Didelot et al., 2017; Uhlemann et al., 2014).

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 (Hu   et al., 2005), however, the data are insufficient to estimate the direction of transmission between any two sampled individuals (Leitner and Romero-Severson, 2018). For this reason, most available methods construct phylogenies from Sanger sequences, and use mathematical models to describe the process that generates the phylogeny. Thus, these approaches infer the flow between population sub-groups 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 the time of sampling (such as location or age).

Specifically, the migration model (Lemey et al., 2009) describes the states of viral lineages at any time 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 Markov Chain Monte Carlo (MCMC) sampling, as well as posterior estimates of the phylogeny and the states of its lineages, which are latent variables in the model. Despite the efficiency in computation, this model is strongly sensitive to sampling bias (De Maio et al., 2015). The MultiTypeTree model (Vaughan et al., 2014) relaxes the independence assumption and allows the coalescent rate explicitly dependent on states of the lineage, however, the induced model complexity often renders sampling the latent phylogeny and state history computationally infeasible. The computational challenge was addressed by 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 a pre-specified ordinary differential equation (ODE) model of disease spread. The flow parameters are obtained as by-products of the estimated latent states and parameters of the compartment model, and in general vary in time over the phylogenetic history. Müller et al. (2017) and others (Poppinga et al., 2015) have clarified the assumptions that underlie the marginalisation of state histories into lineage state probabilities, which led to more accurate maximum likelihood and Bayesian estimation routines (De Maio et al., 2015; Müller et al., 2018; Volz and Siveroni, 2018). With this marginalisation trick, a greater range of flow models and data sets can be analysed, though computational run-times often remain on the order of several weeks for data from hundreds of individuals (Rasmussen et al., 2018; Ratmann et al., 2017; Stadler and Bonhoeffer, 2013). Another popular family of model is to describe the generative process of phylogenies by forward-time epidemiological models, rather than the backward-time coalescent models. Stadler et al. (2012) applies birth-death models to describe the phylogeny, and it is generalised to varying population sizes (Stadler et al., 2013) and structured population (Kühnert et al., 2016), from which the flow parameters can be evaluated from birth rates.

An alternative strategy is to conduct a phylogenetic analysis of multiple distinct pathogen sequences per infected host because such data make it possible to attribute sets of viral lineages to individuals, analyse the topological relationship between the sets, and infer the ancestral relationships between individuals, which can provide direct evidence into the direction of transmission between two individuals (Leitner and Romero-Severson, 2018; Romero-Severson et al., 2016). Such analyses are becoming broadly applicable to fast-evolving pathogens like HIV because standard deep sequencing protocols generate thousands to millions of distinct pathogen sequence fragments (Bonsall et al., 2018; Gall et al., 2012; Zhang et al., 2020b). In contrast to classical approaches augmenting phylogenetic data with epidemiological data such as sampling dates and contacts through mathematical models, deep sequence data prompt us to obtain transmission pairs without external self-reported data and extra modelling assumptions (Didelot et al., 2017; Hall et al., 2015; Jombart et al., 2014; Volz and Frost, 2013), and thus reduce the bias in estimating transmission flows. PhyloScanner (Wymant et al., 2017) is the first software to systematically explore transmissions from multiple genomes per host.

Viral sequence data has been relatively frequently used on data from North America and Europe, but such studies are rare in sub-Saharan Africa because of the limited number of available viral sequences. The Phylogenetics And Networks for Generalised Epidemics in Africa consortium (PANGAEA) has substantially increased the number of available full genome

sequences in sub-Saharan Africa, and through these data are contributing to understanding HIV epidemics in sub-Saharan Africa. PANGEA provides access to deep sequences from > 20,000 HIV-infected individuals from sub-Saharan Africa, including dense population-based samples from locations where the burden of HIV is among the highest in Africa (Abeler-Dörner et al., 2019). Prior work focused on software development (Skums et al., 2018; Wymant et al., 2017), validation of the bioinformatics protocol for inferring the direction of transmission (Ratmann et al., 2019; Rose et al., 2019; Todesco et al., 2019; Zhang et al., 2020b), and reconstruction of partially observed transmission networks at the population level in the case of HIV (Ratmann et al., 2019).

1.3 Aim, contributions and structure

This thesis aims to provide statistical models to understand population-level transmission dynamics from HIV deep sequence data. The emerging next-generation sequencing (NGS) technology enables new approaches for phylodynamic analyses. In particular, *phyloscanner* (Wymant et al., 2017) is an automatic software for this purpose and reconstructs the directed transmission events from deep sequence data only. Here we first improve the bioinformatics pipeline for processing the deep sequence data. Then we focus on the reconstructed networks and interpret the network through Bayesian modelling. This may prompt us to better understand the risk of subpopulations in transmitting and acquiring HIV, and in turn, guide intervention strategies. Our method offers a new option to explore phylodynamics, and potentially adds value to HIV research. For instance, we are able to address the problem of substantial spatial sampling heterogeneity, infer the transmission dynamics at a high resolution and potentially identify transmission-associated risk factors, and these are difficult to solve in standard phylodynamics methods due to the scalability. The thesis was motivated by data obtained as a part of the Rakai Community Cohort Study (RCCS), and in turn, we applied the developed methods to derive insights on HIV transmission dynamics in the Rakai setting. Apart from HIV, the flow modelling is transferable to Coronavirus Disease 2019 (COVID-19) where we have no viral sequences, which will be explained in Section 1.4.

The main contributions of the thesis are summarised as follows.

- Chapter 3 extends the approach in Ratmann et al. (2019) to be scalable for larger datasets, examines the existing options in *phyloscanner* (Wymant et al., 2017) and adds new options to correct the misclassification.

- Chapter 4 transfers the ideas in the meteorology to infectious disease transmissions, and fits a flexible Gaussian process model to the source-recipient pairs reconstructed with fewer assumptions. This has never been explored in the phylodynamics literature, mainly because of unavailable source-recipient pairs from the standard methods.
- Chapter 5 - 7 proposes a flexible Bayesian hierarchical model to analyse the outputs from phyloscanner (Wymant et al., 2017) and infer the transmission flows between subpopulations. The novelty of our model is the ability to adjust for the bias introduced by the sampling procedure and thus consider the hard-to-reach populations. In addition, leveraging the information from the neighbouring age groups provides more accurate and robust inference of high-resolution age-specific transmission dynamics. In standard phylodynamics methods, it is usually difficult or even impossible to achieve these features. When building the statistical models, we utilise the system of Gamma likelihoods to facilitate the mixing under the multinomial likelihoods. Moreover, we are one of the first applications of Hilbert Space Gaussian process approximation (HSGP) and explore the practical choices of hyperparameters in our setting.
- Chapter 8 is another similar problem of interests that characterises COVID-19 age-specific transmissions. Rather than using the viral sequences, it combines multiple data sources including mortality, mobility and high resolution contacts. In light of this, we draw a comparison between approaches based on viral sequences and epidemiological surveillance data.

The thesis is structured as follows. Chapter 2 introduces the study procedure of the RCCS and demographic characteristics that will be used in the thesis. Chapter 3 processes the deep sequence data and reconstructs transmission events through phylogenetic analyses. Chapter 4 conducts an exploratory analysis of the reconstructed transmission networks and investigates the spatial structure of transmission flows. Chapter 5 proposes a more appropriate semi-parametric model to describe the population-level transmission flows between population strata. The key feature of the model is to adjust for unequal or disproportionate sampling by data augmentation. Chapter 6 further extends the model to leverage the flows across one-year-increment age bands through the smoothing techniques, and this allows for estimating flows with high granularity. Chapter 7 considers several candidate models and provides the approach to select the one that best describes the data. Chapter 8 transfers the flow modelling to COVID-19, again highlights importance of the flows, and draws a comparison with the viral sequence based approaches. Chapter 9 summarises the key findings and discusses the future research directions.

A full list of publications related to my contributions to research on phylodynamics and HIV includes Xi et al. (2022), Ratmann et al. (2020) and Bbosa et al. (2020). I am the first author of Xi et al. (2022). I contributed to the statistical model development and analysis in Ratmann et al. (2020) and Bbosa et al. (2020).

1.4 Other projects

Apart from the phylodynamics theme that forms the core of my thesis, I have been involved in the Imperial College COVID-19 Response Team since February 2020. Early on, my main responsibilities were to (1) collect COVID-19 data in China mainland from provincial health commission websites, help coordinate the collection work, control the qualities of data, and visualise data (Fu et al., 2021); (2) translate research reports and outcomes to Chinese in order to communicate the findings to the public; and (3) check Chinese control policies and news to aid interpretation of research at Imperial. As the first wave of epidemics in mainland China ended, I contributed to several projects led by our research group in the Department of Mathematics. Here, my main contributions were to help with model checking, design and implementation of components of our age-specific transmission model in Stan, as well as data collection. My contribution to our Science paper involved coding as generated quantities of the model, the estimated transmission flows and sources of COVID-19 infection, and other smaller tasks, which overall took several months. Chapter 8 provides the details, reflect my contributions to this work and compare with my main sequences-based works.

The studies on COVID-19, to which I contributed more substantially, are Monod et al. (2021), Fu et al. (2021), Unwin et al. (2020), Ainslie et al. (2020), Geidelberg et al. (2021) and Knock et al. (2021). I am the co-first author of Monod et al. (2021) and Fu et al. (2021). I contributed to model checking (Unwin et al., 2020), data collection (Ainslie et al., 2020), policy checking (Geidelberg et al., 2021) and data validation (Knock et al., 2021) in other works.

Other studies, to which I contributed principally through being a member of the COVID-19 Response Team, participating in weekly meetings, and the review of early drafts, include Verity et al. (2020), Bhatia et al. (2020), Thompson et al. (2020), Grassly et al. (2020), Hogan et al. (2020), Walker et al. (2020) and Nouvellet et al. (2021).

Chapter 2

Data

My research in this thesis is based on human immunodeficiency virus (HIV) deep sequence data from the Rakai Community Cohort Study (RCCS). The RCCS is an open observational study that was initiated in 1994 by the Rakai Health Science Programme (RHSP). The study follows a cohort of adults aged 15-49 from 40 communities in Rakai (Chang et al., 2016; Grabowski et al., 2017) (Figure 2.1). The communities consist of 36 inland communities (green circles) and 4 fishing sites around Lake Victoria (blue triangles). The slightly darker orange line in Figure 2.1 (right) is the Masaka–Mutukula Road which passes over the district in the north-south direction. This is the main road in Rakai district and we called it in brief as the highway in the thesis.

The RCCS continuously surveyed inland residents over the 18 survey rounds between February 1994 and May 2018, of which each lasted for roughly 18 months. It began to include fishing sites since 2011. To identify eligible participants, a household census was conducted preceding each survey round, during which all the persons regardless of their presence or not in the home were enumerated. As an open cohort, it enrolled new eligible individuals such as new 15-year-olds and immigrants in each round (Santelli et al., 2021). After the census, the RCCS surveyed all the residents who were present and within the eligible age range (15-49) and who provided written informed consent. The participants were interviewed for demographic characteristics, sexual behaviours, antiretroviral therapy (ART) use, and male circumcision status. Also, a serum sample was collected at the time of each survey for HIV testing and potential HIV-1 viral sequencing in the subsequent analysis. Since 94% of those who self-reported ART status are also viral-load suppressed (<1000 Ribonucleic acid (RNA) copies per millilitre) (Grabowski et al., 2017), we used self-reported ART use to approximate actual viral suppression and focused on those who did not report themselves on ART for viral

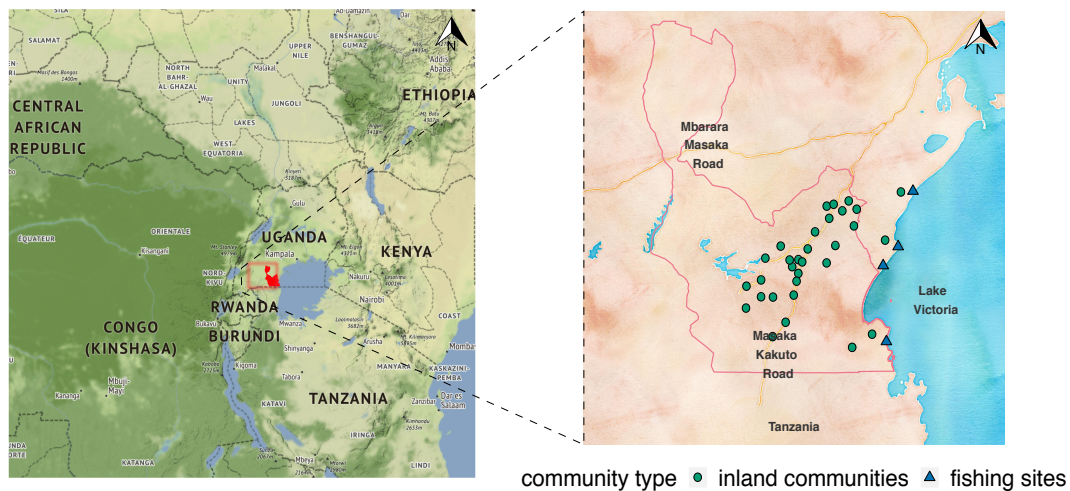


Figure 2.1 Location of the Rakai Community Cohort Study. Left: Location of Rakai District (red) is in south-eastern Uganda at the shores of Lake Victoria. Right: HIV surveillance data were obtained from 2 survey rounds in 36 inland communities of the Rakai Community Cohort Study in Rakai district (green circles) and three survey rounds in the main 4 fishing communities within 3km of Lake Victoria (blue triangles) between August 10, 2011, and January 30, 2015.

sequencing. Viral genomes were deep sequenced for the self-reported ART naïve individuals under the Next-Generation Sequencing (NGS) protocol (Gall et al., 2012).

The analyses in the thesis are based on short viral sequence fragments (reads) from deep sequencing. Ratmann et al. (2019) processed the deep sequences that were collected during August 2011 and January 2015 (rounds 15 to 16), and we will use the reconstructed transmission pairs in the subsequent statistical analysis in Chapter 4 to 7. Recently, the Phylogenetics And Networks for Generalised Epidemics in Africa consortium (PANGEA) provided us with more data, and we are able to extend the study period to 2011.8 - 2018.5 (rounds 15 to 18). However, it requires an approach that scales the existing methods by Ratmann et al. (2019) up to the larger dataset.

We next highlight the demographic data that will be used between Chapter 4 and Chapter 7, in order to interpret the reconstructed transmission networks. The detailed data on the sampling procedure are available in the RCCS through household census, which well characterises the potential bias introduced by participation and sequencing. The majority of non-participants were away for work and study. As a result, participation increased with age, women tended to be more engaged, while young men tended to be hard to reach. In addition,

fishing sites had lower ART coverage than inland communities, and thus more individuals were sequenced. In particular, ART services have been provided to inland communities since 2004 by the RHSP through the support from the US President's Emergency Plan for AIDS Relief (PEPFAR). In contrast, HIV prevention and treatment services were provided by non-governmental organisations to fishing sites in the past, and have been offered by RHSP since fishing sites were added to RCCS in 2011. Therefore, it is essential to stratify the population by age, gender and location and capture the heterogeneous sampling fractions across population groups.

Apart from the sampling differences, fishing and inland communities may show great heterogeneity in risk factors and prevalence (Chang et al., 2016). It is thus stimulating to model the transmission risks in the two locations. Migration is a primary factor that drives the viral dissemination (Grabowski et al., 2020). To account for this, we considered the sampling location and the origin of transmissions separately. If an individual migrated into Rakai within two years, we assume s/he was infected in the origin of migration. Thus, the individual was infected in the migration origin and sampled in the migration destination. Moreover, the age in our analysis refers to the age of individuals at the midpoint of the study, i.e. 2013-05-05. As a result, the age differences are comparable across all the individual pairs. Another option is to use the age at the time when the recipient was infected, and this reflects more accurate age dynamics at the time of infection. It would be a question of epidemiological interests and requires the estimated time of infection, which has recently been possible with deep sequence data in the upcoming package HIVPhyloTSI by Dr. Tanya Golubchiki. Here we focus on the first option, and another option is for future research.

Chapter 3

Bioinformatics pipeline for processing HIV deep sequence data

3.1 Introduction

The aim of this Chapter is to describe updates to the human immunodeficiency virus (HIV) deep-sequence analysis protocol that are based on my research. This Section first describes the analysis procedure proposed by Ratmann et al. (2019). The remaining Sections highlight the changes we made with the purpose of handling larger data and improving performance.

Ratmann et al. (2019) included the individuals if read depths are ≥ 30 on at least 750 positions. The authors divided the individuals into batches of sizes 50 to 75, and performed *phyloscanner* (Wymant et al., 2017) on all possible pairs of batches. In the first stage of analysis, the authors implemented *phyloscanner*, identified potentially phylogenetically close pairs and divided individuals into potential transmission networks in which the individuals were connected by at least one common phylogenetically close individual. They added couples and ten closest individuals as controls to each network. In the second stage of the analysis, *phyloscanner* was used to confirm the transmission pairs and determine transmission directions.

We adopted the similar approach as Ratmann et al. (2019). Due to the inclusion of sequences in two additional survey rounds, the strategy of *phyloscanner* run on all possible pairs of batches becomes implausible. Instead, we used the summary of deep sequences, to quantify the closeness between individuals and thereby identify the potential transmission networks in the first stage analysis. The summary was provided by the most common nucleotide base at the genome position per host, and this is called HIV consensus sequence. In the second stage, we applied *phyloscanner*, and updated some options based on the

experience of Dr. Matthew Hall. Furthermore, we revisited the ancestral relationships classification, spotted some misclassifications, and fixed these issues by incorporating branch lengths. We aim to organise the latest methods into an automated workflow that can be easily applied to other PANGEA-HIV cohorts. The codes are available at https://github.com/olli0601/Phyloscanner.R.utilities/tree/master/misc_data_analysis_RCCS1519/software and the instructions are in https://github.com/olli0601/Phyloscanner.R.utilities/blob/master/misc_data_analysis_RCCS1519/software/phylosanner_run.html.

I summarise our proposed workflow before digging into the detailed procedures.

1. Identify potential transmission network of phylogenetically close individuals, based on analysis of the HIV consensus genome obtained from individuals. The Oxford team pre-processed the HIV deep sequences, mapped them to the reference genome tailored to each sample and reconstructed the consensus sequences in *shiver* (Wymant et al., 2018). We applied the simplest measure of closeness, Hamming similarity, to evaluate similarity scores between consensus sequences and give a rough indication of potential close individuals, where Hamming similarity is defined as the proportion of same nucleotide between two sequences. Identification of potential close individuals requires the pair of individuals that report a long-term sexual relationship, named couples, as the training data where the transmission intermediates are unlikely. And we called all the other pairwise combinations of individuals who do not report a long-term sexual relationship or report a short-term sexual relationship as other pairs. The genetically potential close individuals were considered to be linked in a graph, and naturally broke individuals into several small clusters, on which it became feasible to conduct the subsequent *phyloscanner* analysis. This step is described in detail in Section 3.2, and the codes are in https://github.com/olli0601/Phyloscanner.R.utilities/blob/master/misc_data_analysis_RCCS1519/Rk1519_02_make_potential_transmission_networks.R.
2. Conduct the phylogenetic analysis in each potential transmission network, based on the deep sequence data. Specifically, we discussed how to select a subset of all available background sequences while maintaining the same performance as the entire set. After adding the selected background sequences to each of the potential transmission networks, we then aligned the deep sequence, quantified its alignment quality and made several attempts to enhance the quality. We disabled the functionality of *phyloscanner* to make trees upon the alignment, and built the phylogenetic trees in *IQ-TREE* (Nguyen et al., 2015). The step is detailed in Section 3.3, and the code

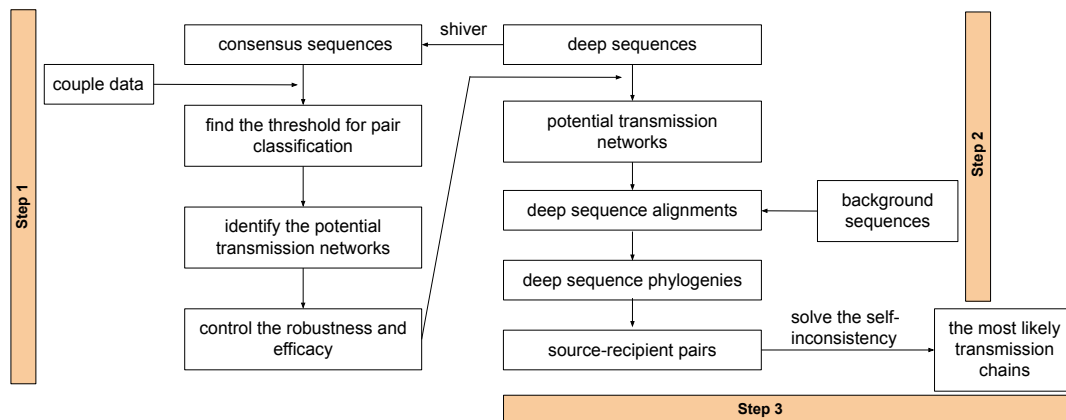


Figure 3.1 **Sketch of bioinformatics pipeline.** The consensus sequences were summarised from deep sequences and used as an initial step to identify the potential transmission networks. To do this, we explored the couple pairs and set up the standard for classification. The potential networks were further processed to improve the scalability and the robustness. Next, the deep sequences were divided by the potential networks. We added the background sequences to each network, aligned the sequences and constructed the phylogenies. These phylogenies were analysed to obtain the pairwise relationship between the hosts and construct the most likely transmission chains.

is available at https://github.com/olli0601/Phyloscanner.R.utilities/blob/master/misc_data_analysis_RCCS1519/Rk1519_03_make_trees.R.

3. Reconstruct the source-recipient pairs, based on the deep sequences phylogenies. We mainly introduced a new option for `phyloscanner` which may reduce the misclassification of pairwise relationships in the phylogeny. In addition, we experimented with the existing options in `phyloscanner` and discussed their performance. We chose the command options based on these experiments and the experience of our collaborators. The summarised pairwise relationship across the trees provided evidence on who infected whom. The inferred source-recipient pairs were processed to eliminate the cycles and thereby reconstruct the most likely transmission chains. The step is in Section 3.4, and available at https://github.com/olli0601/Phyloscanner.R.utilities/blob/master/misc_data_analysis_RCCS1519/Rk1519_04_pairwise_relationship.R.

These procedures are summarised and sketched in Figure 3.1, and their correspondence to the three steps of the pipeline is shown in the orange boxes.

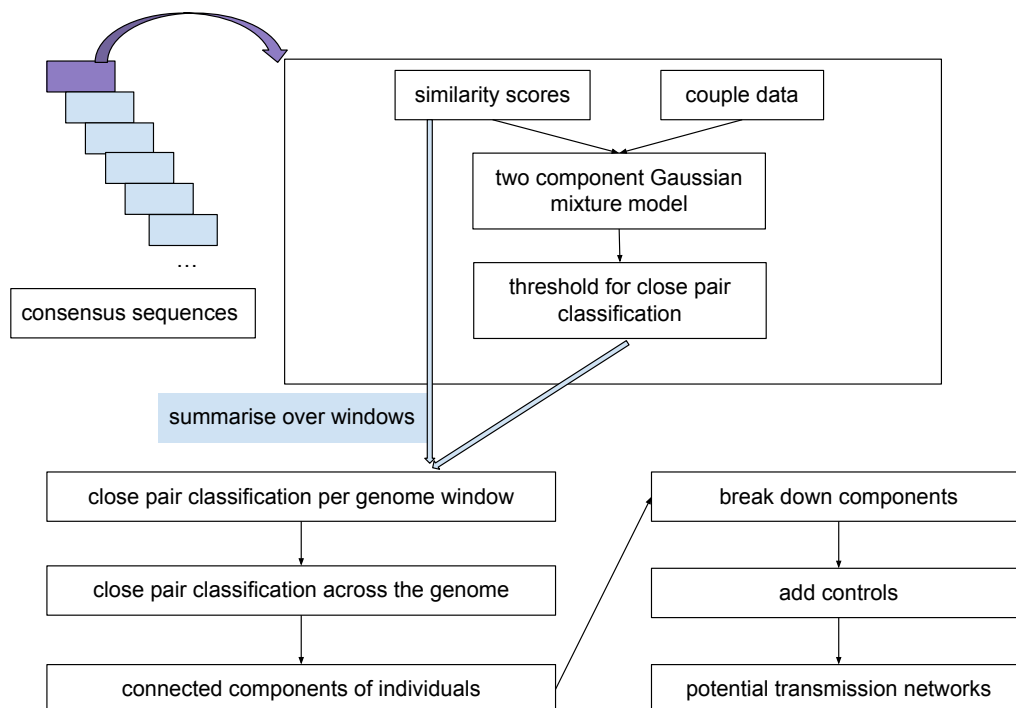


Figure 3.2 **Sketch of making potential networks.** Potential networks were identified through consensus sequences. These sequences were divided into overlapping windows. In each genome window, we evaluated similarity scores, characterised a Gaussian mixture model through the scores among couples, and extracted the threshold for identifying close pairs. Summary of pairwise relationships across the genome provides the closeness between two individuals. If we consider a graph composed of individuals as nodes and the close relationships as edges, individuals were divided into several connected components. Large components were further divided in order to be processed by phyloscanner and controls were added in case any close pairs were missing.

3.2 Potential transmission networks

In this Section, we produce the potential transmission networks using the consensus sequences. The procedure is summarised in Figure 3.2, and the details will be explained in the Subsections.

3.2.1 Similarity scores between consensus sequences across windows

In principle, all analyses were performed on overlapping windows across the HIV genome as is common practice in phyloscanner, because different parts of the HIV genome are subject to different mutation rates, different sequencing success, and most important subject

to recombination, so it often does not make much sense to compare entire HIV consensus sequences. In particular, we divided the whole genome into 99 overlapping windows of length 500 base pair (bp), $[1, 500]$, $[101, 600]$, $\dots$, $[9801, 10257]$.

The first step is to introduce a measure describing the similarity between two consensus sequences in any given, pre-defined genome window, say nucleotide position 1,000 to 1,500 in the HIV Human immunodeficiency virus type 1 group M subtype B (HXB2) reference genome. The software *shiver* (Wymant et al., 2018) generated de-novo references that have the correspondence with the HXB2, and we compared the de-novo reference and the HXB2 to figure out the alignment of the consensus sequences. Upon the aligned consensus sequences, we used a simple measure to quantify how similar two sequences are at each position. The measure is similar to the raw genetic distance (Paradis et al., 2004) and Hamming distance (Hamming, 1950). Suppose the nucleotide of the sequence j and sequence k at position i is respectively $X_{i,j}$ and $X_{i,k}$, the similarity score $S_{i,jk}$ on the position i for the pair of sequences j and k is defined by

$$S_{i,jk} = \begin{cases} 1 & X_{i,j} = X_{i,k} \text{ and } X_{i,k} \in \{a, g, c, t\} \\ \frac{1}{2} & \begin{aligned} &X_{i,j} = r, X_{i,k} = a/g \text{ or } X_{i,j} = y, X_{i,k} = c/t \text{ or } X_{i,j} = w, X_{i,k} = a/t \\ &\text{or } X_{i,j} = m, X_{i,k} = a/c \text{ or } X_{i,j} = s, X_{i,k} = g/c \text{ or } X_{i,j} = k, X_{i,k} = g/t \end{aligned} \\ \frac{1}{3} & \begin{aligned} &X_{i,j} = h, X_{i,k} = a/c/t \text{ or } X_{i,j} = d, X_{i,k} = a/g/t \\ &\text{or } X_{i,j} = b, X_{i,k} = g/c/t \text{ or } X_{i,j} = v, X_{i,k} = a/c/g \end{aligned} \\ 0 & \text{otherwise.} \end{cases} \quad (3.1)$$

Nucleotide here is represented by the universally accepted notation, the International Union of Pure and Applied Chemistry (IUPAC) (Cornish-Bowden, 1985). The score at a location is defined if neither of the two sequences at that position has unknown nucleotide or gaps inserted by the alignment software. To define the similarity scores in a genomic window, the most straightforward approach is to average similarity scores across all the positions in the window where the scores are well-defined. That is, the similarity score for a pair of sequences j and k on the window starting at position w in the HXB2 reference genome is defined as

$$\tilde{S}_{w,jk} = \frac{1}{L_{w,jk}} \sum_{\substack{w \leq i \leq w+499 \\ i: S_{i,jk} \text{ is defined}}} S_{i,jk}, \quad (3.2)$$

$$L_{w,jk} = |\{i : w \leq i \leq w+499 \text{ and } S_{i,jk} \text{ is defined}\}|,$$

where $w = 1, 101, \dots, 9801$. $\tilde{S}_{w,jk}$ is considered if $L_{w,jk} \geq 250$. The reason is that the limited number of positions cannot reliably inform the similarity measure. In addition, too many gaps and insertions in the consensus sequences suggest too many gaps in the deep sequences, and we may construct the unreliable phylogenetic trees in the further analysis. For the same reason, we applied additional stringent selection criteria on $\tilde{S}_{w,jk}$,

1. consider the sequences with read depths > 30 on at least 250bp genome positions in one genome window.
2. consider the pair of sequences if the scores that are available on ≥ 4 genome windows, and as a result, both sequences in the pair satisfy the first criterion in at least 4 common windows.

in order to ensure the pairs identified in the potential network satisfy the requirement that we can build the phylogenies on a sufficient number of windows.

3.2.2 Training data for classifying pairs of individuals into close and not close

We were motivated by Ratmann et al. (2019) and investigated the similarity scores from 828 couples in the Rakai Community Cohort Study (RCCS) that were visited between rounds 1 and 17, and in which both partners were HIV positive and virus were deep-sequenced. It is much easier to characterise phylogenetic relationships between infected individuals with a known, long-term sexual relationship than among individuals between whom nothing about sexual contacts is known. Indeed, the phylogenetic distance of HIV sequences from couples tends to be either very short or very long (Eshleman et al., 2011), because in a couple it is highly unlikely that one partner passed on the infection to someone else, who then passed it back to the other partner (possibly via other intermediaries), rather than direct transmission between the two partners. Small phylogenetic distances then indicate transmission between the two partners, and long phylogenetic distances indicate that either both individuals were infected prior to partnership formation, or that they had other partners that are themselves epidemiologically unconnected. These circumstances lead, at the population level, to a bimodal distribution of phylogenetic distances among couples, and this pattern is exploited here to define phylogenetic cut-off thresholds. Figure 3.3 illustrates the histograms of logarithm similarity scores for five equidistant windows, starting from genome position 1, 2, 001, 4, 001, 6, 001, 8, 001 respectively, and all the distributions have two modes. It also

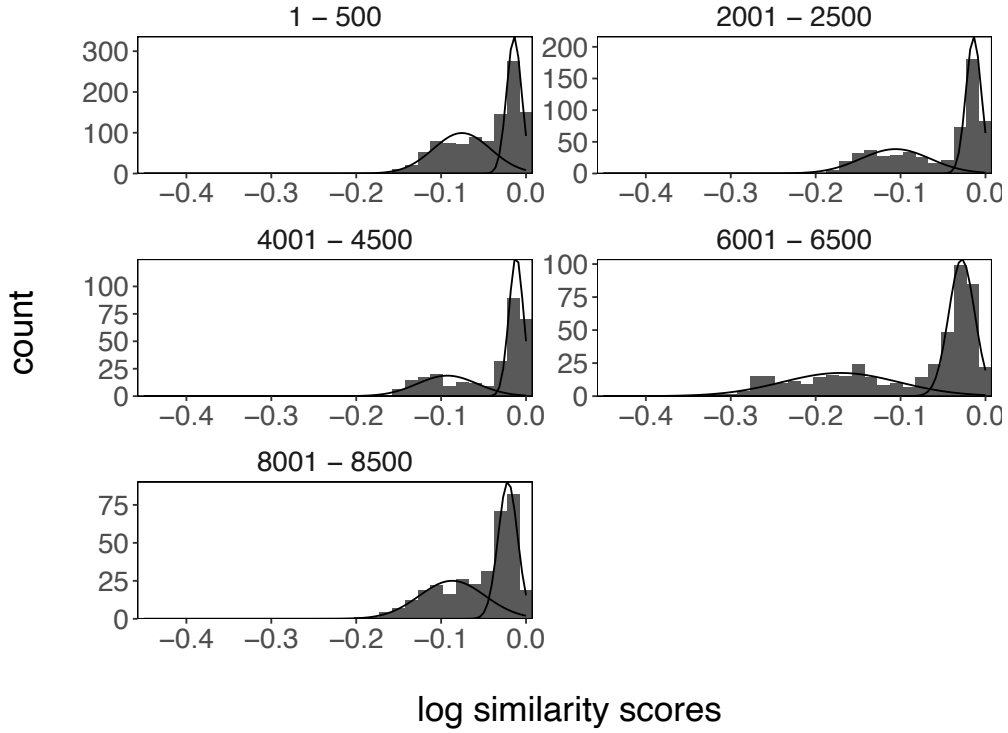


Figure 3.3 **Histograms of log similarity scores among couples.** The five panels show the histograms of log similarity scores among couples in five equidistant windows. The start and end of the windows are specified in the sub-titles. Each panel shows the bi-modal pattern of the similarity scores across the couples. The high peaks around 0 represent that a majority of individuals acquire infections from its partner, and the low peaks indicate some individuals are infected outside the couple relationship. The mixture model in Section 3.2.3 was used to fit the two-component mixture distributions, and the fitted distributions are shown in the black lines.

demonstrates the varying scores in the five genomic windows, and the reason may be the variation in HIV evolutionary rates across the genome (Alizon and Fraser, 2013).

We compared distributions of similarity scores (3.2) in the training data between Ratmann et al. (2019) (rounds 15-16) and our data (rounds 15-18) along the entire genome. Figure 3.4 reveals the weaker bimodal signal during August 2011 and May 2018 than that during August 2011 and January 2015. Ratmann et al. (2019) included 509 scores from the couples, while we have 747 additional scores. About one-third scores are from the couples whose samples have already been collected during rounds 15-16, and the rest belong to new couples. We first focused on the sequences from the existing couples in Ratmann et al. (2019) and compared the distributions of scores by sampling time < 2015.1 . We spotted few differences in the distributions, except for the slightly reduced similarities in 270 new sequences in Figure 3.5

(left). This is likely due to virus evolution at a certain rate and higher divergences between sequences that were sampled later. Therefore, the differences between scores in Figure 3.4 are not explained by the sampling time of existing couples in Ratmann et al. (2019). Next we compared 747 scores of new samples by whether the two hosts have ever been sequenced during rounds 15 and 16. In contrast to Figure 3.5 (left), the two distributions appear distinctly different. It is evident that samples from newly sequenced couples have stronger similarity and do not appear bi-modal. Though we are uncertain about the reasons, the contrasting behaviours of sequences of new couples and existing couples make the bi-modal signal weaker.

3.2.3 Classifier: Gaussian mixture model on log similarity scores

In this Section, we describe modelling log similarity scores of distant and close pairs with a two-component mixture model,

$$\begin{aligned}
 p(\log(\tilde{S}_{w,jk})) &= \phi_w \text{Normal}(\log(\tilde{S}_{w,jk}); \mu_{w1}, \sigma_{w1}^2) \\
 &\quad + (1 - \phi_w) \text{Normal}(\log(\tilde{S}_{w,jk}); \mu_{w2}, \sigma_{w2}^2) \\
 \mu_{w1}, \mu_{w2} &\sim \text{Normal}(0, 1) \\
 \sigma_{w1}, \sigma_{w2} &\sim \text{Exponential}(5) \\
 \phi_w &\sim \text{Beta}(5, 5),
 \end{aligned} \tag{3.3}$$

where $\tilde{S}_{w,jk}$ was defined in (3.2), μ_{w1}, μ_{w2} denote the means of normal distributions for the two components, σ_{w1}, σ_{w2} denote their standard deviations, and ϕ_w represents the probability that observations are from the first component, i.e. the component for distant pairs. The likelihood describes the log score $\log(\tilde{S}_{w,jk})$ in the window w is drawn from the distant component characterised by a Gaussian distribution with mean μ_{w1} and variance σ_{w1}^2 with the probability ϕ_w , and from the close components, $\text{Normal}(\mu_{w2}, \sigma_{w2}^2)$, with the probability $1 - \phi_w$. It is typical to constrain the order of the component means $\mu_{w1} \leq \mu_{w2}$ in the mixture model to make the prior non-exchangeable and make the model identifiable (Betancourt, 2017c). When two components are not well separated, it is recommended to use an informative prior to improve the identifiability. The priors we chose are based on the observations in Figure 3.3. They centre the component means at 1, constrain the component standard deviations within the reasonable 95% CI (0.005 - 0.738) and restrict weights within the 95% CI (0.212 - 0.788).

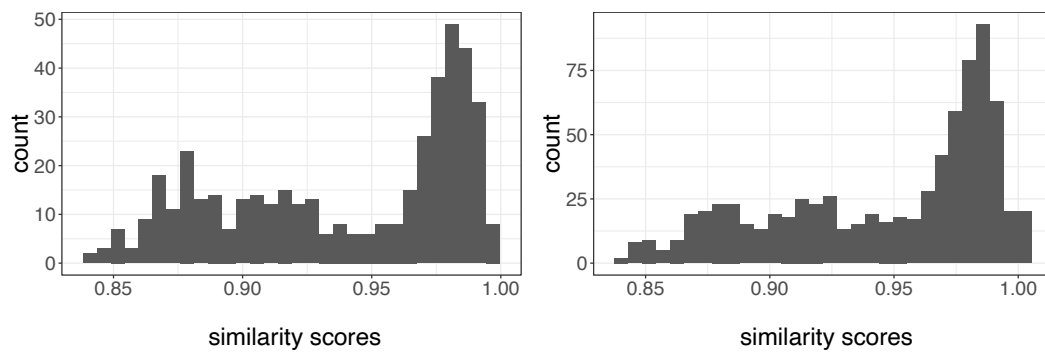


Figure 3.4 **Comparison of score distributions among couples using the sequences that were collected prior to January 2015 and all the sequences.** We evaluated the similarity scores from couples over the entire genome, either using sequences collected prior to January 2015 or all the sequences. The scores are shown in the histograms for the samples that were collected before January 2015 in the left and those between August 2011 to 2019 in the right.

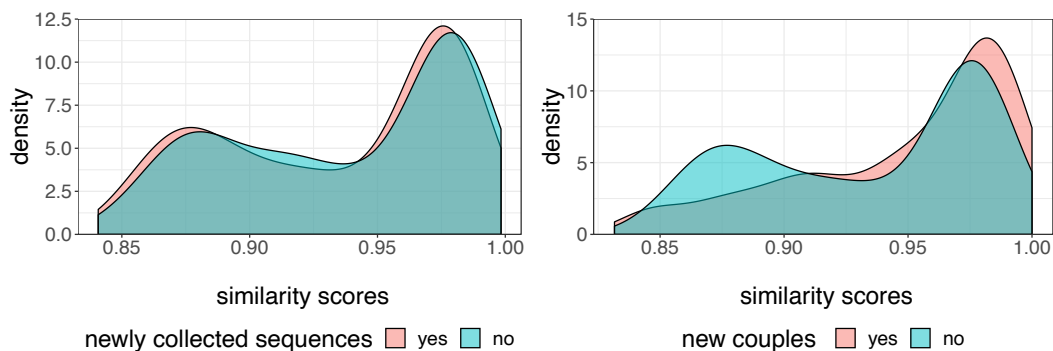


Figure 3.5 **Decomposition of score distributions among the couples.** We first focus on the couples that were analysed by Ratmann et al. (2019) in the left, and compare the density of similarity scores using the sequences that were collected before January 2015 in blue and after January 2015 in red. We next analyse the new samples, and visualise the density of the similarity scores by whether the couple was involved before January 2015 (in blue) or not (in red) in the right.

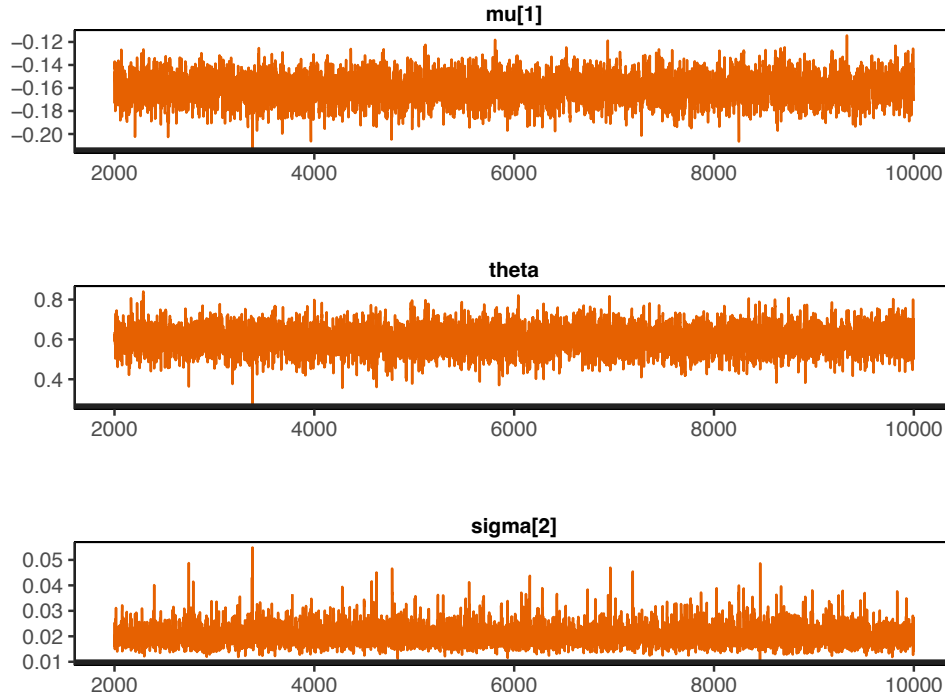


Figure 3.6 **Trace plots for the three parameters with the lowest ESSs in the 68th window.** Stan’s dynamic HMC was run for 10^4 iterations on the scores in each genome window. The minimum ESS across the windows was reached in the 68th window. Shown are trace plots for the three parameters with the lowest ESSs after the burn-ins, in order to present the chains clearer and mitigate the effect of initialisation.

Numerical inferences were made in Stan 2.21.0 using Stan’s no-U-turn sampler (NUTS) Hamiltonian Monte Carlo (HMC) algorithm (Carpenter et al., 2017). All the chains converge well ($\hat{R} < 1.01$), and effective sample sizes (ESS) are over 1,000 out of 8,000 iterations after the warm-up period. In specific, the minimum ESSs of parameters range from 1397.31 to 3499.73 across the windows. Here we show the trace plot where the minimum ESS takes the value 1397.31. Three parameters with the lowest ESSs are presented in Figure 3.6, and the chains mix and converge well.

To visualise the goodness of fit, Figure 3.3 compares the histograms and the fitted densities and illustrates the model (3.3) delivers a reasonable performance on fitting the log scores. In addition, it also indicates adequate sequence pairs are available to characterise the bi-modal pattern in all the five windows. Therefore, no additional information is needed from the neighbouring windows to enhance the signal.

We discussed the weaker bi-modal pattern of the new data than those during rounds 15-16 in Section 3.2.2, and are concerned about the misuse of bi-modal patterns. Thus, we fitted



Figure 3.7 Comparison of the estimated mixture distribution parameters in the whole genome using the sequences that were collected before and after January 2015. Estimated means (top panels) and standard deviations (bottom) of two components (the lower component in the left and the upper component in the right) are shown along the genome positions (in the x-axis). Estimates are described by the median estimates in the lines and the 95% CIs in the shaded areas.

the model (3.3) to the consensus sequences that were collected during 2011.8 - 2015.1 and 2011.8 - 2018.5, and plotted the estimated component means (top) and standard deviations (bottom) by sample collection time (in colour). Comparison between the estimates by sampling time suggests the two credibility intervals (CI) depicted in the shaded areas roughly match. Therefore, even if the bi-modal signal becomes weaker using samples collected in 2011.8 - 2018.5, we are still able to identify the two components as previous done.

3.2.4 Applying the classifier to similarity scores from all possible pairs of individuals

Next, we determine the threshold that separates close pairs from distant pairs in each window, based on the parameters of the fitted model (3.3) in Figure 3.7 (blue). The upper component of the mixture model corresponds to the similarity scores of phylogenetically close couples, between whom transmission is very likely to have occurred. We thus consider the probability that a phylogenetically close pair in which transmission likely occurred has a similarity score above the 5% quantile, where the false positive error should be approximately 5% if the model is reasonably accurate. This interpretation motivates us to use the 5% quantile of the upper component as the threshold which avoids the exclusion of too many close pairs.

Figure 3.8 visualises the thresholds along the genomic windows. We investigated the properties of the estimated thresholds. First, the cutoffs reflect the evolution rates across the genome. In particular, Figure 3.9 shows two distance measures against the estimated cutoffs in each genome window. The measures include the average similarity score (3.2) across all the pairs in the left panel, and the median tree-based distance which was calculated via HIV compendium alignments from Los Alamos National Laboratory (LANL) in 2015 in the right panel (Foley et al., 2015) and stored in the phyloscanner repository (Wymant et al., 2017). We observed the thresholds are positively correlated with the average similarity scores, and negatively correlated with the tree-based distances. This is expected given that the thresholds were determined by similarity scores, and similarity and distance are opposite measures. Second, we explored the role that the threshold plays beyond the couples. To do this, we show the histogram of scores across all the possible pairs in the first genome window in Figure 3.10 (left), and focus on the scores around the threshold in the right panel. In contrast with couples, We do not observe the bi-modal pattern near the threshold among all the pairs. This reflects much more noises in the whole population, and suggests we are not able to prove the direct transmissions among them. The finding is consistent with Ratmann et al. (2019). For this reason, the pairs we identified might be indirect and through intermediates. For example, we are not able to exclude female-to-female transmissions that are unlikely in HIV (Chan et al., 2014), and interpret these pairs as transmissions via unobserved intermediates. Third, the thresholds do not fluctuate substantially across the genome and further address the reasonable performance of (3.3) which models windows independently. Last, we show the proportion of pairs classified by the threshold as close across the genome. Approximately half of couples are consistently classified as close. However, the values are much lower and more variable among other pairs.

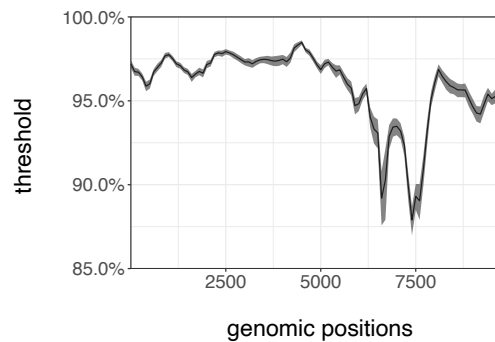


Figure 3.8 **Cutoffs for viral linkage across the genome.** We considered the 5% quantile of the upper component in the mixture model to be the threshold, which separated distant pairs and close pairs. The thresholds are shown along with the genome positions in the line, and the 95% CIs are in the shaded area.

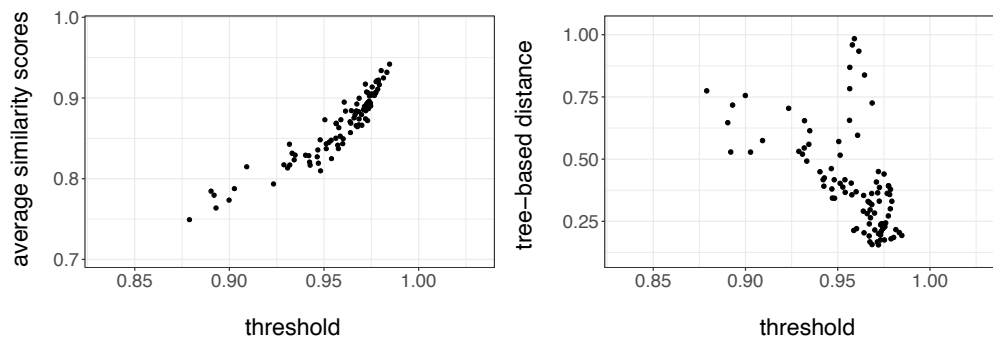


Figure 3.9 **Relationship between cutoffs and distances.** Average similarity scores (3.2) and tree-based distances in each window are shown against the cutoff values in the left and right panels respectively. Each point represents the distance measure and the threshold in a genome window.

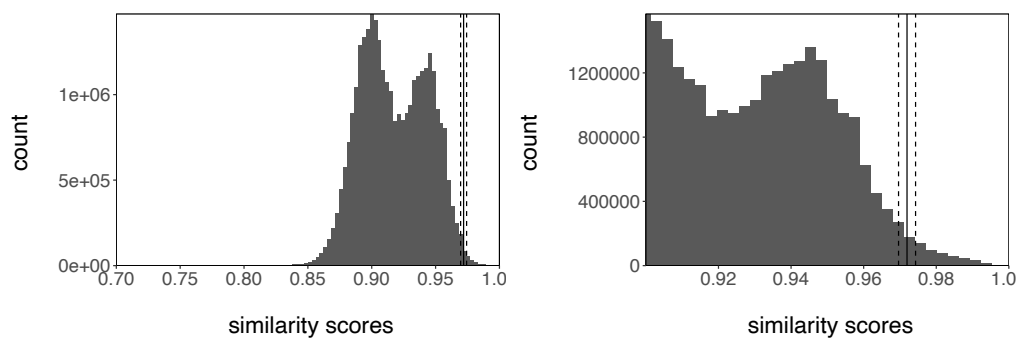


Figure 3.10 **Histogram of similarity scores among all the pairwise combinations of individuals in the first window.** We calculated the similarity scores for all the pairwise combinations in the first window, and show the histogram in the left panel. We then focus on the region around the threshold in the right panel. The vertical black lines refer to the threshold on the first window that was depicted in Figure 3.8, and the dashed lines show the 95% CI of the estimated threshold.

3.2.5 Considering classifications for all windows

We then aggregated the information across the windows and obtained one classification per sequence pair. To do this, we evaluated the proportion of windows indicating the close relationship for each sequence pair. We explore the distribution of the proportion among couples (left) and all the pairwise combinations (right) in Figure 3.12. We find a strong bi-modal distribution among couples with clear peaks at 0% and 100%, but there is a grey area in between. A similar pattern is observed across all the pairs except the increasing moderately distant pairs with the proportion between 0% and 25%.

We attempted three criteria: if (A) one window (B) at least 50% windows (C) at least 80% windows suggest the sequence pair is close, then the pair is considered to be close. To assess the performance, we show the score distributions by the classification upon three criteria and couple status. The distribution is characterised by the median (darkest line), the interquartile range (dark shaded area) and the 95% CI (light shaded area). The window-specific thresholds in Figure 3.8 are added (black line) as a reference. Figure 3.13 (left) indicates the 95% CIs of close and distant pairs are roughly separated, and the reference line is in between, while the CIs are separated to a varying extent in Figure 3.13 (right). The two interquartile ranges substantially overlap when A is applied, and the reference line is above the CIs, which makes no sense. Once B and C are applied, the two 95% CIs are well separated, and the threshold is between two medians. Even though the separation is the most obvious using C, B provides the separation as well as the left panels (couples). In the rest of the Chapter, we focus on C, but we will try B as a sensitivity analysis beyond the thesis. B may be better at accommodating recombinations and thus resulting in more conservative networks, but may require more computational resources.

A few individuals might be deep sequenced at multiple time points, and the sequences at different times are subject to sequencing errors, contamination, or recombination. If at least one sequence pair from two individuals is close, we consider the individual pair to be close. This makes the analysis more conservative than summarising sequences across the time per individual pair. Specifically, 998 individuals have more than one consensus sequences, and as a result, 5,455,821 pairs of individuals have at least two consensus sequences, of which 5,454,901 (99.98%) are consistently supported for the distant relationship, 532 (0.0097%) are consistently supported for the close relationship, and the remaining 388 (0.0071%) provide mixed evidence (i.e. there are some sequence pairs that suggest close relationships, and others that do not).

Consequently, we identified 2525 close individuals in the RCCS and explored their characteristics. In terms of sampling dates, both individuals in 1669 (66.10%) pairs were

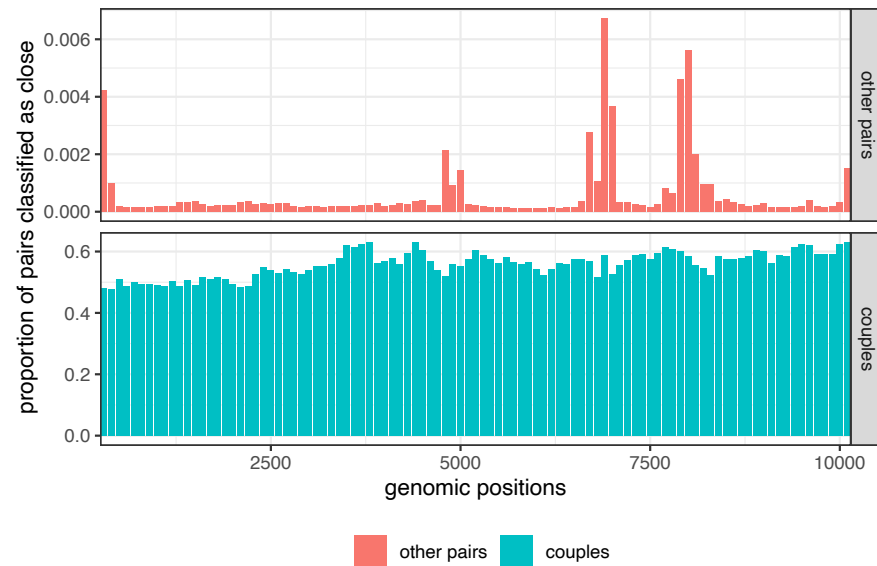


Figure 3.11 **Proportion of pairs classified as close along the genome.** We evaluated the proportion of pairs classified as close in each genome window, using the thresholds in Figure 3.8. This figure shows the proportions by couples (in blue, bottom) and all the other pairwise combinations (in red, top).

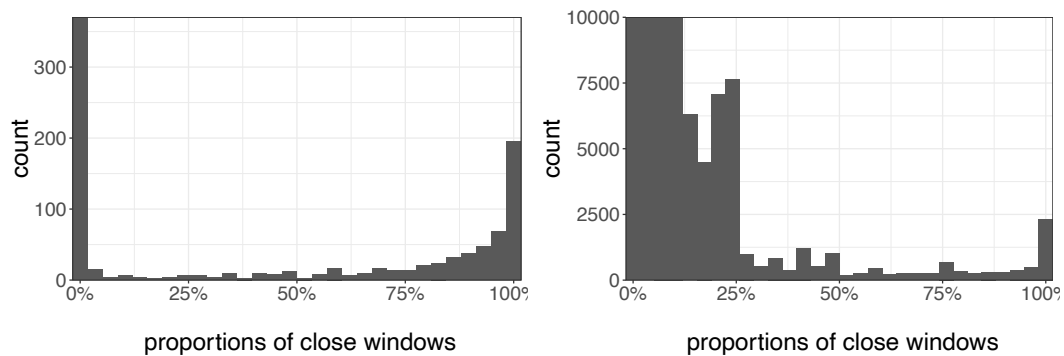


Figure 3.12 **Histogram of the proportion of windows that support the close relationship.** We classified one pair of consensus sequences on each genome window by the proportion of windows supporting the close relationship. We show the histogram of the proportions for the couples in the left panel and all the pairwise combinations in the right panel. Note that loads of individuals (over 3×10^7) have no windows classified as close among all the samples, and we thus cut the y-axis to 10^4 for the clear visualisation of other parts of the diagram.

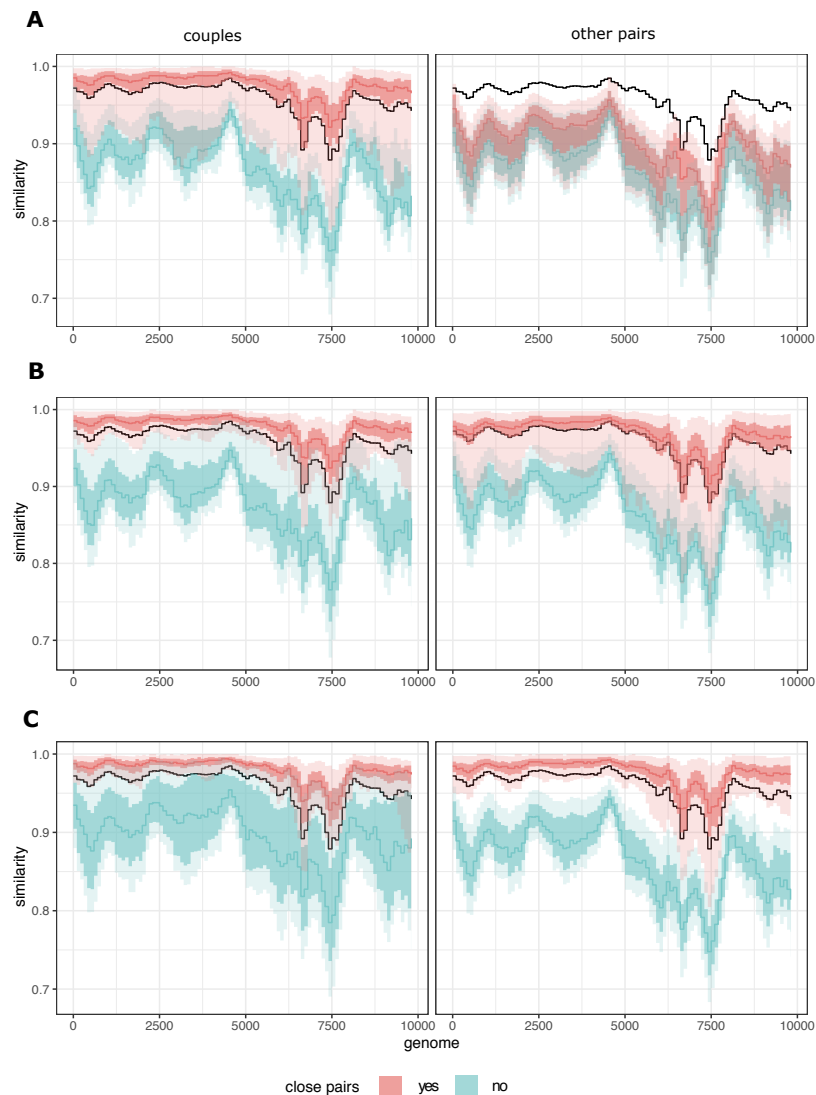


Figure 3.13 The score distributions by the couple status and the classification for the close pairs along the genome windows. We aggregated the shreds of evidence across the windows to classify an individual pair into close and distant. Here we show the score distributions by the classified relationship and the couple status, under the three standards. The classifications are presented in colour, distant in blue and close in red. The couple status is shown in the panels, couples on the left and other pairs on the right. The three standards are in the top, middle and bottom panels respectively. If at least (A) one (B) 50% (C) 80% windows indicate the close relationship for the two hosts, the pair is considered as close. The score distribution is specified by the median (the darkest line), the interquartile range (the dark shaded region) and the 95% CI (the light shaded region). The black line represents the estimated window-specific threshold in Figure 3.8.

Table 3.1 Size distributions of identified networks.

Cluster size	2	3	4	5	6	7	8	9	10	14	291
Count	546	129	45	31	16	9	6	2	2	1	1

sampled between August 2011 and January 2015, those in 257 (10.18%) pairs were sampled between January 2015 and May 2018, and the other 599 (23.72%) pairs have one sampled before January 2015 and another after January 2015. As for gender, 1367 (54.14%) pairs are heterosexual, 676 (26.77%) are within women, 433 (17.15%) are within men, and 49 (1.94%) involve at least one individual with unknown genders. Concerning the number of sequence pairs per individual, 1040 have multiple consensus sequences from at least one individual, of which 632 (60.77%) have the consistent evidence towards the close relationship, while the remaining 408 (39.23%) individual pairs are supported by the mixed evidence.

3.2.6 Building potential transmission networks

Close pairs indicate the closeness between individuals, which prompts us to group the close individuals into networks using a single linkage criterion, i. e. grouping individuals with at least one common person into the same network. The close individuals were then naturally divided into several potential transmission networks. The size distribution of the networks is shown in Table 3.1. Most networks have less than 5 individuals, while there is a large cluster of size 291. We visualise the largest components in Figure 3.14 and spot densely connected nodes in the centre and loosely connected nodes elsewhere.

3.2.7 Analysing large potential transmission networks

It is challenging and time-consuming to reconstruct deep-sequence phylogenies of more than 10,000 unique reads, and so it is problematic to estimate deep-sequence phylogenies of large networks that involve more than 50 individuals. The observation in Section 3.2.6 allows decomposition of large networks that exceed 50 individuals. The division into groups is based on the network structure measure, modularity. Modularity is defined as the contrast between the edges within groups and the counterpart in an equivalent random network, and measures the relative connectivity within groups in contrast with that between groups (Newman, 2006). We optimised the modularity through one of the standard algorithms for modularity optimisation, the multi-level modularity optimisation algorithm (Blondel et al., 2008). It assigns each individual to a community composed of itself and re-assigns one individual to another community if the change improves modularity. Next, the procedure

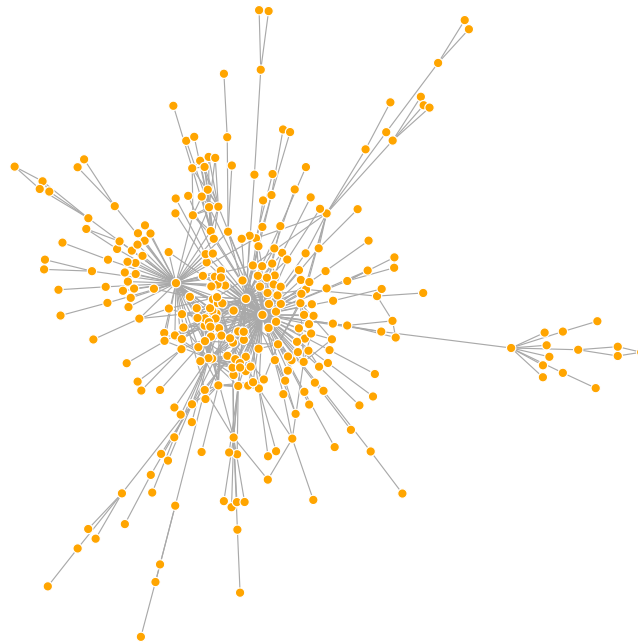


Figure 3.14 **Visualisation of the largest connected component.** We identified close pairs by analysing the genetic distance between consensus sequences. The close relationship performs as links and breaks individuals into connected components. Here we show the largest connected component in Table 3.1. The yellow nodes represent individuals, and the black edges indicate the consensus sequences from two individuals are close. There are 291 individuals in the largest component and it is difficult to process them in one go in the further analysis.

repeats on the community level instead of node level, until no improvement of modularity can be achieved. Figure 3.15 depicts the decomposition of the largest network into several sub-networks through the modularity optimisation algorithm, and shows the sub-networks are connected through red links. We referred to the ends of the red edges as bridging individuals, as they acted as "bridges" between sub-networks. We grouped the bridging individuals into another network in case the interaction between sub-networks was missing. Where necessary, we repeated this strategy recursively until all the networks were of at most the pre-determined

Table 3.2 Size distributions of networks after application of the multi-level modularity optimisation algorithm.

Cluster size	2	3	4	5	6	7	8	9	10	11	13	14
Count	546	129	46	33	16	9	6	3	4	1	1	2
Cluster size	15	16	17	18	21	22	26	37	38	45	50	55
Count	1	1	1	1	1	1	1	1	1	1	1	1

target size. In this Chapter, we applied the strategy twice and the resulting size distribution is in Table 3.2.

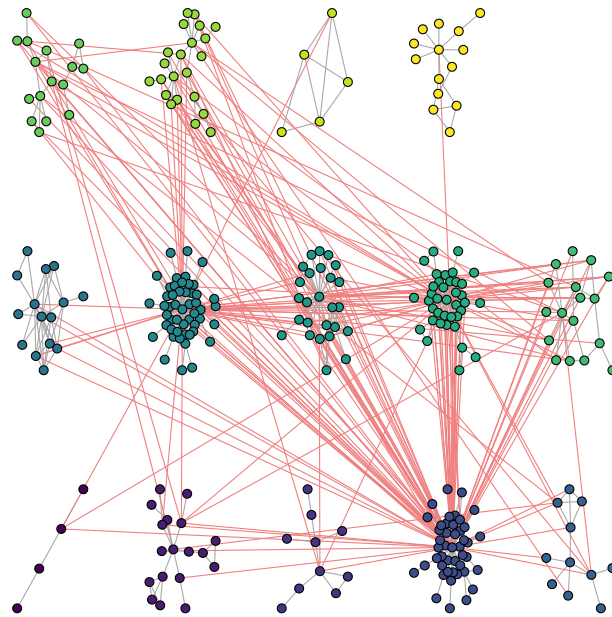


Figure 3.15 **Decomposition of the largest connected component.** We identified the structure of the largest cluster and decomposed it into 14 densely connected clusters. The clusters are displayed by the blocks of individuals that are represented by the nodes in the same colour. The grey edge represents the within-group link and the red edge shows the link between two clusters. The two ends of the red link are called bridging individuals which connect the clusters. The division of the largest component requires a separate cluster composed of all the bridging individuals, to avoid missing the strongly linked individuals between clusters.

3.2.8 Adding controls to potential transmission networks

In the end of this Section, we added controls to each potential network to ensure the robustness and quality of the analysis. In particular, we identified and added three individuals that had the maximum similarity to all the individuals in a network on average and were inferred not to be close to anyone in the network. We also added the known partner in couples as controls, because we were always interested in the phylogenetic relationship of couples. In addition, in order to reduce the number of phyloscanner runs, we merged small-size clusters to one phyloscanner run composed of up to 8 individuals.

In summary, this Section starts from the consensus sequences, and assigns individuals to potential transmission networks based upon the closeness. These networks make the subsequent phyloscanner analysis computationally feasible. The computations in this Section were paralleled on Imperial High performance computing and took approximately one day.

3.3 Deep sequence alignments and phylogeny reconstruction

In this Section, we perform deep sequence analysis on each of the potential transmission networks constructed in Section 3.2, and generate phylogenetic trees in each network on each genomic window. The main workflow is summarised in Figure 3.16.

3.3.1 Background sequences

Inclusion of additional standard sequences, with the name of background sequences, is a common practice in phylogenetic analysis, and it aids the interpretation of sequences of interests, e.g. the deep sequences in the RCCS, and to root the phylogeny. For example, the false links can be inferred in the phylogeography without additional background sequences (Bbosa et al., 2019). We used two sets of whole-genome sequences as background sequences. The first set was provided by the LANL and included standard HIV-1 subtype consensus sequences of non-recombinants. Several Circulating Recombinant Form (CRF) only have two complete sequences from which consensus sequences cannot be made, and one of the CRF sequences was chosen under this circumstance. In addition, we included complete or near-complete genomes from Kenya, Uganda and Tanzania. In total, the background

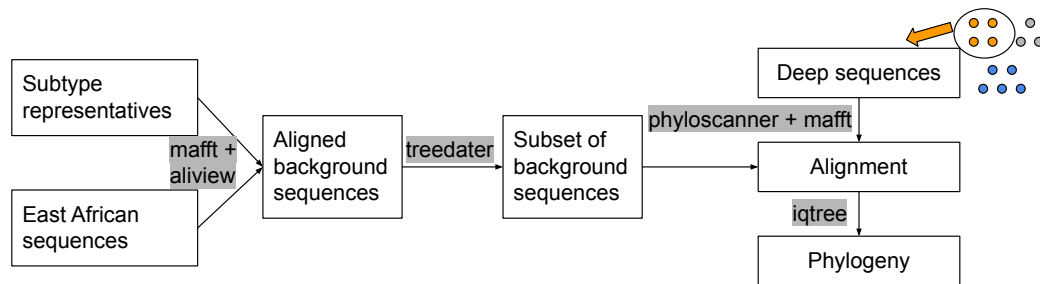


Figure 3.16 **Sketch of deep sequence phylogenetic analysis.** We started from the deep sequences in one potential network and added the background sequences to each network. Background sequences included subtype representatives and east African whole genome sequences, and were aligned to HXB2. The hundreds of background sequences may slow down the phylogenetic analysis, and we thus made a selection while remaining the outcomes unaffected. Then we made deep sequence alignments together with the selected background sequences, and phylogenies were built upon the alignments. The software in all the steps is shown in grey shaded texts.

sequences include 113 representatives of whole-genome HIV subtype or HIV CRFs and 522 further whole-genome sequences from east Africa.

3.3.2 Alignment of background sequences

We aligned all the background sequences from east Africa through MAFFT version 7.475 (Kato and Standley, 2013). Multiple Alignment using Fast Fourier Transform (MAFFT) is a fast program for aligning sequences by considering the evolutionary events. As the alignments will be added to each network for the further analysis, we hope to guarantee its quantity. Specifically, we inspected the alignment and improved it by manual curation. We compared the alignment with HXB2 and moved the nucleotide to make it match with HXB2 in the alignment editor AliView version 1.26 (Larsson, 2014). Additional efforts were invested on the consecutive gaps which suggest the potential failure in alignment. It is worth mentioning that we deleted the insertions to the HXB2 for simplicity. However, a better practice is to not delete any nucleotides. This keeps the sequence variations in sub-Saharan Africa, and potentially contributes to the better alignment of target sequences with respect to background sequences.

3.3.3 Selection of background sequences

The 635 background sequence may increase the computational complexity and slow down the process in all the networks. Therefore, several attempts were made to reduce the number of background sequences while keeping the phylogenies unaffected.

In order to evaluate whether the reconstructed phylogenies are reasonable, we built phylogenies and inferred their time to most recent common ancestor (TMRCA) with `treedater` version 0.5.0 (Volz and Frost, 2017), using the alignment in one potential transmission network on one genome window. Faria et al. (2014) suggests the time of HIV introduction to humans was in the 1920s, and thus we consider the earlier TMRCA estimates impossible and as mistakes in the phylogeny construction. We undertook the assessment on the first potential transmission network, and show the estimated TMRCAs over the genomic windows in red lines. We observed fluctuations in the estimates over the windows, and for the interests of the overall pattern, we smoothed the estimates using local weighted regression (LOESS).

Moreover, it is possible that the phylogenetic analysis failed because of mistakes in the alignment. Alignment of sequences was completed by inserting gaps between nucleotides so that identical and similar characters would align in successive genomic positions. The alignment software may struggle to align the nucleotide to the correct position, resulting in consecutive gaps. As a result, we identified six consecutive positions where more than 80% were filled by gaps, which we called the "chunk of gaps", excluded the windows where chunks of gaps appeared and repeated the TMRCA analysis using the remaining genome windows.

We next assess if estimates of TMRCA can better align with known estimates of HIV introduction in humans based on four different subsets of background sequences, (1) Set1: A/D associated subtype representatives, (2) Set2: all the subtype representatives, (3) Set3: A/D associated subtype representatives and east African sequences, (4) all the subtype representatives and east African sequences. We display the smoothed TMRCA estimations against genome windows in Figure 3.17, where we use all the windows (left) or exclude the windows where potential alignment issues arise (right). As illustrated in both panels, TMRCAs were estimated to be in the 1950s when all the background sequences were used. None of the subsets produced the same curve as the whole set, and Set 1 gave the most biased estimates of TMRCA. This suggests that we cannot exclude sequences from east Africa or sequences of subtypes other than the most prevalent A and D.

We then investigated the effect of using the whole set of subtype representatives and subsets of 522 east African sequences in the TMRCA analysis, in particular, 50, 100, 200, 300, 400, 500 random samples from the east African sequences. An analogy of Figure 3.17

is shown in Figure 3.18 for this investigation. It illustrates that at least 200 random samples of east African sequences are adequate to generate the almost same estimates as the whole set in both panels.

3.3.4 Generating deep-sequence alignments

In this Section, we align deep sequences in each potential network together with the background sequence alignments. Sequence alignment is to match sequences as much as possible for further analysis of their dissimilarity as a result of the evolutionary process as in Section 3.3.2. The overall strategy is to automate the alignment procedure through phyloscanner and MAFFT, and spare no efforts on manually improving alignment quality, taking the enormous alignment tasks into account.

To quantify the alignment quality, we extracted the length of the largest chunk of gaps because the chunk of gaps is an indicator of mistakes in sequence alignment (misalignment) aforementioned in Section 3.3.3. To obtain the quantity, we calculated the proportion of gaps at each position, compared the proportion to a reference value, say 95%, to determine whether one position mainly consisted of gaps and counted the number of successive gaps-dominated positions. As inspired by Section 3.3.3, another quality measure is the TMRCA estimates, as the misalignment is likely to pose problems to the phylogenetic trees and potentially produce the unrealistic TMRCA estimates.

We investigated the relationship between the two indicators. The length of the largest chunk of gaps and the TMRCA estimate were extracted from each alignment on a window in a potential network. We plot the TMRCA against the length for all the windows in one of the largest networks, the 404th network, in Figure 3.19 (right). The negative correlation suggests the alignment with a large chunk of gaps potentially has an unreasonably early TMRCA estimate. In addition, long branch lengths were perceived to reflect erroneous sequence data in the phylogenetic analysis (Mai and Mirarab, 2018). We thus plot this known measure against length of the largest gap chunk, and observed a positive correlation. This shows the length indicator reflects the errors in the alignment. Also, the length is negatively correlated, but not perfectly correlated, with the TMRCA, and we thus consider both indicators for the complementary purpose.

Following the definition of alignment quality indicators, we attempted several strategies to improve the quality and examined the effect on two networks, the first network with $< 1,000$ taxa and the 404th network with > 2000 taxa. Generally, we applied phyloscanner to make deep sequence alignments, which makes use of MAFFT. To guarantee the alignment

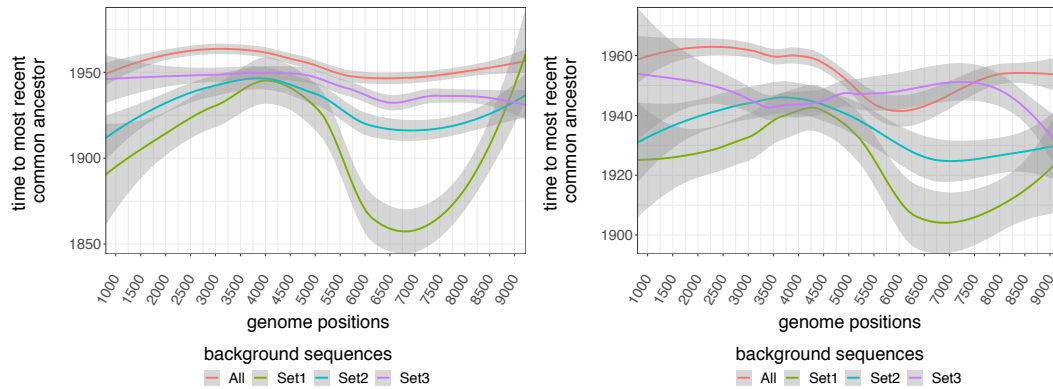


Figure 3.17 Estimated TMRCAs using four subsets of background sequences. We estimated the TMRCa from the reconstructed phylogeny in each genome window for the first potential transmission network. The estimates were smoothed along the genome. The smoothed estimates and the 95% confidence intervals are shown in the coloured lines and grey shaded areas. The colour represents the use of different subsets of background sequences in phylogeny reconstruction. The left panel shows the impact of different types of background sequences and the right panel shows the impact when the chunks of gaps were deleted.

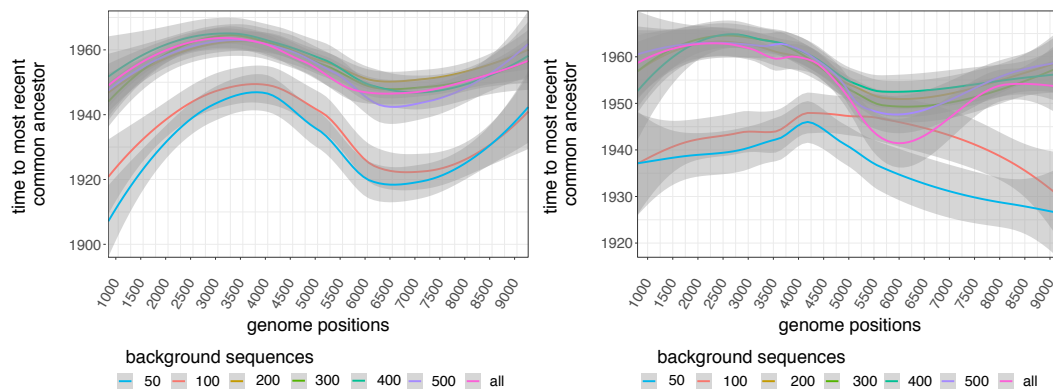


Figure 3.18 Estimated TMRCAs using the downsampled subsets of background sequences. We estimated the TMRCa from the reconstructed phylogeny in each genome window for the first potential transmission network. The estimates were smoothed along the genome. The smoothed estimates and the 95% confidence intervals are shown in the coloured lines and grey shaded areas. The colour represents the subset of downsampled background sequences in phylogeny reconstruction. The left panel shows the impact of down-sampled background sequences, and the right panel shows the impact when the chunks of gaps were deleted.

quality, we opted for a slow but accurate algorithm in MAFFT, iterative refinement method using weighted sum-of-pairs which optimise the goodness of alignment and consistency scores which accommodate the discrepancy between the multiple alignments and pairwise alignments. There are several algorithms for the pairwise alignment tailored for different alignment features. We assumed the entire genome can be globally aligned without a large chunk of unaligned sequences in the boundary, and thus specified the G-INS-i option. Figure 3.20 (left) is an example of a subset of aligned sequences in the window between 5,850 and 6,099 and presents mistakes in the alignment. The challenge might be that massive insertions with respect to HXB2 makes software struggle to shift nucleotides to the right position. The first strategy we explored is to change arguments in the software. For instance, gap opening penalty and gap extension penalty represent the cost for opening or extending the gaps respectively. Increasing the two arguments may reduce the gaps in the alignment. We increased the gap opening penalty from the default value of 1.53 to several values between 2 and 10. Alternatively, we brought up the gap extension penalty from the default value of 0 to 0.123. None of these parameter specifications makes Figure 3.20 (left) better aligned. Second, we tested if deleting gap chunks from the alignment leads to the improvements in TMRCA estimates. We attempted two approaches (1) deletion of all the gaps-dominated positions (2) deletion of the chunk of gaps only. Figure 3.21 illustrates the TMRCA estimates when adopting either strategies, and we spotted no significant differences in the TMRCA estimates with no deletion. Third, we performed the alignment by using exactly the same algorithm in MAFFT, G-INS-i, on all the aligned sequences. Figure 3.20 (right) presents the improved alignment after the realignment where the gaps become less. Apart from this single example, we examined the two quality indicators in the potential networks (the first and 404th) in Figure 3.22 and Figure 3.23 with and without the realignment. They demonstrated the realignment strategy (red boxplots) effectively makes the largest chunk of gaps shrink in Figure 3.22, and reduces the early TMRCA estimates below the dash black lines in Figure 3.23.

The previous quality check mainly focused on the first and 404th potential transmission network. Next, we apply the two metrics to check the alignment quality for all the networks. First, we summarised the metrics in the two networks, and compared their metrics with all the other networks. The largest gap chunks have length 1 (0 – 12) and 2 (0 – 33) respectively in the first and 404th network across the genomic windows. The 95% CI of the medians across the networks is 0 (1 – 2.8), and does not substantially deviate from the medians in the two networks, with the values of 1 and 2. In addition, the TMRCA in the two networks are 1956.70 (1854.14 – 1985.63) and 1953.84 (1866.69 – 1986.31)

respectively across the genomic windows. The 95% CI of the median TMRCAs across the networks is 1957.45 (1951.91 – 1960.10), and does not largely deviate from 1956.70 and 1953.84.

Apart from the medians, we evaluated the ranges, the minimum and maximum, of the two measures across the windows, and draw the histograms of these summary statistics based on all the potential transmission networks. We compared the histograms and the two measures calculated from the first and 404th network depicted in the dashed lines. Figure 3.24 shows most of the networks have similar alignment qualities as the two networks we intensively studied. However, there are some windows that have either considerable gap chunks and very early estimates of ancestor ages. For instance, the runs 61 and 221 around the position 8,850, and the 6th run around the position 9,175 have the largest gaps. We inspected these runs and found there are massive insertions with respect to the HXB2 sequence in these alignments. Also, we observed very early root ages in the windows starting from roughly 5,900 in runs 12, 22 and 339. The background sequences around 5,900 were quite noisy, and the target sequences in RCCS were generally well aligned. Overall, we do not identify major alignment problems. Deep sequence alignment was paralleled at Imperial High performance computing and took one week in our case.

3.3.5 Reconstructing deep-sequence phylogenies

In this Section, we discuss how to build the phylogenetic tree from the deep sequence alignment. A phylogenetic tree is a branching diagram built from the alignments, and reflects the evolutionary relationships between the organisms. RAxML (Stamatakis, 2014) and IQTREE (Nguyen et al., 2015) are both fast software to infer the maximum likelihood phylogeny for a large data set. We updated the phylogeny construction software from RAxML (Ratmann et al., 2019) to IQTREE version 2.0.3. RAxML models the rate heterogeneity over sites by a gamma model (Yang, 1994), while IQTREE relaxes the assumption of Gamma distributed evolutionary rates and provides an option for a free rate model (Yang, 1995). Dr. Matthew Hall found that this fixes the previous problem that estimates of phylogenetic branch lengths increase alignment size. Therefore, we disabled the default option in phyloscanner using RAxML, and opted for IQTREE. It takes about one week to reconstruct deep sequence phylogenies.

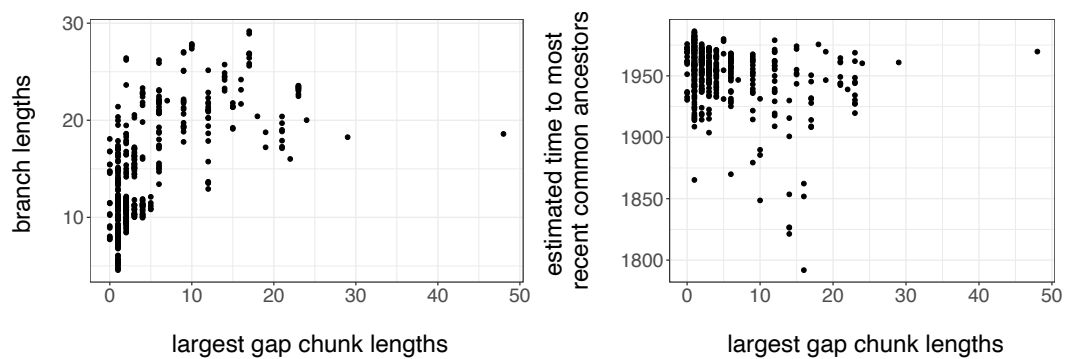


Figure 3.19 Measures of alignment quality. We aligned the sequences, built the phylogenetic tree, and inferred the time of common ancestry in the trees in one of the largest potential networks (404th). We focused on two measures of alignment quality, the length of the largest gap chunks in the alignment and the estimated time of common ancestry. The long branch length is a common indicator for the mistakes in the phylogeny reconstruction. We first show the branch length against one measure of inferior alignment quality on the left, and then display the relationship between the two measures. Each point represents quality measures of the alignment in one genome window.

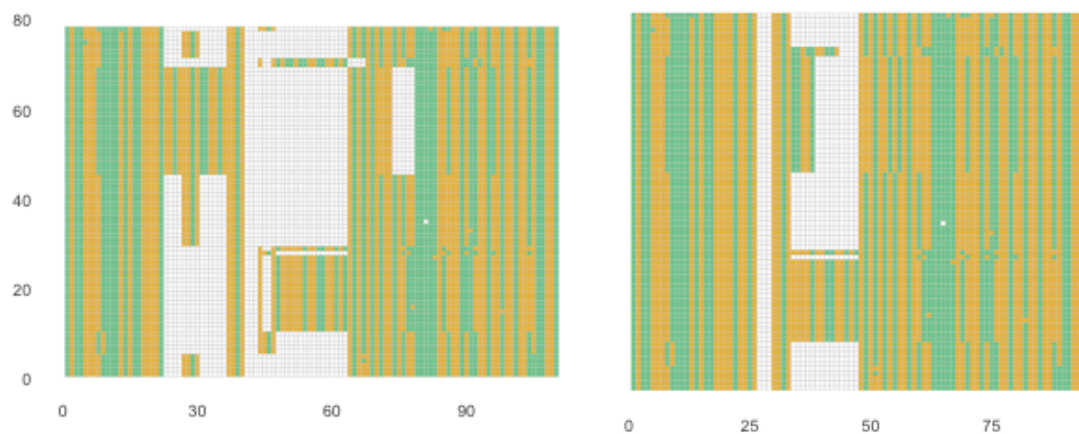


Figure 3.20 Example of misalignment and the proposed solution. We applied the first part of phyloscanner to generate alignment for the sequences in the 404th potential network. Here we took the genome window between 5850 and 6099 as an example. We visualise the subset of the aligned sequences in the left panel and observe too many insertions and blocks of sequences in the incorrect positions. We proposed to re-align the sequences in the left using MAFFT, and this produced the alignment in the right which has the enhanced alignment quality.

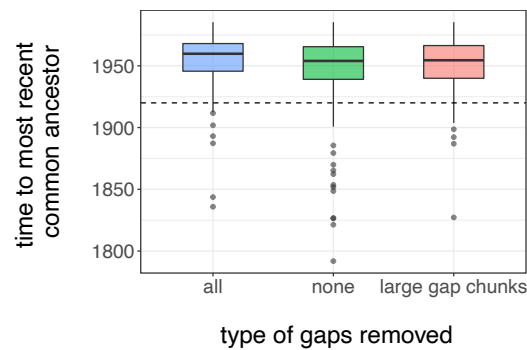


Figure 3.21 **Effect of gap removal on the estimated time of the most recent common ancestry.** We used phyloscanner and specified the MAFFT option as "iterative refinement methods using weighted sum-of-pairs and consistency scores" to align the sequences. The resulting alignments sometimes have a large chunk of gaps, especially when we consider the relatively large potential transmission networks. We removed the gap chunks, rebuilt the trees, and estimated the age of common ancestry in one of the largest networks, the 404th network. Then we compare the estimates in boxplots, when (1) all the positions with $\geq 95\%$ gaps were removed, abbreviated as "all" (2) no positions were deleted, abbreviated as "none" (3) only large gap chunks taking at least 6 positions were removed, abbreviated as "large gap chunks".

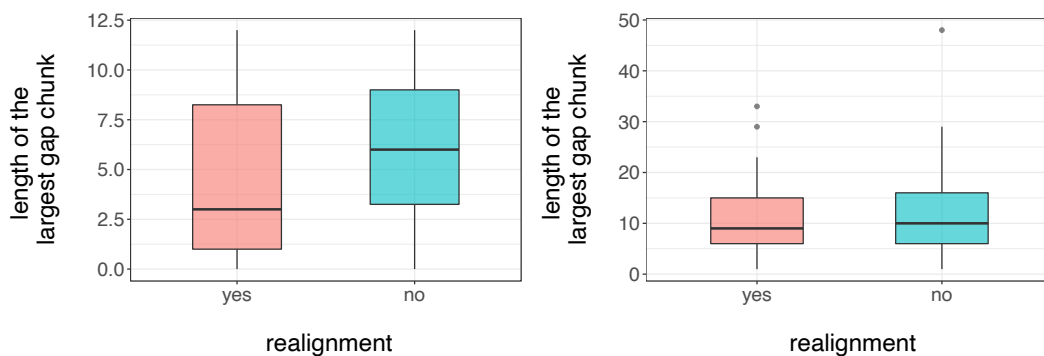


Figure 3.22 **The effect of realignment strategy on the length of the largest gap chunk.** We aligned the sequences from the first and 404th potential transmission networks in phyloscanner, and show the length of the largest chunk of gaps in the box plots (in blue). To mitigate the problem in Figure 3.20, we applied the realignment to the phyloscanner output and plot the same type of box plots in red. The left shows the comparison for the 1st potential network which includes $< 1,000$ taxa, and the right is for the 404th potential network which includes $> 2,000$ taxa.

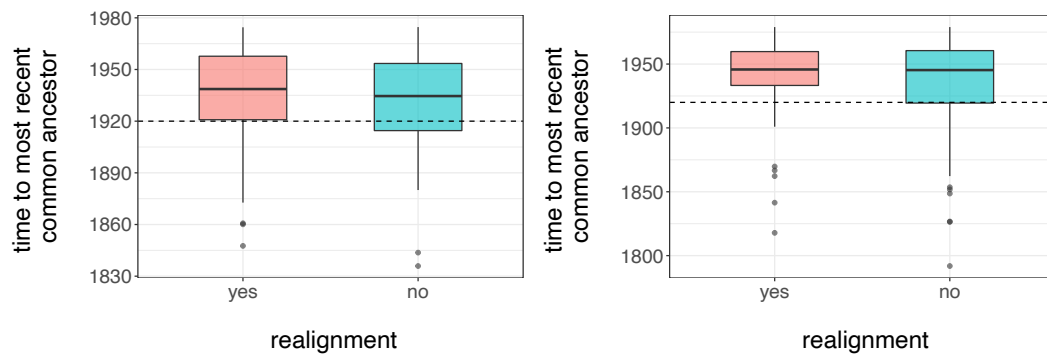


Figure 3.23 The effect of realignment strategy on the estimated time of common ancestry. We aligned the sequences from the first and 404th potential transmission networks in phyloscanner, built the phylogeny and show the time of common ancestry in the box plots (in blue). To mitigate the problem in Figure 3.20, we applied the realignment to the phyloscanner output and plot the same type of box plots in red. The left shows the comparison for the 1st potential network which includes < 1,000 taxa, and the right is for the 404th potential network which includes > 2,000 taxa. The dashed black lines act as the reference and represent the known time of HIV passed to humans.

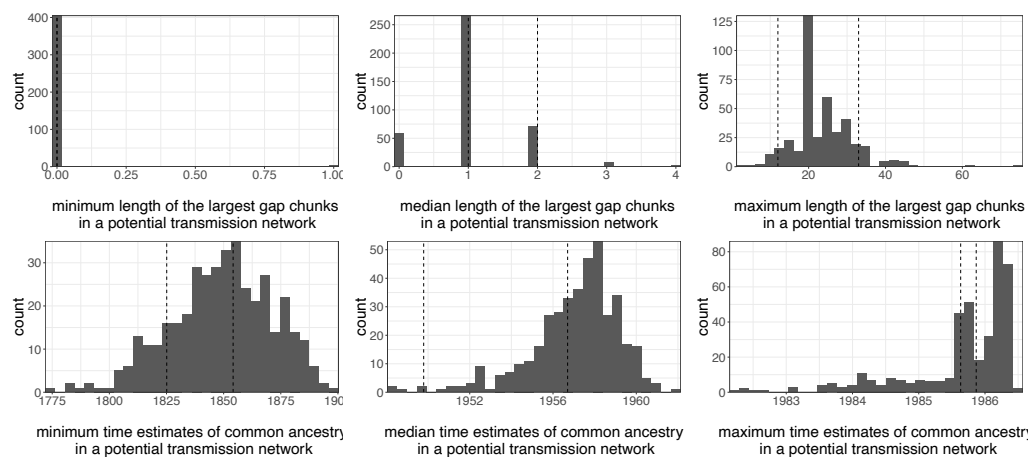


Figure 3.24 Histograms of summarised quality measures across the potential networks. We calculated the minimum (left), median (middle) and maximum (right) of the length of the largest gap chunks (on the top) and the estimated time of common ancestry (in the bottom) for each potential network, and plot the histograms. The reference lines are provided by the corresponding quantities in the first and 404th potential transmission networks, and depicted in the black dashed lines.

3.4 Identifying source-recipient pairs

In this Section, we analyse the deep sequence phylogenies in 3.3 in order to reconstruct the source-recipient pairs of transmissions.

3.4.1 Rules for deriving pairwise phylogenetic relationships between individuals in deep-sequence phylogenetic trees

We extract information from the deep-sequence phylogenies in order to reconstruct the source-recipient pairs. This topic is thoroughly discussed in the previous works (Hall et al., 2021; Ratmann et al., 2019; Wymant et al., 2017).

First, *phyloscanner* applies the maximum parsimony principle while penalising the within-host diversity, in order to determine the state (colour) of internal nodes. The coloured phylogeny is divided into subgraphs, the contiguous set of nodes that belongs to the same individual. The distance between two subgraphs is defined as the shortest patristic distance (the sum of branch lengths) between any two nodes, of which each is from one subgraph. For two hosts i and j , three quantities can be extracted,

- minimum distance between subgraphs: the minimum patristic distance between subgraphs u from i and v from j , denoted by Δ_{ij} ,
- adjacency: the shortest path between at least one subgraph from i and j is not across other individuals except i and j , denoted by $A_{ij} = 1$,
- path: the number of subgraphs from j which have a subgraph from i as its ancestor, denoted by P_{ij} .

According to the three quantities, we define four types of ancestral relationship between any two hosts i and j in one phylogeny,

- i is ancestral to j , abbreviated as ancestral if $\Delta < \Delta^*$, $A_{ij} = 1$, $P_{ij} \geq 1$ and $P_{ji} = 1$,
- i is descendant to j , abbreviated as descendant if $\Delta < \Delta^*$, $A_{ij} = 1$, $P_{ij} = 0$ and $P_{ji} \geq 1$,
- complex or no ancestry, abbreviated as complex if $\Delta < \Delta^*$, $A_{ij} = 1$, $P_{ij} \geq 1$ and $P_{ji} \geq 1$ or $\Delta < \Delta^*$, $A_{ij} = 1$, $P_{ij} = 0$ and $P_{ji} = 0$,
- not close or not adjacent, abbreviated as unlinked if $\Delta \geq \Delta^*$, or $A_{ij} = 0$,

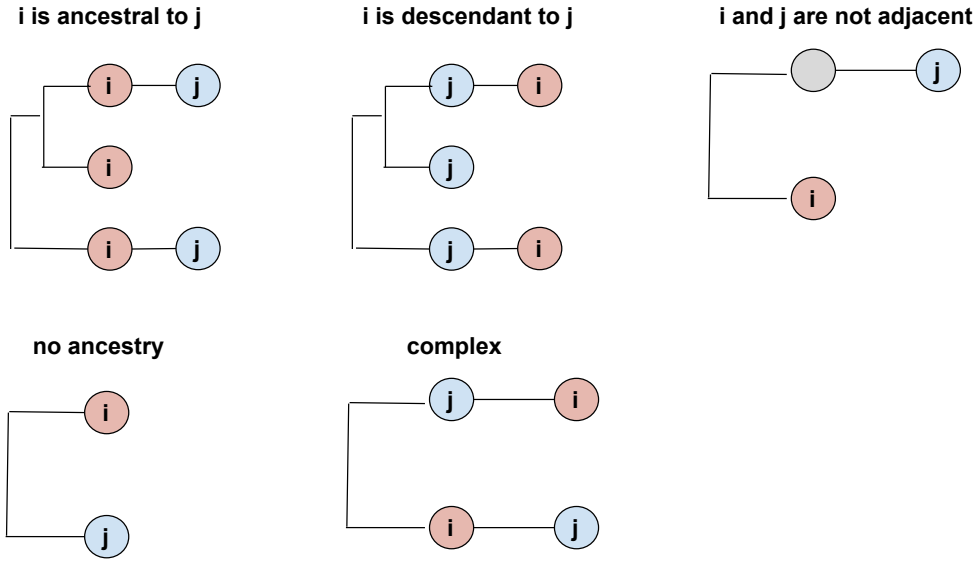


Figure 3.25 **Illustration of four types of relationships.** The top row illustrates i is ancestral to j , i is descendant to j , i and j are not adjacent. The bottom row shows the complex or no ancestry relationship. The subgraphs of host i are in red, j in blue, and any other hosts in grey.

where Δ^* is a user-defined threshold for excluding distant pairs. A visual summary of ancestral relationships is in Figure 3.25.

Usually there are more than one phylogenies that include the hosts i and j . Our next step is to summarise the ancestral relationships across the phylogenetic trees by counting the trees indicating each relationship. Assume the number of trees that suggested the pairwise relationship between i and j to be ancestral, descendant, complex and unlinked is

$$\mathbf{k}^{\text{pr}} = (k_a^{\text{pr}}, k_d^{\text{pr}}, k_c^{\text{pr}}, k_u^{\text{pr}}), \quad (3.4)$$

and the total number of trees that involve i and j is then

$$k^+ = \sum_{i \in \{a, d, c, u\}} k_i^{\text{pr}}. \quad (3.5)$$

To determine the probability of each pairwise relationship π^{pr} , a simple approach is to model the counts by a multinomial distribution,

$$\mathbf{k}^{\text{pr}} \sim \text{Multinomial}(k^+, \pi^{\text{pr}}), \quad (3.6)$$

and the maximum likelihood estimate (MLE) is then

$$\hat{\pi}_i^{\text{pr}} = \frac{k_i^{\text{pr}}}{k^+}. \quad (3.7)$$

However, (3.6) makes the assumption that all the trees are independent. This assumption does not hold because the genome windows are overlapping in phyloscanner analysis. Ratmann et al. (2019) tackled this problem by adjusting the counts k_i^{pr} for overlap in the alignment. The adjusted counts are then modelled by a real-valued extension of multinomial distribution.

We then ascertain the transmission pairs and their directions upon $\hat{\pi}^{\text{pr}}$. Define the linkage score, the probability that i and j are phylogenetically close, as

$$\hat{\pi}_l = \hat{\pi}_a + \hat{\pi}_d + \hat{\pi}_u, \quad (3.8)$$

where the individual indices i and j are omitted for the simplicity in illustration. We next apply a hard threshold to $\hat{\pi}_l$. That is, if $\hat{\pi}_l$ is above the pre-defined threshold π_l^* , the hosts i and j are considered to be phylogenetically close, and we will investigate their transmission direction. In the literature, two thresholds are used, 0.5 by Hall et al. (2021) and 0.6 by Ratmann et al. (2019).

If the individual i and j are linked, the directed score is defined to hard threshold the transmission direction. For example, Hall et al. (2021) defined the directed score as

$$w = \max\{\hat{\pi}_a^{\text{pr}}, \hat{\pi}_d^{\text{pr}}\}, \quad (3.9)$$

and if $w > \zeta = 0.33$, the pair composed of i and j are directed. i is confirmed to transmit to j if $\hat{\pi}_a^{\text{pr}} > 0.33$, and similarly j transmits to i if $\hat{\pi}_d^{\text{pr}} > 0.33$. Another definition is provided by Ratmann et al. (2019),

$$w = \max\left\{\frac{\hat{\pi}_a^{\text{pr}}}{\hat{\pi}_a^{\text{pr}} + \hat{\pi}_d^{\text{pr}}}, \frac{\hat{\pi}_d^{\text{pr}}}{\hat{\pi}_a^{\text{pr}} + \hat{\pi}_d^{\text{pr}}}\right\}, \quad (3.10)$$

and the threshold is set to $\zeta = 0.6$. We refer to the two definitions as Matthew's rule (M) and Oliver's rule (O) in the rest of the Chapter, and assess the impact of two definitions.

So far, we have resolved the transmission pairs, and these pairs form transmission networks. In the network, there are at most three types of edges, each indicating the evidence for either transmission direction or no directions resolved. A problem with the networks is the existence of cycles and non-unique sources per individual. To address the issue, phyloscanner implements Edmonds algorithm (Edmonds, 1965) to maximise the weights on the edges and identify the most likely directed graph, i.e. transmission chain.

3.4.2 Down-sampling of reads in case of read imbalance

Phyloscanner has the option to exclude some tips from the phylogeny before reconstructing ancestral states, and the inferred ancestral relationship (3.4) may depend on the option choices in `phyloscanner`. In this Section, we investigate the impact of one option on the outcome. To quantify the accuracy, we introduce the seroconversion pairs. If two individuals i and j are a couple, phylogenetically linked, the first positive date of i is no later than the last negative date of j , and j turns positive then, we confirm from the clinical data that i transmits to j and i and j are called seroconversion pairs.

The option we explored is the downsampling. Downsampling means the uneven read counts potentially bias the inference of transmission directions (Purvis and Agapow, 2002) and one may want to remove some tips to make reads more balanced. In the RCCS data, we show the window counts for one individual pair whose reads are imbalanced (read differences > 100) in Figure 3.26, and observe the varying situations across the pairs. `phyloscanner` provides the downsampling option that removes the entire unique reads at random from one phylogeny to a pre-defined value. In our experiment, we assumed it was more likely to infer the heavily sampled individuals as ancestors, performed the `phyloscanner` analysis with and without downsampling, counted the frequency of an individual with more reads classified as the ancestor in both scenarios and evaluated the ratio. If the ratio is < 1 , the downsampling is effective in reducing bias of ancestor inference towards the heavily sampled individuals. The ratio is 0.8907 (0.8844 - 0.8919) among the seroconversion pairs, 0.9106 (0.9080 - 0.9131) among the couples, and 0.9303 (0.9289 - 0.9321) across all the pairs. This experiment validates our assumption and suggests it is useful to use downsampling to mitigate the bias introduced by uneven reads.

Furthermore, we investigated the effect of down-sampling on the accuracy of ancestral relationship inference, across the pairs with known transmission directions. We compared the estimated transmission directions with the known directions by the clinical data when down-sampling was or was not applied. In specific, there were 66 seroconversion pairs in Ratmann et al. (2019), and we assessed the accuracy of the estimated transmission direction using all the linked pairs. We repeated the down-sampling for 10 times to take randomness of down-sampling into account. The results are summarised in Table 3.3. The impact of down-sampling on accuracy depends on the rules. In the case of Oliver's rule, down-sampling reduces the pairs of which the direction cannot be resolved, increases the correct inference of transmission directions, but increases the incorrect inference as well. However, down-sampling reduces the incorrect inference, improves the estimated directions that are consistent with the clinical evidence and decreases the unclear transmission directions under

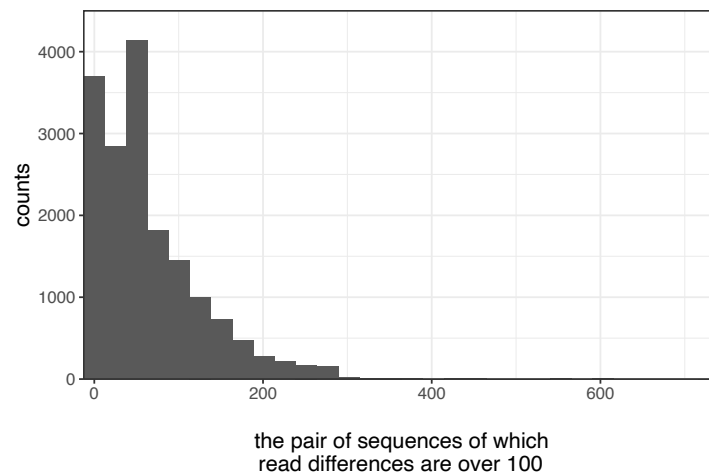


Figure 3.26 **Histogram of the number of sequence pairs of which the differences in reads are greater than 100.** We gathered the pair of sequences with read differences > 100 across the phyloscanner runs, counted the number for each pair of individuals, and show it in the histogram.

Table 3.3 Inference accuracy on the seroconversion pairs when down-sampling is applied or not.

down-sampling	rule	consistent	inconsistent	unclear	total
no	O	19	3	14	36
yes	O	21 (19.23 - 21)	5 (4.22 - 5)	9.5 (9 - 10.78)	35
no	M	20	5	14	39
yes	M	21 (20 - 22)	4.5 (4 - 5)	12 (12 - 13)	38

Matthew's rule. Overall, the down-sampling is recommended, taking both ratios and the accuracy measures into account.

3.4.3 Handling ancestral subgraphs with no branch length differences

Next, we inspect the phylogenies of pairs inferred to be in the opposite direction to the known direction from clinical data, and observed some misclassified ancestral relationships. The definition of ancestry in phylogenetic subgraphs depends on the paths between subgraphs, which in turn are defined by the ordering of branching events in the deep-sequence phylogeny and the maximum parsimony lineage assignments on the tree topology. The ordering of branching events is defined regardless of the length of the phylogenetic branches, and in some cases, this can be problematic. One example is given in the left subfigure of Figure 3.27,



Figure 3.27 Classification of multiple ancestral relationships. We display the parts of a phylogeny which contain AID3114 (in blue), AID6098 (in pink), AID1311 (in yellow), and AID0719 (in red) in the genome window 1,250bp-1,499bp. Tips indicate unique read fragments, and the number in the tip label reflects the copy number.

which shows one subgraph from AID3114 in blue, which is inferred to be ancestral to multiple subgraphs from individual AID6098 in red. This inference is problematic because the branch lengths between some subgraphs of both individuals are (essentially) zero. Thus, while the multifurcating root is assigned to the individual shown in blue, the underlying topological ordering is arbitrary.

We developed new options to handle such situations. First, we classified a pair as complex if the number of paths from one host to another is greater than a pre-specified value. This compromises the existing options. The reason for the new option is that a large number of paths may reflect more intermingled relationships between hosts. However, the option did not work, and a counterexample is illustrated in Figure 3.27. In particular, we visualise a phylogeny that includes AID3114, AID6098, AID1311 and AID0719 in the genome window between 1250bp and 1499bp. In this phylogeny, we excluded the individuals with less than 30 reads after the topology analysis and did not apply the down-sampling. In the parts of the phylogeny, we observed 4 paths from AID3114 to AID6098 and 5 paths from AID1311 to AID0719. However, the phylogeny reflects the intermingled relationship between AID3114 and AID6098 and the ancestral relationship between AID1311 and AID0719.

An alternative approach is to distinguish the two cases in Figure 3.27. Importantly, we want to identify the case in Figure 3.27 (left) where the most recent common ancestor of tips from AID3114 has more than two children, which is called a multifurcation. This indicates the order of these children and the ancestor cannot be determined, and the two hosts are more intermingled. This motivates our alternative solution.

The starting point is that we observe multiple paths from host i to host j and no paths from j to i . The new option we proposed is as follows. First, we focus on the subgraphs of the potential ancestral host i and identify the ancestors in the subgraphs. Next, we follow the same procedure for the possible descendant host j . The question is whether order of the subgraphs from the two hosts can be determined. To tackle the problem, we calculate the branch lengths between the ancestors in the subgraphs from the two hosts. Any $< 10^{-5}$ branch lengths suggest the branch lengths are essentially 0, since the minimum branch length generated by IQTREE is $< 10^{-5}$. If so, the two ancestors are in the same multifurcation, and we classify i and j as complex. If all the branch lengths are $< 10^{-5}$, we ascertain the ancestral relationship between i and j . This approach works for the two examples in Figure 3.27.

We applied this new option to the seroconversion pairs and compared with the current option in *phyloscanner* that recognised the multiple ancestral relationships as an ancestral relationship. Once again, we concentrate on the pairs that are linked applying either option. The inference accuracy is shown in Table 3.4. We observed a decrease in the transmission directions that can be estimated correctly or incorrectly and a rise in the pairs whose directions cannot be resolved, under Oliver’s rule (3.10) when downsampling was applied. However, we observed no changes when downsampling was not applied. In addition, the findings under Matthew’s rule (3.9) regardless of downsampling were the same as those under Oliver’s rule (3.10) with downsampling. Therefore, our new option managed to fix the misclassifications, though the correct inference was also brought down.

As the drawback of our modification lies in the failure to recover several directions that the current option can, we further assessed the cost of applying the new option. Table 3.5 shows the changes in the number of pairs that are estimated to be linked or directed, using both options. It reveals the growing number of directed pairs under Oliver’s rule. The potential reason is that our modification generates an increasing number of complex relationships, which makes denominators in (3.10) drop, and in turn amplifies the signal of directionality. However, given the reduction in the correct inference among seroconversion pairs, we advise to be cautious with the rule. In contrast, we observed roughly 100 – 200 reductions in the number of pairs whose direction can be resolved under Matthew’s rule, for the reason that the quantities in (3.9), $\hat{\pi}_a^{\text{pr}}$ and $\hat{\pi}_d^{\text{pr}}$, both decrease.

Table 3.4 Inference accuracy on the seroconversion pairs when zero branch lengths between subgraphs are taken into account or not in determining the pairwise relationship.

down-sampling	zero length adjustment	rule	consistent	inconsistent	unclear	total
no	no	O	26	5	3	34
no	yes	O	25	3	6	34
yes	no	O	24	3	7	34
yes	yes	O	24	3	7	34
no	no	M	23	4	10	37
no	yes	M	20	0	17	37
yes	no	M	24	3	11	38
yes	yes	M	22	1	15	38

Table 3.5 Linked and directed pairs when zero branch lengths between subgraphs are taken into account in determining the pairwise relationship.

down-sampling	zero-length adjustment	rule	linked pairs	directed pairs
no	no	O	928	734
no	yes	O	930	749
yes	no	O	913	681
yes	yes	O	913	716
no	no	M	1055	706
no	yes	M	1057	594
yes	no	M	1040	704
yes	yes	M	1040	524

We then assess the impact of down-sampling, when the new option is applied. We calculated the quantities as in Section 3.4.2, and found the risk ratios of a heavily sampled individual in the pair classified as parents when applying down-sampling are 0.9071 (0.8920 - 0.9149), 0.8907 (0.8844 - 0.8919) and 0.8714 (0.8699 - 0.8728) among seroconversion pairs, couples and all the pairs respectively. Table 3.6 demonstrates the impact of downsampling on the inference accuracy. Although the risk ratios suggest that downsampling mitigates the issue of heavily sampled individuals being inferred as ancestors, the seroconversion pairs suggest no significant changes in the inference accuracy.

3.4.4 Identified transmission networks and source-recipient pairs

On present evidence, we used the adjustment for zero branch lengths between ancestral subgraphs throughout the analysis, and applied the two rules and down-sampling or not as sensitivity analysis. We first reported the transmission networks. We identified 640 and

Table 3.6 Inference accuracy on the seroconversion pairs when down-sampling is applied or not, taking the zero branch lengths between ancestral subgraphs into account.

down-sampling	rule	consistent	inconsistent	unclear	total
no	O	21	4	12	37
yes	O	20 (20 - 21.77)	4 (3.23 - 5)	10 (9.23 - 11)	35
no	M	19	2	18	39
yes	M	19 (18 - 20.55)	2 (1 - 2)	17 (15.45 - 18)	38

Table 3.7 Cluster size distributions of transmission networks.

down-sampling	rule	2	3	4	5	6	7	8
yes	O	441	125	37	22	6	6	3
no	O	431	126	40	23	7	7	3
yes	M	457	120	35	17	6	5	-
no	M	453	120	35	18	8	5	1

637 networks with and without down-sampling respectively under Oliver's rule, while the numbers were always 640 under Matthew's rule. Table 3.7 shows the cluster size distributions in the four scenarios by rules and down-sampling strategies. The maximum cluster size is 8, while most clusters of size are 2 to 3. The networks in Table 3.7 are based on all the sequences collected during August 2011 and May 2018, and are thus not comparable with the previously constructed networks on samples between August 2011 and January 2015 (Ratmann et al., 2019).

Then, we identified the most probable chain spanning all the nodes in the transmission network, and explored the reconstructed transmission pairs involving individuals enrolled between August 2011 and January 2015. Table 3.8 presents the sex composition of these pairs. In general, we identified the substantial impact of the rules (3.9) and (3.10), and Matthew's rule (3.9) generated more conservative results. In contrast, the down-sampling had a relatively marginal effect and resulted in fewer source-recipient pairs.

Last, we compared our experiment using downsampling and Oliver's rule (3.10) (the first row) and the previous experiment (Ratmann et al., 2019) (the last row). We obtained the similar numbers of heterosexual pairs in the two studies. The minor discrepancies may be caused by the introduction of new options. As discussed in Section 3.4.3, the new option may increase the pairs of which the transmission direction cannot be resolved. Furthermore, sequencing was performed again by the Fraser Pathogen Dynamics Group. We may obtain more sequences for one individual pair that appear in rounds 15-16, or sequences may be

Table 3.8 Reconstructed pairs in the most likely transmission chain when zero lengths are adjusted.

down sampling	rule	source-recipient pairs	heterosexual	male-to-female
yes	O	407	299	176
no	O	427	307	185
yes	M	326	238	143
no	M	362	262	154
yes, 2019	O	402	293	176

improved to a higher quality. In addition, as the down-sampling is subject to randomness, we might sample slightly different sets of tips, and the inferred transmission directions may be marginally altered.

3.5 Summary

In this Chapter, we constructed from deep-sequence data collected in Rakai transmission networks with information on the direction of transmission and highly supported source-recipient pairs, that is phylogenetically likely transmission pairs in which a particular direction of transmission is also highly supported. The source attribution is purely established on the phylogenies that are build through deep sequences data, and does not require extra knowledge. In comparison to Ratmann et al. (2019), we included new sequence data collected after January 2015, used a different and computationally less demanding approach for constructing potential transmission networks, and updated several algorithmic rules along the way to identify phylogenetically highly supported source-recipient pairs.

Chapter 4

The non-parametric Gaussian process flow model

4.1 Introduction

In this Chapter, we present the methods for estimating spatial transmission flows at the population level from phylogenetic data on the direction of transmission between two individuals that are geo-located to their place of residence. The phylogenetic data and transmission events with likely direction of transmission were presented in Chapter 3. The Global Positioning System coordinates of the households were recorded in the census, and the household members were enumerated, so we are able to figure out the residence location of individuals involved in a transmission event. The geo-locations were randomised within 2 kilo-meters to ensure anonymity and protect the privacy of study participants. Figure 4.1 visualises a subset of the 293 directed transmission events reconstructed in Ratmann et al. (2019) as arrows on a map where the place of residence is used to geo-locate the two individuals in a phylogenetic transmission event. In this Chapter, we aim at exploring the spatial structure of transmission flows by fitting a Gaussian process (GP) model to the observed directed transmission events.

Gaussian process regression has emerged as a powerful approach for spatial modelling (Gelfand and Schliep, 2016). In particular, it can be used to model the spatial surface for continuous variables such as temperature (Banerjee et al., 2003), or discrete variable such as the abundance or number of cases at a location by replacing the Gaussian likelihood by an exponential family member (Diggle et al., 1998). In addition, multivariate spatial Gaussian processes are applied to multivariate measurements at each location (Gelfand et al., 2003), such as oxygen, carbon monoxide and particulate matter 2.5.

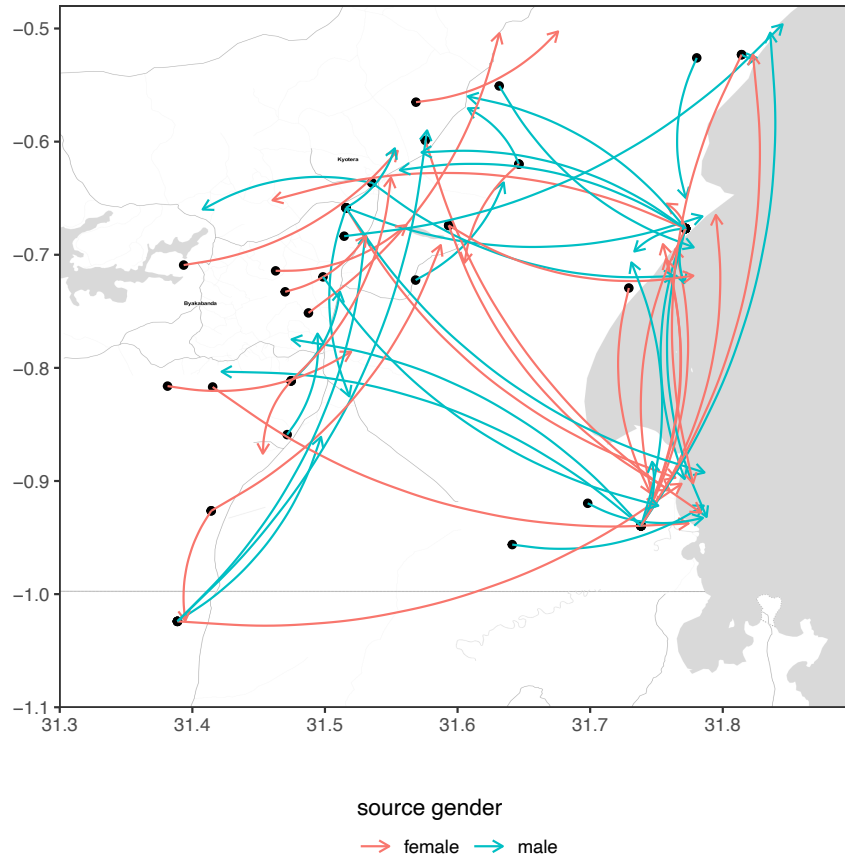


Figure 4.1 **Observed transmission events.** The map shows the reconstructed directed transmission events by the residence geo-locations that are randomised by $\pm 2\text{km}$. Arrow colours represent the genders of sources. There are substantial transmissions along the main road in the north-south direction and between the lake (in grey) and the area away from the coast. However, the general pattern stays unclear.

In the context of spatial epidemiology, Vanhatalo and Vehtari (2007) models high-resolution spatial aggregated deaths associated with cerebral vascular and alcohol by the Poisson likelihood. Its intensity incorporates the relative risk of a disease, on which a Gaussian process prior is imposed over the region. Bhatt et al. (2017) maps the prevalence of *Plasmodium falciparum* in sub-Saharan Africa through the Gaussian processes, where the mean function is accurately informed by state-of-art machine learning methods using location-specific covariates. Ak et al. (2018) predicts the cases of Crimean–Congo hemorrhagic fever over the spatial locations in Turkey. To model the spatial temporal kernel efficiently, the Kronecker product is used for decomposing the spatial and temporal effects. In these examples, Gaussian process and its variations are demonstrated to be effective in tackling the spatial modelling

problems, and estimate epidemiological quantities such as prevalence, cases, and relative risks at a high resolution.

The goal of this Chapter is to predict virus movement at each spatial location. The problem fits well into bivariate Gaussian process regressions and may benefit from spatial interpolation. Indeed, the problem we consider is similar to an emerging topic, Gaussian processes on directional data. Directional data θ , also called circular data, are defined on the unit circle, periodical and have no magnitudes. They are popular in meteorology (wind directions) and oceanography (wave directions). Jona-Lasinio et al. (2012) transforms a linear variable X to be circular, for example by using modulus 2π , i.e. $\theta = X \bmod 2\pi$, and the Gaussian process prior on the linear variable X induces the warped Gaussian process on the circular data θ . An alternative approach (Wang and Gelfand, 2014) introduces a latent radius variable r , and the polar coordinate (θ, r) can be transformed to the Cartesian coordinate (X, Y) . A bivariate Gaussian process prior on (X, Y) induces the projected Gaussian process on θ where the latent variable r is marginalised out. In contrast with the wrapped Gaussian process, the project Gaussian process is through a non-linear transformation, and is thus more flexible, where the circular distribution can be asymmetric and bimodal. Jona Lasinio et al. (2020) develops the first R package for spatial interpolation of circular data through Gaussian processes.

Despite the similarity of concepts with the directional data problems (Jona-Lasinio et al., 2012; Wang and Gelfand, 2014), we have in our application the geo-locations of the household of individuals at the time of community census. Therefore, the changes in latitudes and longitudes of residence associated with a phylogenetic source-recipient pair are both available. As a result, we are able to work on Cartesian coordinates, because the angles and radius are both known. With no need of marginalising angular data analytically, our model does not require a specialised package developed by Jona Lasinio et al. (2020) and can be straightforward integrated with the probabilistic programming language, Stan.

4.2 The basic Gaussian process flow model

We started by introducing the bi-variate Gaussian process in the spatial context. The bi-variate Gaussian process was applied to model the quantity of interests such as the log prevalence y_i at the i th spatial location $x_i = (x_{1,i}, x_{2,i})$. Here our aim is to estimate the viral movement from the i th phylogenetically likely source location $x_i = (x_{1,i}, x_{2,i})$. The viral movement is two dimensional, and defined by either the i th phylogenetically likely recipient

location $z_i = (z_{1,i}, z_{2,i})$ or the change in location from the i th source to the i th recipient,

$$y_i = (y_{1,i}, y_{2,i}) = (z_{1,i}, z_{2,i}) - (x_{1,i}, x_{2,i}), \quad (4.1)$$

where $i = 1, \dots, n^+$ and n^+ is the number of directed pairs. We chose the location changes y_i as they are less dependent on the scale of longitudes and latitudes. The collection of source locations is denoted by $\mathbf{x} = (x_i)_{i=1, \dots, n^+} \in \mathbb{R}^{n^+ \times 2}$, and that of changes in locations is $\mathbf{y} = (y_i)_{i=1, \dots, n^+} \in \mathbb{R}^{n^+ \times 2}$. Several mathematical terms are defined prior to the model development.

- Half-Normal: If $Y = |X|, X \sim \text{Normal}(0, \sigma_X^2)$, then $Y \sim \text{Half-Normal}(0, \sigma_X^2)$.
- Kronecker product: If $\mathbf{A} \in \mathbb{R}^{m \times n}$ and $\mathbf{D} \in \mathbb{R}^{p \times q}$, then the Kronecker product $\mathbf{A} \otimes \mathbf{D}$ is defined as

$$\mathbf{A} \otimes \mathbf{D} = \begin{bmatrix} A_{1,1}\mathbf{D} & \cdots & A_{1,n}\mathbf{D} \\ \vdots & \ddots & \vdots \\ A_{m,1}\mathbf{D} & \cdots & A_{m,n}\mathbf{D} \end{bmatrix}. \quad (4.2)$$

- Hadamard product: If $\mathbf{A} \in \mathbb{R}^{m \times n}$ and $\mathbf{D} \in \mathbb{R}^{m \times n}$, then the Hadamard product $\mathbf{A} \circ \mathbf{D}$ is defined as

$$\mathbf{A} \circ \mathbf{D} = \begin{bmatrix} A_{1,1}D_{1,1} & \cdots & A_{1,n}D_{1,n} \\ \vdots & \ddots & \vdots \\ A_{m,1}D_{m,1} & \cdots & A_{m,n}D_{m,n} \end{bmatrix}. \quad (4.3)$$

- The Cholesky decomposition of a symmetric positive-definite matrix $\mathbf{A}$, $\mathbf{L}_\mathbf{A}$ satisfies $\mathbf{L}_\mathbf{A}\mathbf{L}_\mathbf{A}^T = \mathbf{A}$ where $\mathbf{L}_\mathbf{A}$ is a lower triangular matrix with positive diagonal entries.

We start by modelling the changes in latitudes and longitudes as noisy observations of a non-parametric vector-valued function $\mathbf{f}$ to which we assign a Gaussian process prior,

$$\begin{aligned} \mathbf{y} &\sim \text{Normal}(\mu_x \mathbf{1}_x + \mathbf{f}, \sigma_y^2), \\ x &= \text{latitude, longitude}, \\ \mathbf{f} &\sim \text{GP}(\mathbf{0}, \mathbf{k}), \\ \sigma_y &\sim \text{Half-Normal}(0, 1), \end{aligned} \quad (4.4)$$

where μ_{latitude} and $\mu_{\text{longitude}}$ represent the baseline of changes in latitudes and longitudes, $\mathbf{1}_{\text{latitude}}$ and $\mathbf{1}_{\text{longitude}}$ are indicators of the type of coordinates, σ_y denotes the noise of observations. The likelihood makes the assumption of the similar changes in both coordinates for

the flow starting from the same source locations. The assumption does not usually hold in practice, but it might be useful to detect the overall spatial pattern of transmission events. The covariance matrix of $\mathbf{f}(x_i) \in \mathbb{R}^2$ and $\mathbf{f}(x_j) \in \mathbb{R}^2$ is

$$\mathbf{k}(x_i, x_j) = \begin{bmatrix} k_{1,1}(x_i, x_j) & k_{1,2}(x_i, x_j) \\ k_{2,1}(x_i, x_j) & k_{2,2}(x_i, x_j) \end{bmatrix}, \quad (4.5)$$

where $k_{1,1}(x_i, x_j)$ denotes the covariance between the changes in latitudes at the source locations x_i and x_j , and $k_{1,2}, k_{2,1}$ and $k_{2,2}$ are defined in a similar way. We consider the separable covariance structure as Wang and Gelfand (2014) that separates the covariance between the outputs and that between the inputs. (4.5) becomes

$$\mathbf{k}(x_i, x_j) = \Sigma_{\mathbf{C}} \otimes k(x_i, x_j), \quad (4.6)$$

where $\Sigma_{\mathbf{C}}$ is the covariance matrix of horizontal and vertical changes, k is the univariate covariance function and $k(x_i, x_j)$ is the covariance of baseline changes at the source locations x_i and x_j . This decomposition supposes f_1 and f_2 have the same length-scales, and the correlation between the two functions are homogeneous across the input spaces. The covariance matrix $\Sigma_{\mathbf{C}}$ can be expressed by standard deviations and a correlation matrix,

$$\Sigma_{\mathbf{C}} = \text{diag}(\alpha_{\mathbf{C}}) \mathbf{C} \text{diag}(\alpha_{\mathbf{C}}), \quad (4.7)$$

where $\alpha_{\mathbf{C}} = (\alpha_{C_1}, \alpha_{C_2})$ represents standard deviations of the changes along the latitudes and longitudes, $\mathbf{C} \in \mathbb{R}^{2 \times 2}$ is a symmetric and positive definite correlation matrix, e. g. $C_{1,2}$ represents the correlation between the changes in the latitude and longitude. If $\mathbf{C}$ is an identity matrix,

$$\mathbf{k}(x_i, x_j) = \begin{bmatrix} \alpha_{C_1}^2 k(x_i, x_j) & 0 \\ 0 & \alpha_{C_2}^2 k(x_i, x_j) \end{bmatrix}. \quad (4.8)$$

Equivalently, f_1 and f_2 are independent,

$$\begin{aligned} f_1 &\sim \text{GP}(0, \alpha_{C_1}^2 k), \\ f_2 &\sim \text{GP}(0, \alpha_{C_2}^2 k). \end{aligned} \quad (4.9)$$

Then for the input $\mathbf{x} \in \mathbb{R}^{n^+ \times 2}$, $\mathbf{k}(\mathbf{x}, \mathbf{x})$ is decomposed into the covariance between rows of $\mathbf{f}(\mathbf{x})$ and between columns of $\mathbf{f}(\mathbf{x})$. There is no need to work on the covariance matrix $\mathbf{k}(\mathbf{x}, \mathbf{x}) \in \mathbb{R}^{2n^+ \times 2n^+}$, and instead we are able to work on $k(\mathbf{x}, \mathbf{x}) \in \mathbb{R}^{n^+ \times n^+}$ and $\mathbf{C} \in \mathbb{R}^{2 \times 2}$,

which is computationally easier,

$$\begin{aligned} \text{vec}(\mathbf{f}(\mathbf{x})) &\sim \text{Normal}(\mathbf{0}, \mathbf{k}(\mathbf{x}, \mathbf{x})), \\ \mathbf{k}(\mathbf{x}, \mathbf{x}) &= (\text{diag}(\boldsymbol{\alpha}_{\mathbf{C}}) \mathbf{C} \text{diag}(\boldsymbol{\alpha}_{\mathbf{C}})) \otimes k(\mathbf{x}, \mathbf{x}). \end{aligned} \quad (4.10)$$

For the purpose of smoothing and uncovering the structure, one option of the covariance function k is the square exponential kernel,

$$k(x_i, x_j | \sigma_k, \ell) = \sigma_k^2 \exp\left(\frac{1}{2} \sum_{d=1}^2 \frac{(x_{id} - x_{jd})^2}{\ell_d^2}\right), \quad (4.11)$$

where σ_k is the marginal standard deviation of the univariate kernel k , ℓ_d is the length-scale along the input dimension d . The kernel depends on the distance only and suggests the similarity between responses at two locations does not depend on the source locations, but on the proximity between the source locations. In addition, if the two source locations are close, the changes in latitudes and longitudes of the flow keep similar.

One simple model is to treat longitudes and latitudes as independent and set $\mathbf{C}$ to be an identity matrix (4.8), and we call this Model 1. In an alternative model, we incorporate the correlation between the changes in longitudes and latitudes, and estimate the correlation matrix in (4.6), which is called Model 2. The final model, with the name of Model 3, we consider is to model male-to-female and female-to-male transmissions by distinctive Gaussian processes. In particular, assume the first n_1^+ rows of $\mathbf{x} \in \mathbb{R}^{n^+ \times 2}$ are from men to women and the last n_2^+ from women to men, where $n^+ = n_1^+ + n_2^+$. The model is then

$$\begin{aligned} \mathbf{y} &\sim \text{Normal}(\mu_{x,y} \mathbf{1}_{x,y} + \mathbf{f}, \boldsymbol{\sigma}_{\mathbf{y}}^2), \\ x &= \text{latitude, longitude}, y = \text{mf, fm} \\ \mathbf{f} &\sim \text{GP}(\mathbf{0}, \mathbf{k}), \\ \boldsymbol{\sigma}_{\mathbf{y}} &\sim \text{Half-Normal}(0, 1), \end{aligned} \quad (4.12)$$

where the terms are defined the same as (4.4) except the dependence of the baselines $\mu_{x,y}$ on both coordinates x and genders y . In addition, $\mathbf{k}$ is extended from (4.5) to further incorporate the correlation between male-to-female and female-to-male transmissions,

$$\mathbf{k}(x_i, x_j) = \begin{bmatrix} \mathbf{k}_{1,1}(x_i, x_j) & \mathbf{k}_{1,2}(x_i, x_j) \\ \mathbf{k}_{2,1}(x_i, x_j) & \mathbf{k}_{2,2}(x_i, x_j) \end{bmatrix}, \quad (4.13)$$

where $\mathbf{k}_{d,d'}, d, d' = 1, 2$ are 2×2 covariance matrices between the viral movement from male to female along the dimension d and that from female to male along the dimension d' . As motivated by the separable kernel (4.6), we further separate the covariance by dimensions and by genders. This assumes the correlation between the changes in the two coordinates is homogeneous regardless of sources' gender. That is,

$$\mathbf{k}(x_i, x_j) = \Sigma_{\mathbf{Q}} k(x_i, x_j) = (\Sigma_{\mathbf{C}} \otimes \Sigma_{\mathbf{H}}) k(x_i, x_j), \quad (4.14)$$

where $\Sigma_{\mathbf{Q}}$ is the covariance matrix between the changes by both dimensions and genders, $\Sigma_{\mathbf{C}}$ is interpreted the same as (4.6), and $\Sigma_{\mathbf{H}}$ is the covariance matrix of the male-to-female and female-to-male viral movement. In a similar way to (4.10), for the input $\mathbf{x}$, the covariance matrix is

$$\mathbf{k}(\mathbf{x}, \mathbf{x}) = \Sigma_{\mathbf{C}} \otimes \left(\begin{bmatrix} \mathbf{1}_{n_1^+, n_1^+} \Sigma_{H_{1,1}} & \mathbf{1}_{n_1^+, n_2^+} \Sigma_{H_{1,2}} \\ \mathbf{1}_{n_2^+, n_1^+} \Sigma_{H_{2,1}} & \mathbf{1}_{n_2^+, n_2^+} \Sigma_{H_{2,2}} \end{bmatrix} \circ k(\mathbf{x}, \mathbf{x}) \right). \quad (4.15)$$

The details are provided in Appendix B.

We complete the model specification by describing the prior distributions, and the choices are similar to the Stan Development Team (2018). The priors on the baselines and marginal standard deviations are overall uninformative. Betancourt (2017b) suggested imposing a weakly informative prior on the length scales because the length-scales are non-identifiable beyond the range of covariate distance. Inverse Gamma prior constrains the length-scales within the covariate ranges. A popular prior choice for covariance matrix is inverse-Wishart distribution, mainly because of its favourable conjugate property with the multivariate normal likelihood. However, the advantage is restricted to Gibbs sampling and does not apply to the Hamiltonian Monte Carlo (HMC) sampling. One may thus apply alternative priors such as Lewandowski-Kurowicka-Joe (LKJ) prior (Lewandowski et al., 2009) to covariance matrices in Stan that depend on HMC. An advantage is that it is straightforward to express the prior belief on the covariance matrix. Specifically, if $\mathbf{C}$ follows a LKJ distribution with parameter η ,

$$p(\mathbf{C}|\eta) \propto (\det \mathbf{C})^{\eta-1}, \quad (4.16)$$

where $\eta = 1$ represents the uniform density across the correlation matrices, $\eta > 1$ prefers the identity matrix, and $\eta < 1$ favours the strong correlation. We impose the LKJ (3) on $\mathbf{C}$ and $\mathbf{H}$ that favours the null hypothesis (Model 1), and suggests the 95% credibility interval

of correlation is $(-0.71, 0.71)$. That is,

$$\begin{aligned}\mu_x, \mu_{x,y} &\sim \text{Normal}(0, 10), \\ \text{where } x &= \text{latitude, longitude and } y = \text{mf, fm}, \\ \alpha_{C_d} &\sim \text{Half-Normal}(0, 10), d = 1, 2, \\ \ell_d &\sim \text{InverseGamma}(5, 5), d = 1, 2, \\ \mathbf{C}, \mathbf{H} &\sim \text{LKJ}(3).\end{aligned}\tag{4.17}$$

Note we have not specified the priors for σ_k and $\alpha_{\mathbf{H}}$, and will then discuss in the next Section. The resulting posterior distribution of Model 3 is

$$\begin{aligned}p(\mu, \mathbf{f}, \sigma_y, \alpha_{\mathbf{C}}, \alpha_{\mathbf{H}}, \sigma_k, \ell, \mathbf{C}, \mathbf{H} | \mathbf{y}) &\propto \\ p(\mathbf{y} | \mu, \mathbf{f}, \sigma_y) p(\mu) p(\mathbf{f} | \alpha_{\mathbf{C}}, \alpha_{\mathbf{H}}, \mathbf{C}, \mathbf{H}, \sigma_k, \ell) p(\sigma_y) & \\ p(\alpha_{\mathbf{C}}) p(\alpha_{\mathbf{H}}) p(\sigma_k) p(\ell) p(\mathbf{C}) p(\mathbf{H}). &\end{aligned}\tag{4.18}$$

4.3 Numerical inference of the Gaussian process flow model

4.3.1 Identifiability

The non-parametric Gaussian process flow model was fitted in Stan 2.21.0 using no-U-turn sampler (NUTS) HMC algorithm (Carpenter et al., 2017). The first problem with (4.18) is the model identification. For simplification, we further assume $\alpha_H = \alpha_{H_1} = \alpha_{H_2}$. This assumption is not too restrictive and suggests the deviation of changes along d away from the mean of changes along d is the same between male-to-female and female-to-male transmissions. Equation (4.15) then rewrites to

$$\begin{aligned}\mathbf{k}(\mathbf{x}, \mathbf{x}) &= \Sigma_{\mathbf{C}} \otimes \left(\begin{bmatrix} \mathbf{1}_{n_1^+, n_1^+} \Sigma_{H_{1,1}} & \mathbf{1}_{n_1^+, n_2^+} \Sigma_{H_{1,2}} \\ \mathbf{1}_{n_2^+, n_1^+} \Sigma_{H_{2,1}} & \mathbf{1}_{n_2^+, n_2^+} \Sigma_{H_{2,2}} \end{bmatrix} \circ k(\mathbf{x}, \mathbf{x}) \right) \\ &= (\text{diag}(\alpha_{\mathbf{C}}) \mathbf{C} \text{diag}(\alpha_{\mathbf{C}})) \otimes \left(\alpha_H^2 \begin{bmatrix} \mathbf{1}_{n_1^+, n_1^+} H_{1,1} & \mathbf{1}_{n_1^+, n_2^+} H_{1,2} \\ \mathbf{1}_{n_2^+, n_1^+} H_{2,1} & \mathbf{1}_{n_2^+, n_2^+} H_{2,2} \end{bmatrix} \circ k(\mathbf{x}, \mathbf{x}; \sigma_k, \ell) \right) \\ &= \left((\alpha_H^2 \sigma_k^2 \begin{bmatrix} \alpha_{C_1}^2 & \alpha_{C_1} \alpha_{C_2} \\ \alpha_{C_1} \alpha_{C_2} & \alpha_{C_2}^2 \end{bmatrix}) \circ \mathbf{C} \right) \\ &\otimes \left(\begin{bmatrix} \mathbf{1}_{n_1^+, n_1^+} H_{1,1} & \mathbf{1}_{n_1^+, n_2^+} H_{1,2} \\ \mathbf{1}_{n_2^+, n_1^+} H_{2,1} & \mathbf{1}_{n_2^+, n_2^+} H_{2,2} \end{bmatrix} \circ k(\mathbf{x}, \mathbf{x}; 1, \ell) \right).\end{aligned}\tag{4.19}$$

Equation (4.19) shows the ability to identify the product $\alpha_H^2 \sigma_k^2 \begin{bmatrix} \alpha_{C_1}^2 & \alpha_{C_1} \alpha_{C_2} \\ \alpha_{C_1} \alpha_{C_2} & \alpha_{C_2}^2 \end{bmatrix}$, but the individual terms cannot be identified. Thus we set $\alpha_H = \sigma_k = 1$ and their scales are incorporated in $\begin{bmatrix} \alpha_{C_1}^2 & \alpha_{C_1} \alpha_{C_2} \\ \alpha_{C_1} \alpha_{C_2} & \alpha_{C_2}^2 \end{bmatrix}$. The covariance matrix in Model 3 (4.15) is then

$$\mathbf{k}(\mathbf{x}, \mathbf{x}) = (\text{diag}(\alpha_{\mathbf{C}}) \mathbf{C} \text{diag}(\alpha_{\mathbf{C}})) \otimes \left(\begin{bmatrix} \mathbf{1}_{n_1^+, n_1^+} H_{1,1} & \mathbf{1}_{n_1^+, n_2^+} H_{1,2} \\ \mathbf{1}_{n_2^+, n_1^+} H_{2,1} & \mathbf{1}_{n_2^+, n_2^+} H_{2,2} \end{bmatrix} \circ k(\mathbf{x}, \mathbf{x}; 1, \ell) \right). \quad (4.20)$$

4.3.2 Reparameterisation

Neal (2003) introduces a funnel shaped distribution of which the correlation between variables varies across the space. It is usually inefficient to sample from the funnel distribution through Gibbs sampling, because the much lower conditional variance than marginal variance may make it slow to explore the target distribution (Betancourt and Girolami, 2015). Non-centred parameterisation makes model parameters less correlated and leads to a simpler posterior geometry (Papaspiliopoulos et al., 2007). HMC utilises the gradient information to efficiently explore the geometry. However, it still encounters difficulty exploring the bottom of funnels (Betancourt, 2017a), and the problem is more severe if there are no adequate data (Betancourt and Girolami, 2015). We thus adopt the non-centred parameterisation, and (4.10) rewrites to

$$\begin{aligned} \text{vec}(\mathbf{f}(\mathbf{x})) &= \text{Cholesky}(\text{diag}(\alpha_{\mathbf{C}}) \mathbf{C} \text{diag}(\alpha_{\mathbf{C}}) \otimes k(\mathbf{x}, \mathbf{x}; 1, \ell)) \text{vec}(\beta), \\ &= (\text{diag}(\alpha_{\mathbf{C}}) \mathbf{L}_{\mathbf{C}} \otimes \mathbf{L}_k) \text{vec}(\beta) \\ &= \text{vec}(\mathbf{L}_k \beta (\text{diag}(\alpha_{\mathbf{C}}) \mathbf{L}_{\mathbf{C}})^T), \end{aligned} \quad (4.21)$$

where $\beta \in \mathbb{R}^{n^+ \times 2}$ is a vector of independent identically distributed (i.i.d.) standard normal random variables, and where $\mathbf{L}_k$ and $\mathbf{L}_{\mathbf{C}}$ are Cholesky decompositions of $k(\mathbf{x}, \mathbf{x}; 1, \ell)$ and $\mathbf{C}$ respectively.

4.3.3 Prediction

We then explain how to make predictions on arbitrary geographical positions denoted by $\mathbf{x}^*$. The predictions at the new set of inputs $\mathbf{x}^*$ are denoted as $\mathbf{f}(\mathbf{x}^*)$, and can be drawn from

the following predictive distribution,

$$\begin{aligned} \text{vec}(\mathbf{f}(\mathbf{x}^*)) | \mathbf{y}, \mathbf{x}, \mathbf{x}^* \sim \text{Normal} \Big(& \mathbf{k}(\mathbf{x}^*, \mathbf{x}) (\mathbf{k}(\mathbf{x}, \mathbf{x}) + \sigma_y^2 \mathbf{I})^{-1} \mathbf{y}, \\ & \mathbf{k}(\mathbf{x}^*, \mathbf{x}^*) - \mathbf{k}(\mathbf{x}^*, \mathbf{x}) (\mathbf{k}(\mathbf{x}, \mathbf{x}) + \sigma_y^2 \mathbf{I})^{-1} \mathbf{k}(\mathbf{x}, \mathbf{x}^*) \Big), \end{aligned} \quad (4.22)$$

(Rasmussen and Williams, 2006), and the covariance matrix is (4.10) for Model 1 and Model 2 and (4.15) for Model 3. Thus, the changes in both directions and the location of flow destination can be predicted once a new set of source locations is given. These can be alternatively expressed in the polar coordinate, which decomposes the viral movements into angles and radius for the illustration purpose. The angle of viral movement is defined as

$$\theta = \arctan\left(\frac{z_{1,i} - x_{1,i}}{z_{2,i} - x_{2,i}}\right) = \arctan\left(\frac{y_{1,i}}{y_{2,i}}\right), \quad (4.23)$$

and the distance from the location x_i to reach the location y_i is

$$r_i = \sqrt{(z_{1,i} - x_{1,i})^2 + (z_{2,i} - x_{2,i})^2} = \sqrt{y_{1,i}^2 + y_{2,i}^2}. \quad (4.24)$$

(4.23) and (4.24) are useful to decompose the arrows into directions and distance, and thus clearly explore the structured directions.

4.3.4 Experiments

Next, we conducted several experiments to demonstrate the efficiency of these transformations. First we adopted the centred parameterisation of (4.21),

$$\text{vec}(\mathbf{f}) \sim \text{Normal}(\mathbf{0}, \mathbf{k}(\mathbf{x}, \mathbf{x})), \quad (4.25)$$

where the covariance matrix $\mathbf{k}(\mathbf{x}, \mathbf{x})$ is (4.20) in Model 3. We then wish to compare the centred parameterisation (4.25) and the non-centred one (4.21). As discussed in Section 4.3.2, we would expect (4.21) to be more efficient and avoid potential divergent transitions in HMC. However, we are unable to implement (4.25). The reason might be the inversion of the large matrix $\mathbf{k}(\mathbf{x}, \mathbf{x})$. This shows the additional benefits of (4.21) in avoiding the computationally intensive matrix computation.

We then compared an uninformative prior on the length-scales to the weakly informative inverse Gamma prior,

$$\ell_d \sim \text{Normal}(0, 10), d = 1, 2. \quad (4.26)$$

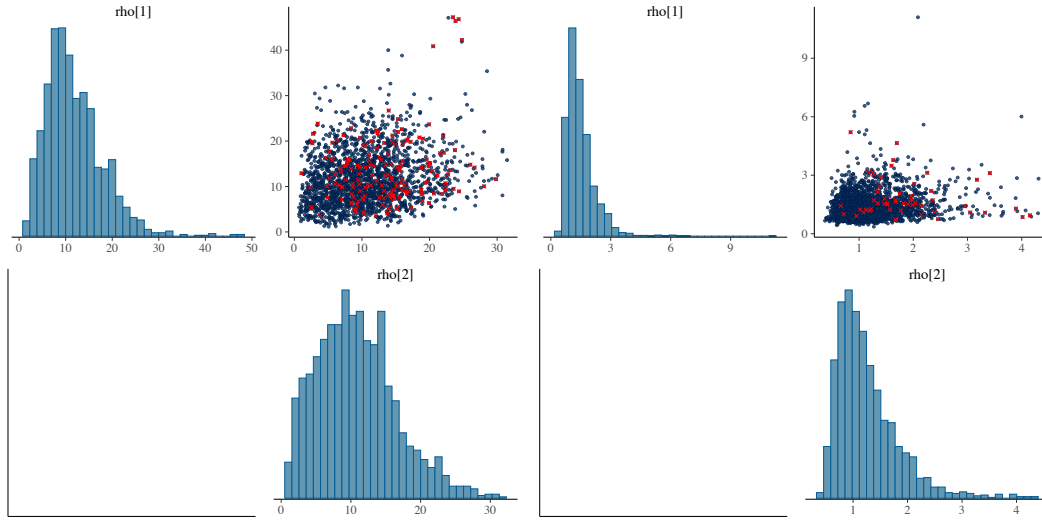


Figure 4.2 **Posterior distributions under weakly informative and uninformative priors.** We imposed weakly informative (left) and uninformative (right) priors on the length-scales and plot pairwise posterior samples of the length-scales. The divergent transitions are highlighted in red.

Figure 4.2 shows the posterior distributions under the weakly informative and uninformative priors have substantially different scales. Length-scale estimates appear highly influenced by the prior specification. Also, there are more divergent transitions (196) under the uninformative prior in Figure 4.2 (left) than the weakly informative prior (52) in Figure 4.2 (right). The divergent transitions are highlighted in red and most occur for the large length-scales. The high number of divergent transitions is undesired and caused by the extreme length-scales that are not identifiable from the covariates domain (Betancourt, 2017b). Moreover, it takes 2.66 days to implement the model with uninformative prior. Therefore, we found weakly informative priors to be more efficient.

4.3.5 Numerical performance on Rakai data

We made inferences of Model 1-3 in Stan, where the numerical performance is documented in Table 4.1. We also show the trace plots of the three worst parameters in Model 3 in terms of effective sample sizes (ESS) in Figure 4.3. We identified no problems in mixing and convergence.

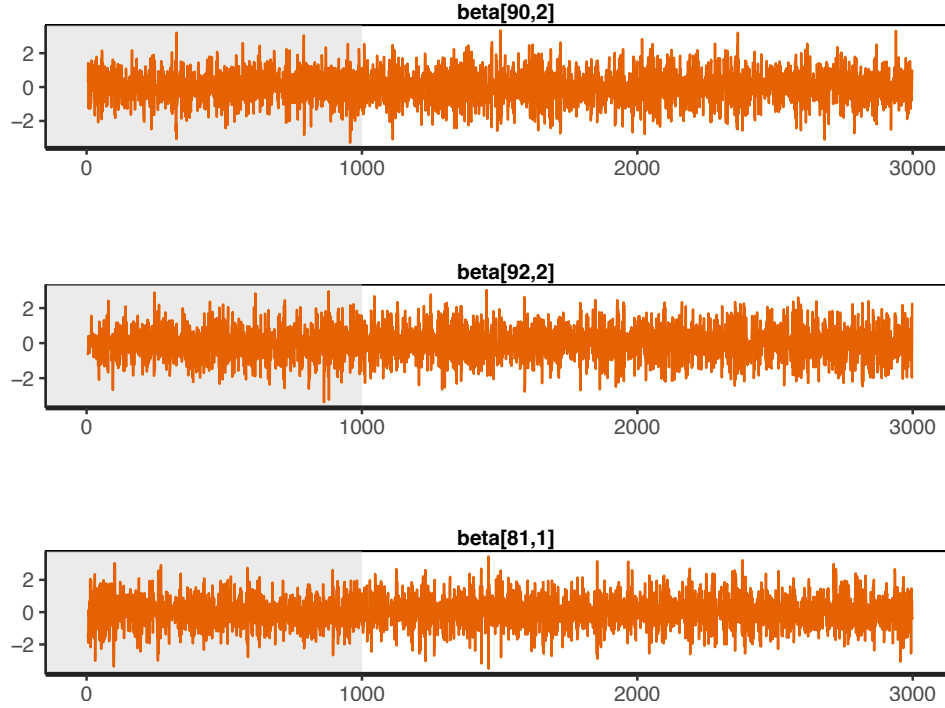


Figure 4.3 **Trace plots for parameter inference on the Rakai data set.** Stan’s dynamic Hamiltonian Monte Carlo was run for 3×10^3 iterations. Shown are trace plots for three parameters with the lowest ESSs.

Table 4.1 Numerical performance of the Gaussian process flow model.

Method	Run time (hours)	Effective Sample Size
1	8.35	2000 (1432.40 - 2872.59)
2	17.86	2000 (1301.48 - 3114.12)
3	11.61	2000 (1358.97 - 2789.73)

4.4 Model performance measures

Simple measures of model performance, including mean absolute error, worst case error, and whether the observed value is covered by 95% posterior credibility interval, are first introduced. Recall the observed changes in the two directions are in the columns of $\mathbf{y} \in \mathbb{R}^{n^+ \times 2}$ and the estimated changes from the fitted model are the columns of $\hat{\mathbf{y}} \in \mathbb{R}^{n^+ \times 2}$. Then, the mean absolute error (MAE) is defined as

$$\text{MAE}(\mathbf{y}, \hat{\mathbf{y}}) = \frac{1}{2n^+} \sum_{i,d} |\hat{y}_{i,d} - y_{i,d}|. \quad (4.27)$$

The worst-case error (WCE) is defined as

$$\text{WCE}(\mathbf{y}, \hat{\mathbf{y}}) = \max_{i,d} |\hat{y}_{i,d} - y_{i,d}|, \quad (4.28)$$

and they quantify the average and largest discrepancies between the observed and predicted changes in both directions on the observed source locations. Denote the 2.5% quantile of $\hat{y}_{i,d}$ as $q_{2.5}(\hat{y}_{i,d})$ and the 97.5% quantile of $\hat{y}_{i,d}$ as $q_{97.5}(\hat{y}_{i,d})$, and the 95% posterior predictive credibility interval (PPCrI) of the i th point on the dimension d is then $[q_{2.5}(\hat{y}_{i,d}), q_{97.5}(\hat{y}_{i,d})]$. Then the proportion of the observation in 95% PPCrI is defined as

$$\text{Data in 95\% PPCrI} = \frac{\sum_{i,d} \mathbb{1}\{y_{i,d} \in [q_{2.5}(\hat{y}_{i,d}), q_{97.5}(\hat{y}_{i,d})]\}}{2n^+}. \quad (4.29)$$

In addition to these simple measures, the continuous ranked probabilistic score (CRPS) quantifies the accuracy of a forecast distribution F , given the observations $\mathbf{y} \in \mathbb{R}^{n^+ \times 2}$, where F can either be in a parametric form or empirically estimated by a sample of draws (Jordan et al., 2017). The CRPS is defined by

$$\text{CRPS}(F, y_{[i,d]}) = \int (F(z) - \mathbb{1}\{y_{[i,d]} \leq z\})^2 dz, i = 1, \dots, n^+, d = 1, 2. \quad (4.30)$$

If the first moment of F is finite, it can be rewritten as

$$\text{CRPS}(F, y_{[i,d]}) = E_F |X - y_{[i,d]}| - 0.5 E_{F,F} |X - X'|, \quad (4.31)$$

where X, X' are independently sampled from F . (4.31) suggests the dependence of CRPS on the magnitude of the observations. To make CRPSs in both dimensions comparable, we apply the relative CRPS (rCRPS),

$$\text{rCRPS}(F, y_{[i,d]}) = \text{CRPS}(F, y_{[i,d]}) / y_{[i,d]}. \quad (4.32)$$

Another useful technique is cross-validation but this typically requires re-fitting the model. For example, leave-one-out cross-validation (LOO) uses one observation as the test set, the rest of the observations as the training set, and validates the accuracy of the model fitted to the training data on the test data. In particular, the expected log pointwise predictive density (ELPD) for a new dataset is defined as

$$\text{ELPD} = \sum_i p(\mathbf{y}_{[i]} | \mathbf{y}_{[-i]}) = \sum_i \int p(\mathbf{y}_{[i]} | \boldsymbol{\theta}) p(\boldsymbol{\theta} | \mathbf{y}_{[-i]}) d\boldsymbol{\theta}, \quad (4.33)$$

where $p(\mathbf{y}_{[i]}|\theta)$ is the likelihood (4.10) evaluated at the i th data, $p(\theta|\mathbf{y}_{[-i]})$ denotes the posterior distribution of model parameters given the training data, and θ refers to the set of model parameters. Vehtari et al. (2017) shows it can be numerically approximated without refitting through importance sampling. The numerical approximation to LOO can be improved through Pareto smoothed importance sampling (PSIS) that addresses the problem of infinite variance of importance weights (Vehtari et al., 2015). The implementation is through `loo` version 2.1.0.

The final measure that can be efficiently calculated from draws via Markov Chain Monte Carlo (MCMC) is the Bayes factor. The marginal likelihood, also called model evidence, is defined as

$$p(\mathbf{y}|M_i) = \int p(\mathbf{y}|\theta)p(\theta|M_i)d\theta. \quad (4.34)$$

The ratio $p(\mathbf{y}|M_1)/p(\mathbf{y}|M_2)$ is the Bayes factor and measures the strength of evidence in favour of M_1 relative to M_2 . $p(\mathbf{y}|M_1)$ is usually intractable, but can be effectively estimated with bridge sampling (Gronau et al., 2017a; Meng and Wong, 1996). In this Chapter, all performance measures except the Bayes factor will be used, as we are not able to make `bridgesampling` version 1.1.2 work. One possible solution to the problem will be described in Chapter 7.

4.5 Preliminary results

We fitted Model 1-3 to the RCCS phylogenetically reconstructed viral movement data, and Table 4.2 compares the inference accuracy. The third model achieved the best results in terms of posterior predictive checks, although the difference to the other models was subtle. Model 3 also has marginally higher ELPD in the leave-one-out cross-validation. To compare ELPD quantitatively, we assume the differences of ELPD are approximately normally distributed in presence of large sample sizes (Vehtari et al., 2017) and assess the p-value of the hypothesis test where the null hypothesis is that the difference of ELPD is 0. We found the differences between Model 3 and Model 2, and between Model 3 and Model 1 are statistically insignificant, with p-values 0.3396 and 0.0669. The subsequent analyses were based on Model 3.

We next discuss our key findings of Model 3. The top row of Figure 4.4 illustrates the predicted angles (4.23) on the left and distances (4.24) on the right for the directed transmissions originated from men (in blue) and women (in red) at the equidistant locations. The corresponding standard deviations are in the bottom row of Figure 4.4.

Table 4.2 Comparison of prediction accuracy in the non-parametric Gaussian Process flow model.

Method	MAE	WCE	Data in 95% PPCrI	rCRPS	ELPD
1	0.0385	0.0385	0.9300	-0.11(-0.47 - 0.01)	-706.2 (42.8)
2	0.0378	0.0378	0.9317	-0.11(-0.47 - 0.01)	-705.8 (43.2)
3	0.0373	0.0373	0.9317	-0.11(-0.47 - 0.01)	-703.3 (43.2)

Table 4.3 Summary of posterior median of flow directions in the regions stratified by the latitude -0.7485917 and longitude 31.70931. The summary is provided by median (minimum - maximum).

Longitude	Latitude	Angles from men	Angles from women
> 31.70931	> -0.7485917	0.99 (0.5 - 1.37)	1.17 (0.5 - 1.51)
> 31.70931	< -0.7485917	-0.75 (-1.28 - -0.04)	-0.91(-1.54 - 1.4)
< 31.70931	> -0.7485917	-0.78 (-1.38 - -0.01)	-0.68 (-1.25 - -0.18)
< 31.70931	< -0.7485917	0.99 (0.53 - 1.44)	0.55 (0.17 - 1.01)

First, we observe marked heterogeneity of transmission directions across the geography, and stratify the space by the longitude 31.70931 and the latitude -0.7485917 in Table 4.3. The two separation lines were not considered in the analysis because of more noises in estimated directions. We focus on the estimated direction of flows originating from men. In particular, most of the flows in the top right block point to the northeast, and this also applies to the bottom left block. In both the top left and bottom right blocks, the flows are towards the southeast in general. The overall direction of flows from women in the four blocks is similar to that of men. However, those in the bottom right block, i. e. the second row in Table 4.3, are less structured and have higher variability.

In addition, we evaluate the difference in flow directions at each of the spatial locations by gender and find the 95% CI. All the intervals contain 0. Given the potential differences by genders in Figure 4.4 and Table 4.3, we investigate the reason for the no differences in flow directions by incorporating the uncertainties. In particular, we focus on two points where the median estimates of angles differ most between the genders, i.e. (-0.5983, 31.7093) and (-0.5482, 31.7093), and visualise the density plot of angles by genders in Figure 4.5. The figure shows the bi-modal distributions which peak at $\pi/2$ and $3\pi/2$, but the mass concentrates differently, those from women at $3\pi/2$ and men at $\pi/2$. The pattern may be explained by Figure 4.1 which has few observations near the longitude 31.7 and mixed evidence in the left along the highway and right along the shore. This consequently leads to the highest standard deviations in the bottom left panel of Figure 4.4.

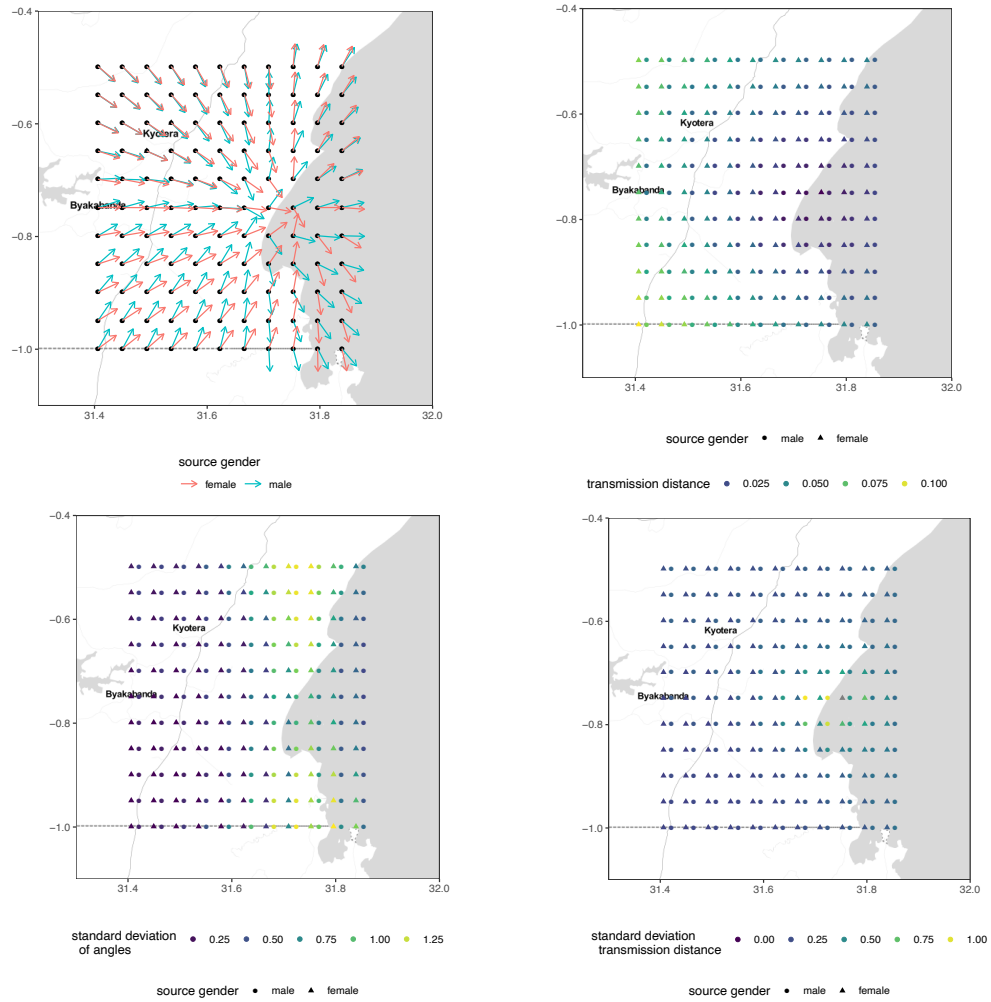


Figure 4.4 **Estimated directed transmission by gender on a regular grid.** A non-parametric model was fitted to the observed directed transmissions. Given a location, we are able to estimate directed transmissions by gender, and decompose these into angles (top left) and distances (top right). Top left: arrows represent transmission directions at a location (points) by gender (in colour). Top right: the distance of transmissions from a location (points) by gender (in shape) is shown in colour. The standard deviations associated with angles and distance are in the bottom left and right panels respectively. The standard deviations of estimates at a spatial location (points) are shown in colour by gender (in shape). As the standard deviations of distance depend on the scale of distance, we considered the coefficient of variations as the uncertainty measure of distance.

Next, we explore the potential impact of the north-to-south main road on the transmission directions. To do this, we calculated the geodesic distance between points on the regular grid and the highway, and classified the points at the grid by whether they are within 1 kilometre from the highway, and summarised the angle estimates by the classes and genders. The angles of transmissions from men are 0.53 (-0.9 - 0.87) and 0.19 (-1.44 - 1.46) at the points near and away from the highway respectively, while the values are 0.18 (-0.77 - 0.45) and 0.01 (-1.44 - 1.49) for women. None of the summaries differs from 0, and this suggested the overall direction of flows was towards Lake Victoria regardless of genders and locations. However, the angle median of male-to-female transmissions is 0.35 radian towards the north than the female-to-male. Therefore, we may argue that the north-to-south highway has a greater impact on shifting the male-to-female transmissions to the north than the opposite counterpart.

Apart from the transmission directions, we then comment on transmission distance and correlation. As seen from the top right panel in Figure 4.4, the distance of phylogenetically reconstructed transmission events decreases as moving from inland to the coast, however, the coefficient of variation slightly increases. Regarding the correlation matrix **C** and **H**, Model 3 suggests the correlation between the coordinates is 0.0650 and that between genders is higher, 0.4400. This means we are more able to borrow information between female-to-male and male-to-female transmissions.

Finally, we illustrate another representation of the estimates which combines the information from the top panels in Figure 4.4, and Figure 4.6 shows the predicted viral movements on the regular grid. We observe the evident transmissions towards the lakeside. Furthermore, only 59 out of 293 (about 20%) transmissions are between households, and most of the transmissions are within the household. If we exclude these transmissions of which the origin and destination are in the same coordinates and repeat the analysis, the predicted transmissions are visualised in Figure 4.7. The arrows are now longer and their structure more obvious. For instance, we are able to spot the opposite transmission directions from the lake and the inland. However, we may overestimate the distances due to the exclusion of the transmissions in the same household.

4.6 Summary

In this Chapter, we estimated the directions and angles of viral movements from the human immunodeficiency virus (HIV) deep-sequence data of the Rakai Health Science Programme

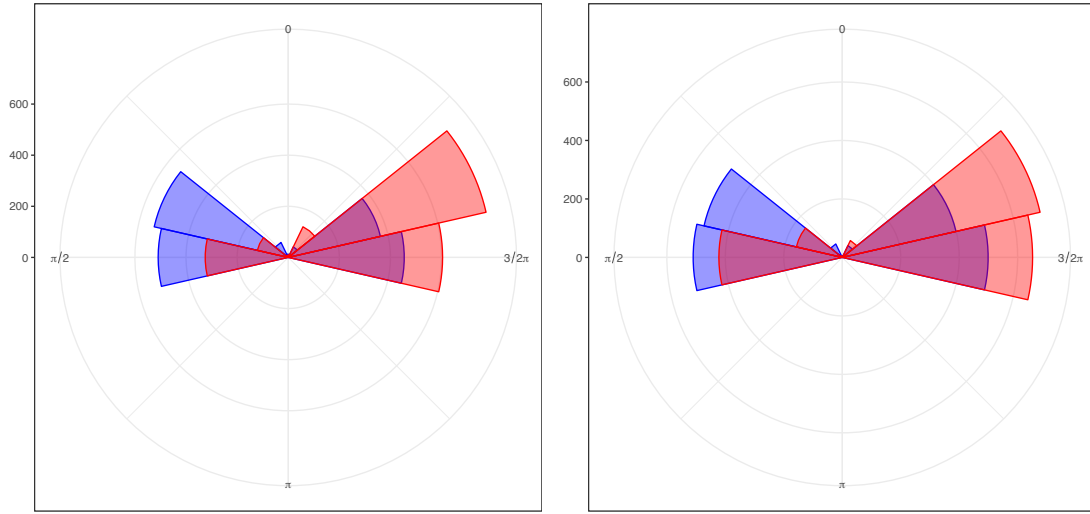


Figure 4.5 **Posterior distributions of angles of transmissions at two points with maximum gender differences in angle median estimates in Figure 4.4.** The two subplots show distributions of transmission angles from men (blue) and women (red) at each of the two locations.

(RHSP). We used a standard non-parametric Gaussian process flow model and fitted the model to the phylogenetically likely source recipient pairs.

We apply the ideas in meteorology to explore the structured transmission directions by gender, and borrow the information between the two types. However, there are several limitations of this study. First, populations live in well-separated towns and villages, and so are not diffused across the entire map. Similarly, the shores of Lake Victoria define the end of settlements, but individuals move frequently and relatively fast over water between fishing villages. As a result, we have doubts about plotting viral movements on a regular grid across the Rakai area, and about how the Gaussian process kernel borrows information across approximate geographic source locations depending on the spatial distance between the source locations. In addition, as a stationary covariance function, the square exponential kernel, presumes the constant covariance structure over the space, as long as the distance between two source locations is fixed. Therefore, this does not work if the flow directions keep similar in some regions and change rapidly over other regions. Also, the location we considered is the coordinate at the time of survey sampling, and this may not be consistent with the location where the transmission occurred and does not incorporate other complicated factors such as mobility. Last, we made the non-parametric inference on the partially observed

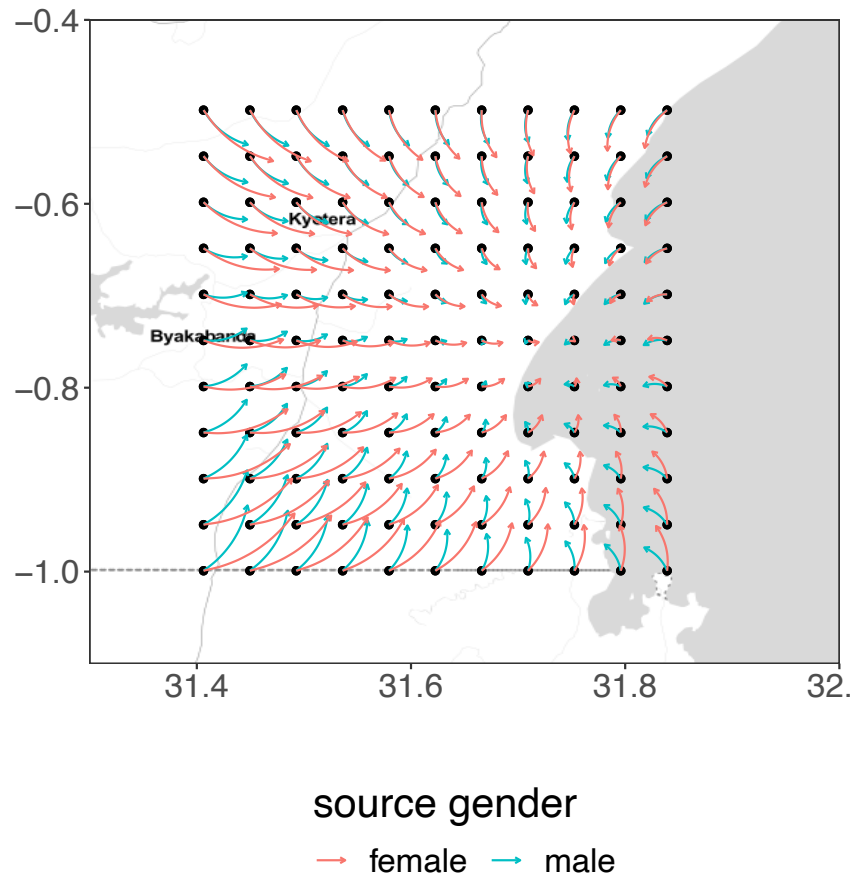


Figure 4.6 **Estimated recipient locations of directed transmissions by gender on a grid of locations.** The transmission destinations were predicted for a grid of points, and are shown on the map of Rakai District. Arrows represent the estimated directed transmissions, and the colour shows the transmissions from men (blue) and women (red).

transmission networks reconstructed from the heterogeneously sampled viral sequence, and thus the inference may be subject to sampling bias.

The main, and surprising, finding of our analysis is that viral flows appear to be towards the high HIV prevalence areas at the lake, rather than in the opposite direction, especially when we included all the phylogenetically source-recipient pairs in the analysis. We did not see how we could address the substantial structural limitations of the Gaussian process model, and we chose to adopt a different methodology that we describe in the next Chapter.

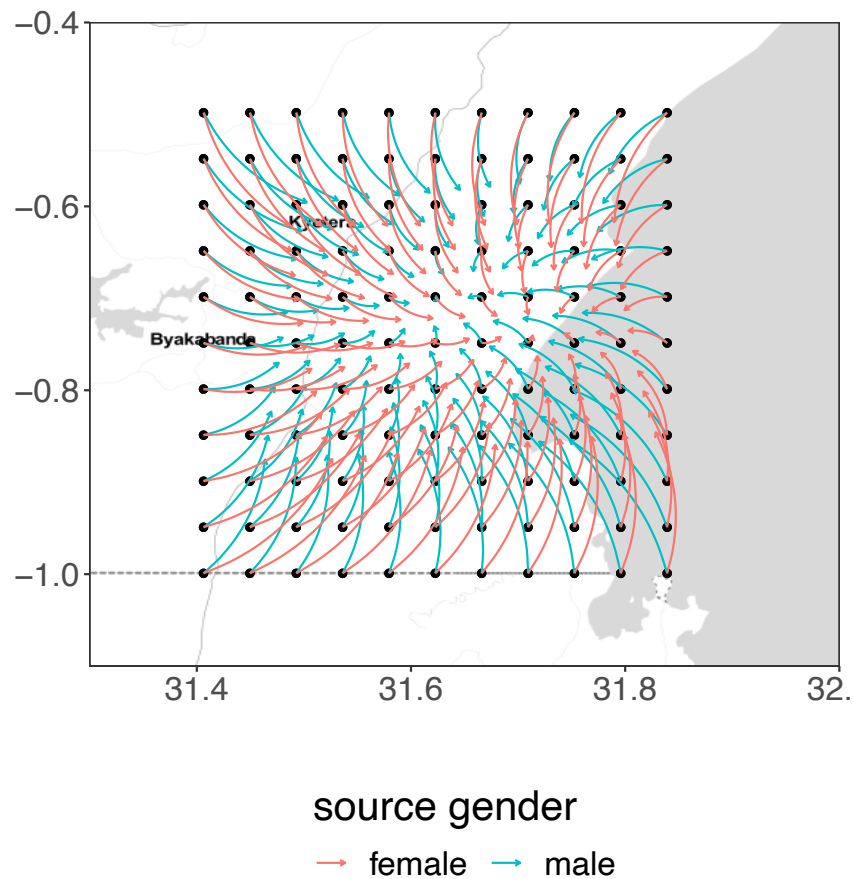


Figure 4.7 **Estimated recipient locations of directed transmissions by gender on a grid of locations, where we included only between household transmissions.** The transmission destinations were predicted for a grid of points, and are shown on the map of Rakai District. Arrows represent the estimated directed transmissions, and the colour shows the transmissions from men (blue) and women (red).

Chapter 5

The basic Bayesian Poisson flow model

5.1 Introduction

In this Chapter, we investigate transmission dynamics by defining and estimating transmission flows. Recall that across Africa, there is an increasing focus on identifying groups of individuals that are at high risk of acquiring human immunodeficiency virus (HIV) and at high risk of spreading the virus, with the goal of targeting control interventions to these groups (UNAIDS, 2018b). 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 5.1). Sources are population groups that disproportionately pass on infection, sinks are groups that disproportionately acquire infection, and hubs are both sources and sinks (Abeler-Dörner et al., 2019). 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. The World Health Organisation (WHO) and others are recommending to target interventions to hotspots to maximise cost-effectiveness of interventions and reduce infection burden in hotspots (Saul et al., 2018; UNAIDS, 2018b). Although often assumed, it is not clear if hotspots are also sources of epidemic spread to the neighbouring lower-prevalence communities (Ratmann et al., 2020). If they are, spatial targeting control efforts could have a substantial knock-on effect in reducing infection in the broader population (Aral et al., 2015). In this application, the population is divided into individuals living in high-prevalence areas (h) and low-prevalence areas (l), and

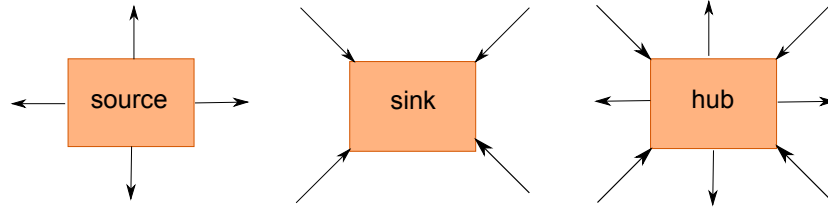


Figure 5.1 **Analysis aims.** The overall aim of phylogenetic source attribution analyses is to infer how pathogens are passed on between population groups. Conceptually populations can be divided into source populations that predominantly transmit disease, sink populations that predominantly receive infection, and hub populations that both disproportionately transmit and receive infections.

then transmission flows are estimated within and between them,

$$\pi = \begin{pmatrix} \pi_{hh} & \pi_{hl} \\ \pi_{lh} & \pi_{ll} \end{pmatrix}, \quad (5.1)$$

where π_{ab} denotes the proportion of transmission flows from group a to group b subject to $\sum_{ab} \pi_{ab} = 1$. Important summary statistics are the vector of sources of infection in group b individuals (δ^b), for example in inland communities; the vector of recipients of infection from group a individuals (ω^a), for example from fisherfolk; 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

$$\delta^b = (\delta_a^b)_{a \in \mathcal{A}}, \quad \delta_a^b = \pi_{ab} / \sum_c \pi_{cb}; \quad (5.2a)$$

$$\omega^a = (\omega_b^a)_{b \in \mathcal{A}}, \quad \omega_b^a = \pi_{ab} / \sum_c \pi_{ac}; \quad (5.2b)$$

$$\gamma_{ab} = \pi_{ab} / \pi_{ba}. \quad (5.2c)$$

5.2 Approaches for quantifying and controlling for sampling bias

As motivated in Chapter 4, an important issue to address is that the interpretation of phylogenetically reconstructed transmission networks is subject to sampling bias. Sampling bias is the bias brought about by the way by which the data were collected, which is not a

rare problem in medical studies. For example, publication bias poses a problem for meta-analysis given the fact that statistically insignificant studies are published less frequently, and unpublished work may introduce bias to the overall effect size of a particular issue (Ioannidis, 2005). In this Section, we will study the literature on publication bias first, and then address the selection bias in phylodynamics.

Jin et al. (2015) summarises the approaches for publication bias adjustment. Most of the methods in the review assume large studies are less likely to be missing, and small studies are more likely to be published if they have a large effect size, e.g. odds ratio. The first type of approach is based on funnel plots which show the effect size against precision (Duval and Tweedie, 2000). Despite of the benefits for interpretation, it is subjective and hard to disentangle the asymmetry brought by study heterogeneity. Another type of approach is model-based and assumes the dependency of publication on sample size and/or p-values. They explicitly model the selection process and estimate the effect size. Givens et al. (1997) proposes a Bayesian hierarchical model and estimates the effect sizes of environmental tobacco smoke on lung cancer incidence, when the publication bias is adjusted through data augmentation. Specifically, the authors apply a random effect model to model the effect sizes of complete study sets, and meanwhile inform the number of missing studies by a selection mechanism, i.e. they partition the p-values into intervals with varying publication probabilities. Silliman (1997) and Larose and Dey (1998) also use a Bayesian hierarchical model which captures the within-study and between-study variability. Nevertheless, Silliman (1997) models the selection by weighted distributions which are either parametric dependant on the size/significance of the study or non-parametric characterised by step functions. Instead of assuming one weight function, Larose and Dey (1998) compares five competing weight functions through Bayes factor, and chooses the best model specification preferred by data. Baker and Jackson (2006) incorporates more information such as sample size, study type, and size of collaboration and journal impact factors to construct a more realistic publication process. Guan and Vandekerckhove (2016) considers four generative publication processes including the model without publication bias, and averaging over the models instead of selecting one. This framework allows for detection of publication bias and meanwhile an accurate estimate of the true effect size. Overall, most of the methods handle the publication bias explicitly by modelling the sampling process, which can be integrated with the model for effect sizes through data augmentation in the Bayesian hierarchical model framework. We will apply the same strategy to our problem.

In the phylodynamics analysis, heterogeneous sampling and incomplete sampling are less addressed (Frost et al., 2015). Volz et al. (2009) assumes that a fraction of the infected popu-

lations is sampled simultaneously at a particular time, derives the probability of observing the transmission induced coalescence and thus considers the incomplete spatial sampling. Stadler (2010) derives the density of a sampled tree generated by birth-death process under certain conditions such as root ages, where the sampling is at present with a probability and over time with another probability. Both studies incorporate the incomplete sampling spatially (Stadler, 2010) and/or temporally (Stadler, 2010; Volz et al., 2009) but not heterogeneous sampling. Hall et al. (2016) demonstrates the spatial and temporal preferential sampling both impact the phylogenetics and phylodynamics significantly through simulated experiments, and Frost et al. (2015) highlights more efforts should be investigated to address the bias introduced by preferential sampling. Karcher et al. (2016) describes the sampling processes by an inhomogeneous Poisson process of which intensity depends on effective population size, and thus accounts for temporal preferential sampling in the coalescent model. Kühnert et al. (2016) generalises birth-death model to multi-type birth-death-sampling model, and thus allows the varying sampling fractions by type and piecewise constant sampling fractions over time, which is further accelerated in Scire et al. (2020). Despite the substantial developments in the phylodynamic inference under heterogeneous sampling, spatial sampling heterogeneity is either not incorporated (Karcher et al., 2016) or computationally challenging to consider (Kühnert et al., 2016), especially in presence of a large number of sequences and demes (See Section 5.5.2 for details).

Apart from phylodynamics, heterogeneous sampling is considered in other phylogenetic analysis. Linkage analysis is useful in identifying risk factors of transmission clusters. Carnegie et al. (2014) assumes random sampling conditional on group membership, where groups are defined by a set of characteristics such as demographical factors, risk behaviours and disease progression, and infers viral linkage probabilities. Ratmann et al. (2016) reconstructs probable transmissions through phylogenetic relationships, determines their stage using clinical data, and thereby quantifies the transmissions from early stage of infections. The authors compare the cohort to a near-complete population cohort and characterise the missing cases by negative binomial distributions. These phylogenetic studies adjust for uneven sampling by assuming sampling at random conditional on group memberships and specifying sampling fractions from external information.

In contrast with the methods based on Sanger sequences, the emerging deep sequence data potentially offer a new avenue to tackle the problem. In particular, we are able to reconstruct partially observed transmission networks as in Chapter 3, describe sampling processes from detailed demographical data, and integrate the process with complete transmission dynamics inference in a Bayesian hierarchical framework, as motivated by the literature

on publication bias. To characterise the sampling fractions in the survey, we adopt the practice in the other phylogenetic studies, assume missing at random within subpopulations stratified by individual characteristics and utilise the external information. It would be more computationally efficient to work on the reconstructed pairs directly rather than the phylogenies.

5.3 Development of the Bayesian Poisson flow model

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 $i = 1, \dots, N$ 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 π with entries $\pi_{ab} = z_{ab}/z^+$ where $z^+ = \sum_{ab} z_{ab}$. The flow matrix is in general not symmetric, has positive diagonal entries, 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 (Chan et al., 2014). We denote the number of structurally non-zero entries by L , which satisfies $L \leq A^2$.

We next put the flow matrix into context of the standard epidemiological model. In general, the flow matrix is time-dependent due to changes in population composition and varying transmission rates (Anderson and 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 ordinary differential equations (ODE) pertaining to the male (m) high-risk population are

$$\begin{aligned}\dot{S}_{mh} &= -\lambda(t)S_{mh}(t) + \mu - \mu S_{mh}(t) \\ \dot{I}_{mh} &= \lambda(t)S_{mh}(t) - \gamma(t)I_{mh}(t) - \mu I_{mh}(t) \\ \dot{T}_{mh} &= \gamma(t)I_{mh}(t) - \mu T_{mh}(t),\end{aligned}\tag{5.3}$$

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 $\mathcal{T} = [t_1, t_2]$ is

$$z_{fh,mh}([t_1, t_2]) = \int_{t_1}^{t_2} \beta_{fh}(t) I_{fh}(t) S_{mh}(t) / N_{fh}(t) dt, \quad (5.4)$$

and the corresponding proportion of transmissions is

$$\pi_{fh,mh}([t_1, t_2]) = \frac{z_{fh,mh}([t_1, t_2])}{z^+([t_1, t_2])}, \quad (5.5)$$

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 $\mathcal{T} = [t_1, t_2]$. The compartment model will not be used in our model, but will be used in the subsequent simulations in Section 5.4.1.

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 in our exposition 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 $w_{ij} \in [0, 1]$ in (3.9) or (3.10). The direction score matrix is not symmetric, diagonal entries are zero, and entries involving unsampled individuals are missing. To estimate the flow matrix π , consider the observed flow counts

$$n_{ab} = \sum_{i \in a, j \in b} \mathbb{1}\{s_i = 1\} \mathbb{1}\{s_j = 1\} \hat{z}_{ij} = \sum_{i \in a, j \in b} \mathbb{1}\{s_i = 1\} \mathbb{1}\{s_j = 1\} \mathbb{1}\{w_{ij} > \zeta\}, \quad (5.6)$$

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. An possible improvement in future work would be to remove the threshold ζ , and instead propagate uncertainty in the direction scores. The counts can be arranged into the $A \times A$ count matrix $\mathbf{n}$, and sum to $n^+ = \sum_{a,b} n_{ab}$. In the previous studies (Hall et al., 2021; Ratmann et al., 2020), ζ takes the values 0.6 or 0.33, depending on which of the decision rule (3.9) or (3.10) is used, and the resulting n^+ is usually between 100 and 500. The naïve flow estimator

is defined by

$$\hat{\pi}_{ab}^n = \frac{n_{ab}}{\sum_{c,d} n_{cd}}. \quad (5.7)$$

If we suppose that each population group a is independently sampled at random with probability ξ_a , and the actual flows from a group a to group b are z_{ab} , then

$$\hat{\pi}_{ab} = \frac{z_{ab} \xi_a \xi_b}{\sum_{c,d} z_{cd} \xi_c \xi_d}. \quad (5.8)$$

This clarifies that the naïve flow estimator (5.7) is only unbiased when the population groups were homogeneously sampled, i.e. ξ_a is the same for all a , which is rarely the case (Ratmann et al., 2020).

5.3.1 Model framework

The data inputs for estimating transmission on flows from pathogen deep-sequence data (5.6) 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 Kastele et al., 2017), prompting us to formulate the statistical model in general terms. Considering the actual, unobserved number of transmissions between all population groups, $\mathbf{z} = (z_{ab})_{a,b \in \mathcal{A}}$, the complete data likelihood that arises under mathematical models of the form (5.3) in a fixed study period is the multinomial

$$p(\mathbf{z} | \mathbf{z}^+, \boldsymbol{\pi}) \propto \prod_{a,b} (\pi_{ab})^{z_{ab}}, \quad (5.9)$$

where for ease of notation we denote by $\boldsymbol{\pi}$ the vector of non-zero elements of the flow matrices (5.1), and similarly for $\mathbf{n}$ and $\mathbf{z}$. Model (5.9) ignores potential second-order correlations between transmission events, for example that a female infected by an older male may be more likely to transmit to men of older age. For computational efficiency, we consider the related Poisson model

$$p(\mathbf{z} | \boldsymbol{\lambda}) \propto \prod_{a,b} (\lambda_{ab})^{z_{ab}} \exp(-\lambda_{ab}) = \left[\prod_{a,b} (\pi_{ab})^{z_{ab}} \right] \left[\eta^{z^+} \exp(-\eta) \right], \quad (5.10)$$

where λ_{ab} can be interpreted as the transmission intensities from group a to group b , $\eta = \sum_{c,d} \lambda_{cd}$, and π_{ab} are recovered via $\pi_{ab} = \lambda_{ab} / \eta$. Equation (5.10) follows from the Poisson trick (Lee et al., 2017) which relates the Poisson distributions to the multinomial distribution

via

$$\begin{aligned} p(\mathbf{z}|z^+, \boldsymbol{\pi}) &= \frac{z^+!}{\prod_i z_i!} \prod_i \pi_i^{z_i} \\ &\propto \prod_i \frac{\lambda_i^{z_i}}{\lambda^{+z_i}} = \frac{\prod_i \lambda_i^{z_i}}{\lambda^{+z^+}} = \frac{\prod_i \lambda_i^{z_i} \exp(-\lambda_i)}{\lambda^{+z^+} \exp(-\lambda^+)}, \end{aligned} \quad (5.11)$$

where $\pi_i = \lambda_i / \lambda^+$, which suggests the product of Poisson distributions could be factored into the product of a Poisson distribution of the overall count and a multinomial distribution of subcounts given the total. As z^+ carries no information on $\boldsymbol{\pi}$, it is equivalent to make inferences of $\boldsymbol{\pi}$ under the multinomial distribution and of $\boldsymbol{\lambda}$ under the product of Poisson distributions.

The actual transmission flows $\mathbf{z}$ are not observed. We assume that individuals are sampled at random within strata (SARWS) in a suitably chosen stratification of the population. SAWRS implies in particular that sampling is independent of transmission status (i.e. being a 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 transmission counts z_{ab} can be conveniently integrated out, yielding for the observed flow counts the thinned Poisson model

$$p(\mathbf{n}|\boldsymbol{\lambda}, \boldsymbol{\xi}) = \prod_{a,b} (\lambda_{ab} \xi_a \xi_b)^{n_{ab}} \exp(-\lambda_{ab} \xi_a \xi_b). \quad (5.12)$$

To understand the basic property of (5.12), we calculate the maximum likelihood estimates (MLE) of $\boldsymbol{\pi}$,

$$\hat{\pi}_{ab} = \frac{n_{ab}}{\hat{\xi}_a \hat{\xi}_b} / \sum_{a,b} \frac{n_{ab}}{\hat{\xi}_a \hat{\xi}_b}, \quad \hat{\xi}_a = \frac{N_a^s}{N_a}. \quad (5.13)$$

(5.13) aligns with the expectation where the flows are calculated from the adjusted counts for sampling proportions, and the proportions are evaluated from sampling associated data.

Next, we will discuss the specification of all prior distributions. Considering population sampling, we exploit additional information on the sampling vector $\mathbf{s}$. We assume that transmission is independent of sampling, allowing us to decompose the joint posterior distribution into

$$\begin{aligned} p(\boldsymbol{\lambda}, \boldsymbol{\xi}|\mathbf{n}, \mathbf{s}, \mathbf{X}) &\propto p(\mathbf{n}|\boldsymbol{\lambda}, \boldsymbol{\xi}, \mathbf{s}, \mathbf{X}) p(\boldsymbol{\lambda}|\boldsymbol{\xi}, \mathbf{s}, \mathbf{X}) p(\boldsymbol{\xi}|\mathbf{s}, \mathbf{X}) \\ &= p(\mathbf{n}|\boldsymbol{\lambda}, \boldsymbol{\xi}) p(\boldsymbol{\lambda}|\boldsymbol{\xi}) p(\boldsymbol{\xi}|\mathbf{s}, \mathbf{X}). \end{aligned} \quad (5.14)$$

A possible limitation of (5.13) is that the counts N_a , N_a^s can be small when the population is finely stratified. It is thus often advantageous to model individual-level sampling probabilities in terms of a linear combination of predictors. Using a logistic regression approach, we obtain

$$p(\xi_a | \mathbf{s}, \mathbf{X}) = \int \text{logit}^{-1}(\mathbf{x}_a \beta) p(\beta | \mathbf{s}, \mathbf{X}) d\beta, \quad (5.15)$$

where $\mathbf{x}_a$ is the row vector of population characteristics that specify group a , β are the regression coefficients, and $p(\beta | \mathbf{s}, \mathbf{X})$ is the posterior density of the regression coefficients, estimated from the sampling status vector $\mathbf{s}$ of all individuals in the study population. Details on how we model the sampling cascade of transmission events are given in Appendix A.

Considering the prior density on the transmission intensities λ , we propose using

$$\lambda_{ab} | \xi_a, \xi_b \sim \text{Gamma}(\alpha_{ab}, \beta), \quad \alpha_{ab} = 0.8/L, \quad \beta = 0.8/z^{+P}(\xi_a, \xi_b), \quad (5.16)$$

where L is the length of the flow vector π , which is equivalent to the number of structurally non-zero entries in the flow matrices (5.1), and z^{+P} is the number of expected transmission events,

$$z^{+P} = \sum_{a,b:n_{ab} \neq 0} \frac{n_{ab}}{\xi_a \xi_b} + \sum_{a,b:n_{ab}=0} \frac{1 - \xi_a \xi_b}{\xi_a \xi_b}. \quad (5.17)$$

This choice is motivated by the fact that (5.16) induces on π an objective Dirichlet prior density with parameters $\alpha_{ab} = 0.8/L$ (Berger et al., 2015), such that the likelihood (5.12) dominates the prior (5.16) regardless of the number of flows to estimate.

The Dirichlet distribution is the conjugate prior for multinomial distribution, and it could be expressed as the product of Gamma distributions,

$$\begin{aligned} p(\pi | \alpha) &= \frac{\prod_i \Gamma(\alpha_i)}{\Gamma(\sum_i \alpha_i)} \prod_i \pi_i^{\alpha_i-1} \\ &\propto \frac{\prod_i \lambda_i^{\alpha_i-1}}{\lambda^{+\alpha-K}} = \frac{\prod_i \lambda_i^{\alpha_i-1} \exp(-\beta \pi_i)}{\lambda^{+\alpha-K} \exp(-\beta)} \propto \prod_i \lambda_i^{\alpha_i-1} \exp(-\beta \pi_i). \end{aligned} \quad (5.18)$$

As the objective prior of multinomial distribution has the parameters $\alpha_{ab} = 0.8/L$ as suggested by Berger et al. (2015) and $\text{Dirichlet}(\pi | \alpha)$ could be expressed as the product of $\text{Gamma}(\lambda_{ab} | \alpha_{ab}, \beta)$, we used the prior $\lambda_{ab} \sim \text{Gamma}(\lambda_{ab} | \alpha_{ab}, \beta)$, $\alpha_{ab} = 0.8/L$ for the independent Poisson distributions. However, β is not specified through this one-to-one correspondence between independent Gamma and Dirichlet parameters.

Then the value of β is determined by the average intensities. That is,

$$\frac{\sum_{a,b} \alpha_{ab}}{\beta} = \sum_{a,b} \frac{n_{ab}}{\xi_a \xi_b}. \quad (5.19)$$

The problem of (5.19) is that zero observed counts produce zero actual transmission counts after the sampling adjustment, so it might underestimate the actual intensities. To address this issue, we adjusted (5.19) to handle zero observations separately. Assume $n \sim \text{Binomial}(z, s)$, then

$$p(n=0|z, s) = \frac{z!}{0!(z-0)!} s^0 (1-s)^{z-0} = z!(1-s)^z, \quad z = 0, 1, \dots \quad (5.20)$$

Then the probability mass function of the actual count z given no observations is

$$p(z) = \frac{(1-s)^z}{\sum_{z=0}^{\infty} (1-s)^z} = \frac{(1-s)^z}{1/(1-s)} = (1-s)^{z+1}, \quad z = 0, 1, \dots, \quad (5.21)$$

which is in the form of the negative binomial distribution, $\text{NegativeBinomial}(0, 1-s)$. Therefore, the expectation of the actual count given no observations is

$$\mathbb{E}[z] = \sum_{z=0}^{\infty} z(1-s)^{z+1} = \sum_{z=1}^{\infty} z(1-s)^z = \sum_{z=0}^{\infty} z(1-s)^z - 0 = \frac{1-s}{s}. \quad (5.22)$$

Therefore, we adapted (5.19) to

$$\frac{\sum_{a,b} \alpha_{a,b}}{\beta} = \sum_{a,b:n_{a,b}=0} \frac{1 - \xi_a \xi_b}{\xi_a \xi_b} + \sum_{a,b:n_{a,b} \neq 0} \frac{n_{ab}}{\xi_a \xi_b}, \quad (5.23)$$

and solved for β .

5.3.2 Numerical inference with Stan

The transmission flows and model parameters can be efficiently estimated with the Hamiltonian Monte Carlo (HMC) sampler of 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. We have used the no-U-turn sampler (NUTS) HMC algorithm in Stan version 2.21. The implementation is available at https://github.com/BDI-pathogens/phyloscanner/blob/master/phyloflows/inst/misc/gamma_prior.stan and Appendix E.2.

We focus on inference of the posterior distribution (5.14) with the objective prior density (5.16) attached to the transmission intensities, i. e.

$$\begin{aligned}
 p(\lambda, \xi | \mathbf{n}, \mathbf{s}, \mathbf{X}) &\propto p(\mathbf{n} | \lambda, \xi, \mathbf{s}, \mathbf{X}) p(\lambda | \xi, \mathbf{s}, \mathbf{X}) p(\xi | \mathbf{s}, \mathbf{X}) \\
 &= p(\mathbf{n} | \lambda, \xi) p(\lambda | \xi) p(\xi | \mathbf{s}, \mathbf{X}) \\
 &= \left(\prod_{a,b} \text{Poisson}(n_{ab}; \lambda_{ab} \xi_a \xi_b) \right) \left(\prod_{a,b} \text{Gamma}(\lambda_{ab}; \alpha_{ab}, \beta) \right) p(\xi | \mathbf{s}, \mathbf{X}),
 \end{aligned} \tag{5.24}$$

where the flow vector π is obtained from the transformation $\pi_{ab} = \lambda_{ab} / (\sum_{c,d} \lambda_{cd})$.

To specify the posterior density $p(\xi | \mathbf{s}, \mathbf{X})$ in Stan, we further approximated its components $p(\xi_a | \mathbf{s}, \mathbf{X})$ through a suitable closed-form density. In our applications, we opted for Beta densities with shape and rate parameters estimated via maximum likelihood approach, implemented by function `fitdistr` in R and compared the fitted distributions with the samples. The use of the Stan model is illustrated in the online example script https://github.com/BDI-pathogens/phyloscanner/blob/phyloflow_v120/phyloflows/vignettes/07_age_analysis.Rmd.

5.3.3 Numerical inference with a custom Markov Chain Monte Carlo algorithm

For comparison, we have also implemented a tailored Metropolis-within-Gibbs algorithm, in fact before we moved to use Stan. The algorithm is also implemented in the `phyloflows` version 1.2.0. The algorithm exploits the factorisation of the posterior density (5.24) into the full conditionals

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

where ξ_a is the sampling probability of the a th population group, λ_{ab} denotes the average transmission counts from population group a to population group b , $\xi_{-a} = \xi \setminus \xi_a$ and $\lambda_{-ab} = \lambda \setminus \lambda_{ab}$. The density $p(\xi_a | \xi_{-a}, \lambda, \mathbf{s}, \mathbf{n}, \mathbf{X})$ is not available in closed form, and we performed Metropolis-within-Gibbs updates for each ξ_a . The density $p(\lambda_{ab} | \xi, \lambda_{-ab}, \mathbf{s}, \mathbf{n}, \mathbf{X})$ is of Gamma form, and so a Gibbs step can be used to update λ_{ab} . The resulting Markov Chain Monte Carlo (MCMC) algorithm iterates through Metropolis-within-Gibbs updates for each ξ_a , followed by a series of Gibbs updates for each λ_{ab} .

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

$$\begin{aligned} p(\xi_a | \xi_{-a}, \lambda, \mathbf{n}, \mathbf{s}, \mathbf{X}) & \propto \prod_{c=a \text{ or } d=a} p(n_{cd} | \lambda_{cd}, \xi_c \xi_d) p(\xi_a | \mathbf{s}, \mathbf{X}) \\ & \propto \prod_{c=a \text{ or } d=a} \text{Poi}(n_{cd}; \lambda_{cd} \xi_c \xi_d) \text{Gamma}(\lambda_{cd}; \xi_c \xi_d) p(\xi_a | \mathbf{s}, \mathbf{X}), \end{aligned} \quad (5.25)$$

which does not have a 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,

$$q(\xi'_a | \xi_a) = p(\xi'_a | \mathbf{s}, \mathbf{X}). \quad (5.26)$$

The resulting Metropolis Hasting ratio is

$$\begin{aligned} r &= \frac{p(\xi'_a | \xi_{-a}, \lambda, \mathbf{n}, \mathbf{s}, \mathbf{X})}{p(\xi_a | \xi_{-a}, \lambda, \mathbf{n}, \mathbf{s}, \mathbf{X})} \times \frac{q(\xi_a)}{q(\xi'_a)} \\ &= \frac{\prod_b p(n_{ab} | \lambda_{ab}, \xi'_a \xi_b) \prod_{b \neq a} p(n_{ba} | \lambda_{ba}, \xi_b \xi'_a) \prod_b p(\lambda_{ab} | \xi'_a) \prod_{b \neq a} p(\lambda_{ba} | \xi'_a) p(\xi'_a | \mathbf{s}, \mathbf{X})}{\prod_b p(n_{ab} | \lambda_{ab}, \xi_a \xi_b) \prod_{b \neq a} p(n_{ba} | \lambda_{ba}, \xi_b \xi_a) \prod_b p(\lambda_{ab} | \xi_a) \prod_{b \neq a} p(\lambda_{ba} | \xi_a) p(\xi_a | \mathbf{s}, \mathbf{X})} \\ & \quad \times \frac{p(\xi_a | \mathbf{s}, \mathbf{X})}{p(\xi'_a | \mathbf{s}, \mathbf{X})} \\ &= \frac{\prod_b p(n_{ab} | \lambda_{ab}, \xi'_a \xi_b) \prod_{b \neq a} p(n_{ba} | \lambda_{ba}, \xi_b \xi'_a) \prod_b p(\lambda_{ab} | \xi'_a) \prod_{b \neq a} p(\lambda_{ba} | \xi'_a)}{\prod_b p(n_{ab} | \lambda_{ab}, \xi_a \xi_b) \prod_{b \neq a} p(n_{ba} | \lambda_{ba}, \xi_b \xi_a) \prod_b p(\lambda_{ab} | \xi_a) \prod_{b \neq a} p(\lambda_{ba} | \xi_a)}, \end{aligned} \quad (5.27)$$

where for brevity we denoted by ξ' 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_a^s and ξ_a^r . Equations (5.25-5.27) generalise straightforwardly to this setting, and the algorithms update the sampling probabilities ξ^s and ξ^r in turn.

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

$$\begin{aligned}
 p(\lambda_{ab}|\xi, \lambda_{-ab}, \mathbf{s}, \mathbf{n}, \mathbf{X}) &\propto p(\lambda_{ab})p(n_{ab}|\lambda_{ab}, \xi_a, \xi_b) \\
 &\propto \exp(-\lambda_{cd}\xi_c\xi_d)(\lambda_{cd}\xi_c\xi_d)^{n_{cd}}\lambda_{cd}^{\alpha_{cd}-1}\exp(-\beta\lambda_{cd}) \\
 &\propto \exp(-\lambda_{cd}(\xi_c\xi_d + \beta))\lambda_{cd}^{n_{cd}+\alpha_{cd}-1} \\
 &\sim \text{Gamma}(n_{cd} + \alpha_{cd}, \xi_c\xi_d + \beta).
 \end{aligned} \tag{5.28}$$

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

The inferences on the Multinomial-Dirichlet model (5.9) are easier via the Poisson-Gamma model (5.10). The posterior distribution is $p(\mathbf{z}, \boldsymbol{\pi}, \boldsymbol{\xi}|\mathbf{n})$ under the multinomial likelihood and Dirichlet prior, which requires update ξ_i and related $\mathbf{z}_i$ in a Metropolis step, followed by Gibb sampling of $\boldsymbol{\pi}$. The Gibb step is straightforward due to conjugacy, but the Metropolis step may involve a large number of correlated parameters per iteration, which may lead to the mixing issues. The posterior is $p(\mathbf{z}, \boldsymbol{\lambda}, \boldsymbol{\xi}|\mathbf{n})$ under the Poisson likelihoods and Gamma priors, and $\mathbf{z}$ is then marginalised out by Poisson thinning property. This can be inferred by sweeping over $\boldsymbol{\xi}$ and $\boldsymbol{\lambda}$, which avoids updating massive correlated latent variables during the Metropolis step and thus makes it easier for inference.

5.4 Results

We conduct several simulation experiments in Section 5.4.1 to Section 5.4.4 to first demonstrate the effect of sampling heterogeneity in the transmission flow inference and validate the performance of our approach. Next we apply this method to the Rakai Community Cohort Study (RCCS) in order to analyse the transmissions between the hotspot and the general population in Section 5.4.5 - Section 5.4.7.

5.4.1 ODE based simulations

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. We considered transmission flows between two population groups $\mathcal{A} = (a, b)$, 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. Transmission chains were simulated under the ODE model (5.3), i.e. not from our simpler likelihood model (5.12), and the simulated flow

Table 5.1 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

matrix π_r^0 was recorded in $r = 1, \dots, 100$ replicate simulations. We then simulated epidemics based on the compartmental model (5.3) among susceptible, infected, and treated individuals in the four population strata. The parameters of the model were specified such that 40% of them were in the group a and 60% were in the group b ; half of the 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/phyloflow_v120/phyloflows/inst/misc/Master.xml. Table 5.1 lists the model parameters that we used in the simulations, where in terms of the notation, we replaced the population subscripts l with a and h with b compared to (5.3). Figure 5.2A illustrates an example of epidemic trajectories in one simulation.

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 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

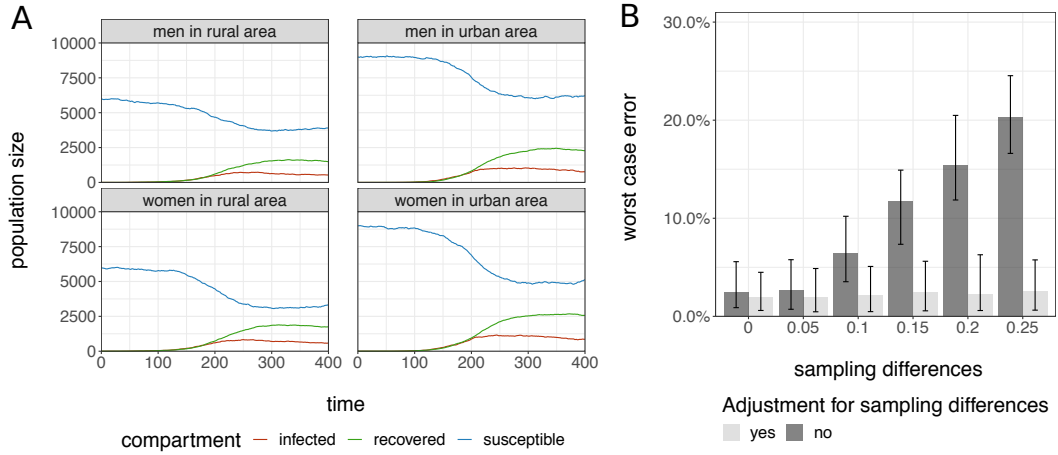


Figure 5.2 Simulation experiments to assess the accuracy of flow estimates under sampling bias. (A) ODE-based models were used to simulate epidemic trajectories of susceptible, infected and treated men and women across two population strata, and transmission flows between them. Panel A visualises one of the simulated trajectories. (B) Transmission flows were estimated under the Poisson likelihood model (5.12), without adjustments for sampling differences (dark grey), and with adjustments for sampling differences (light grey). Accuracy was measured with the WCE between posterior median estimates and the simulated true values, and shown are the average error and 95% range in 100 replicate simulations.

draw with probability ξ_a , and respectively for each individual in b with probability ξ_b , and observed transmissions n_{ab}^r from group a to group b were calculated as in (A.1).

From the simulated data, we first re-estimated the transmission flows while ignoring the sampling heterogeneity. We estimated the posterior distribution (5.14) with the `phyloflows` MCMC-within-Gibbs sampler in Section 5.3.3, with $p(\xi_a|\mathbf{s}, \mathbf{X})$ and $p(\xi_b|\mathbf{s}, \mathbf{X})$ set to the Beta distribution with parameters 25.5 and 25.5. The prior followed the Jeffreys prior and the density is shown in the grey region in Figure 5.3. Thus, the specified sampling densities ignored the sampling heterogeneity. The code is available at https://github.com/BDI-pathogens/phyloscanner/blob/phyloflow_v120/phyloflows/inst/misc/ode_estimation.R. For each simulation, we then calculated the worst case error (WCE) that is defined as $\epsilon_r = \max_{a,b} |\hat{\pi}_{ab,r} - \pi_{ab,r}|$ where r is the index of replicates. The main results are shown in the black bars in Figure 5.2B and indicate that the WCE increased substantially with sampling heterogeneity.

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

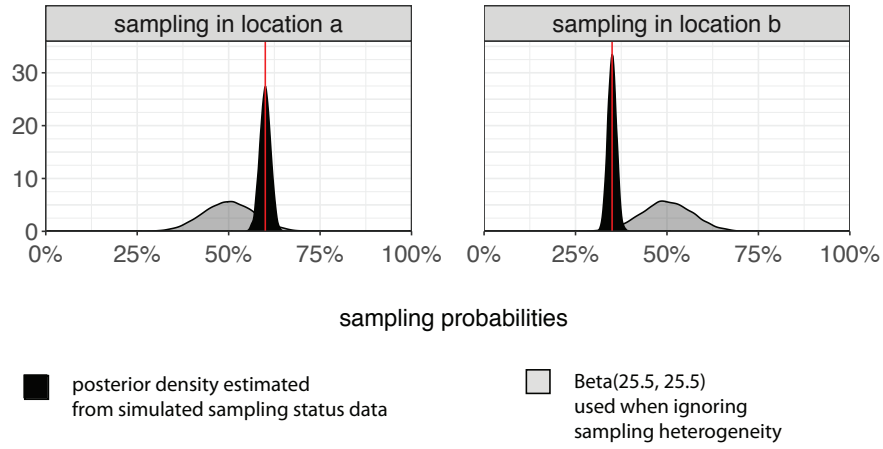


Figure 5.3 **Re-estimated sampling probabilities in the ODE-based simulations.** We simulated transmission flows between men and women in two locations a and b under the ODE model (5.3). From the simulated sampling status data, posterior estimates of the sampling proportions in a and b were obtained with Equation (5.29), and then used in flow inference. We show posterior densities of the sampling probabilities in a and b in black, compared to a $\text{Beta}(25.5, 25.5)$ prior density in grey. Vertical lines indicate the true sampling fractions, $\xi_a = 0.6$ and $\xi_b = 0.35$.

size N_a , through

$$p(\xi_a | \mathbf{s}, \mathbf{X}) = \text{Beta}(\xi_a, N_a^s + \alpha_\xi, N_a - N_a^s + \beta_\xi), \quad (5.29)$$

where we set $\alpha_\xi = 0.5$ and $\beta_\xi = 0.5$ under the Jeffreys prior on the Binomial sampling probabilities. The density $p(\xi_b | \mathbf{s}, \mathbf{X})$ was specified analogously. Figure 5.3 illustrates the estimated posterior sampling distributions, in comparison to the true values. We then estimated the posterior distribution (5.14) with the phyloflows MCMC-within-Gibbs sampler of Chapter 5, using the posterior distribution of sampling probabilities (5.29). The code is available at https://github.com/BDI-pathogens/phyloscanner/blob/phyloflow_v120/phyloflows/inst/misc/ode_estimation.R. Figure 5.2B compares the accuracy in the posterior median flow estimates in terms of the WCE, and illustrates that the Poisson model (5.14) can minimise the impact of sampling heterogeneity on flow estimates. More complex simulation experiments yielded similar results in Section 5.4.3.

In addition to the WCE, we report other performance measures that were defined in Section 4.4, including 95% posterior predictive credibility interval (PPCrI), Continuous Ranked Probabilistic Score (CRPS), Expected Log pointwise Predictive Density (ELPD) and model evidence. These measures are presented in Table 5.2. We observed no significant differences in the first three measures. The potential reason is these assessments on counts instead of

Table 5.2 Additional model performance measures in the ODE based simulation.

Measure	Adjust	35%	40%
Data in	yes	1 (1 - 1)	1 (1 - 1)
95% PPCrI	no	1 (1 - 1)	1 (1 - 1)
CRPS	yes	0.04 (0.01 - 0.09)	0.04 (0.01 - 0.09)
	no	0.04 (0.01 - 0.09)	0.04 (0.01 - 0.09)
ELPD	yes	-16.59 (-17.04 - -16.22)	-16.78 (-17.04 - -16.22)
	no	-16.64 (-17.01 - -16.14)	-16.81 (-17.01 - -16.14)
Marginal likelihood	yes	-643.63 (-723.85 - -565.73)	-566.8 (-723.85 - -565.73)
	no	-246.68 (-274.4 - -221.59)	-255.04 (-274.4 - -221.59)
45%	50%	55%	60%
1 (1 - 1)	1 (1 - 1)	1 (1 - 1)	1 (1 - 1)
1 (1 - 1)	1 (1 - 1)	1 (1 - 1)	1 (1 - 1)
0.04 (0.01 - 0.09)	0.04 (0.01 - 0.09)	0.04 (0.01 - 0.09)	0.04 (0.01 - 0.09)
0.04 (0.01 - 0.09)	0.04 (0.01 - 0.09)	0.04 (0.01 - 0.09)	0.04 (0.01 - 0.09)
-16.97 (-17.04 - -16.22)	-17.16 (-17.04 - -16.22)	-17.09 (-17.04 - -16.22)	-17.17 (-17.04 - -16.22)
-16.98 (-17.01 - -16.14)	-17.15 (-17.01 - -16.14)	-17.09 (-17.01 - -16.14)	-17.14 (-17.01 - -16.14)
-500.95 (-723.85 - -565.73)	-437.27 (-723.85 - -565.73)	-358.78 (-723.85 - -565.73)	-302.78 (-723.85 - -565.73)
-271.64 (-274.4 - -221.59)	-284.35 (-274.4 - -221.59)	-284.48 (-274.4 - -221.59)	-291.03 (-274.4 - -221.59)

flows. These inspections thereby reflect the good fits on the counts regardless of sampling adjustment. Note that the ELPDs may not be reliable due to slightly large parameters in the Pareto distribution (0.7 - 1). It might be fixed by the recently developed option in `loo` (Paananen et al., 2021). Moreover, we are able to evaluate model evidence through `bridgesampler` in this example. However, we find evidence towards the no sampling adjustment. It might be due to the fact that model evidence integrates the likelihood over the prior density, and naturally penalises the complex models. In contrast to the model with no sampling adjustment, the adjusted model includes far more parameters and is intrinsically less favoured. We will not consider these additional measures and will focus on the MAE and WCE in the rest of the simulations.

5.4.2 Sensitivity analysis of the overall sample size of observed transmissions

We further assessed the accuracy of the flow inference as a function of the overall sample size of the observed transmission events, $n^+ = \sum_{a,b} n_{ab}$. We re-considered the ODE-based simulation experiments of Section 5.4.1, 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 5.4 (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 5.4

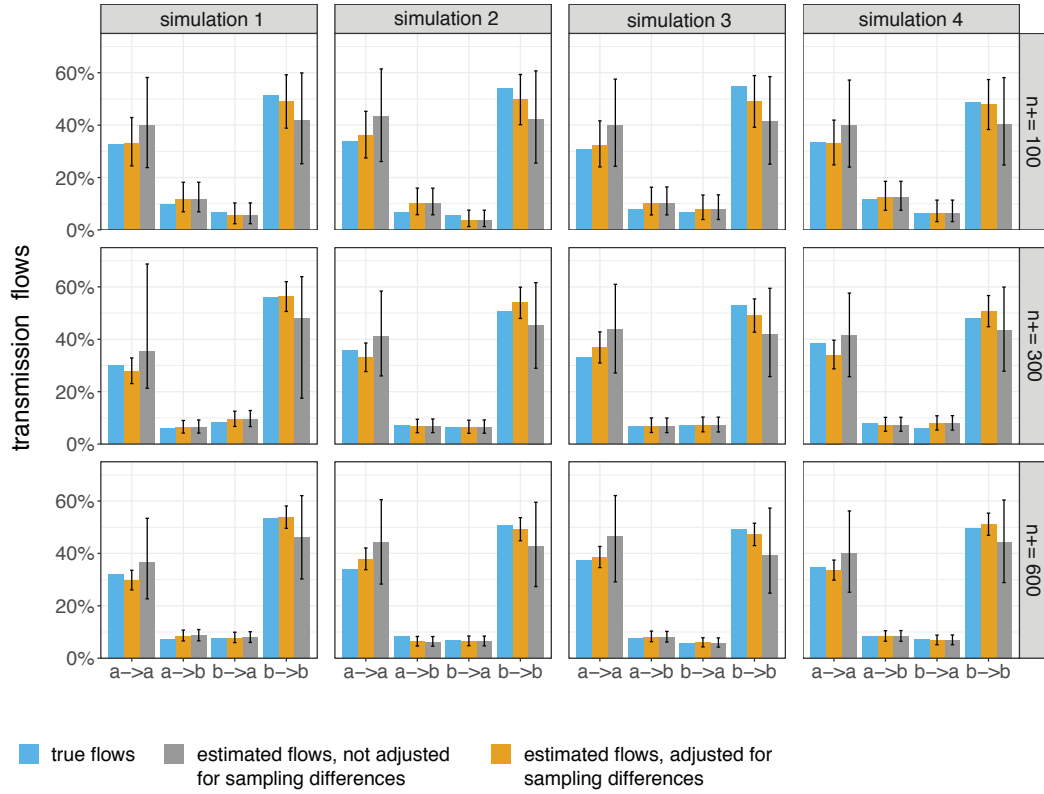


Figure 5.4 **True and re-estimated flows in four ODE-based simulations, varying in sample size.** We simulated transmission flows between men and women in two locations a and b under the ODE model (5.3) in Section 5.4.2. Columns show results for four simulated data sets, and rows show results for increasing samples size. The true transmission flows are shown 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 (5.29) (orange).

(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 (5.29). There was considerable variability in the flow data sets when the sample size was low ($n^+ = 100$). However, information on population sampling substantially improved flow estimates regardless of sample size. Figure 5.5 summarises the WCE in these simulations. When sampling heterogeneity was not accounted for, variability in flow estimates decreased with increasing sampling size, but error magnitude did not decrease. When sampling heterogeneity was accounted for, both the variability in flow estimates and error magnitude decreased with increasing sampling size.

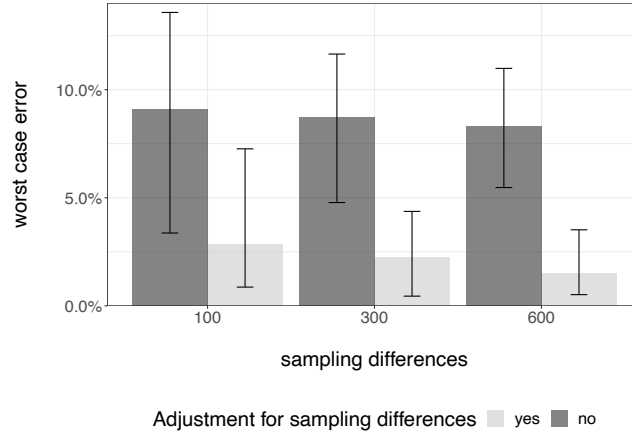


Figure 5.5 Worst case error in estimated transmission flows with increasing sample size. We simulated transmission flows between men and women in two locations a and b under the ODE model (5.3) as described in Section 5.4.1. Transmission flows were re-estimated with the hierarchical Poisson model (5.12) 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 WCE 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.

5.4.3 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 Appendix A.

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) (Ratmann et al., 2020), and simulated 576 transmission flows between the male-female and female-male sub-group combinations. For simplicity, we did not generate epidemic trajectories as in Section 5.4.1 and instead simulated the transmission counts n_{ab} from the pre-specified transmission flows π^0 and sampling probabilities ξ_a which were estimated from the RCCS. For a fixed target sample size n^+ , the total number of actual transmissions z^+ were simulated from $z^+ \sim \text{Poisson}(n^+/\bar{\xi})$, where $\bar{\xi} = \frac{1}{24} \sum_a \xi_a$ was the average pairwise sampling probability in the population. Then, the actual transmission counts between population groups were simulated from $\mathbf{z} \sim \text{Multinomial}(z^+, \pi_0)$. The sampling status of each individual in a was simulated under a Bernoulli draw with probability ξ_a ,

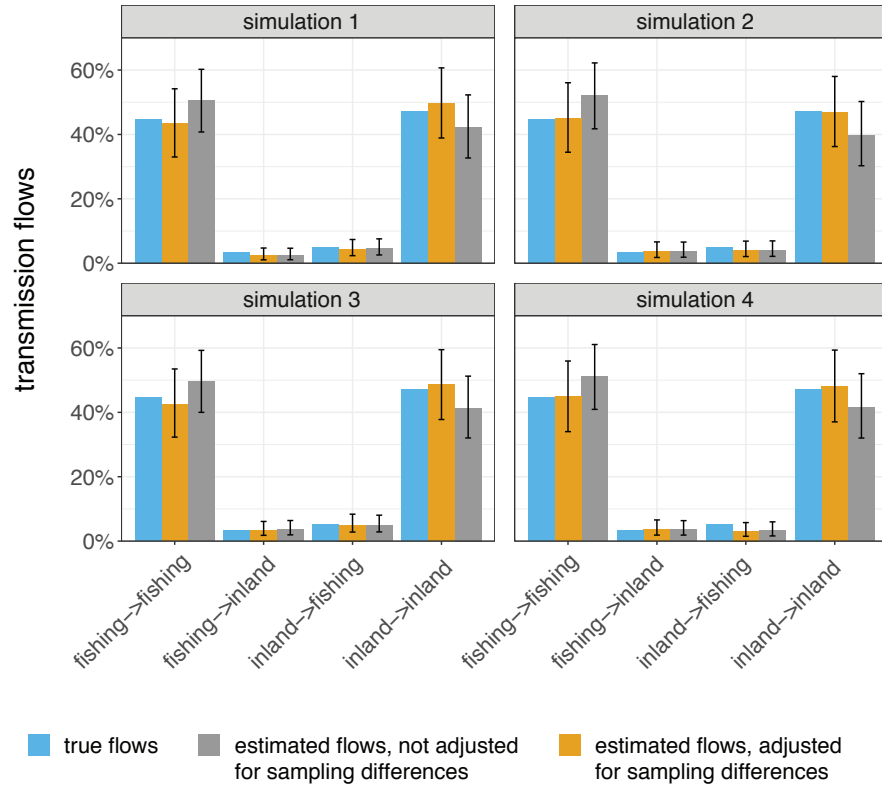


Figure 5.6 True and re-estimated flows in more complex simulations between 24 population groups. We simulated transmission flows by gender, 3 age bands, and location under model parameters similar to those obtained for the Rakai analysis; see Section 5.4.3. Panels show results for four simulated data 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 (5.29) (orange).

and respectively in population b with probability ξ_b , and observed transmissions n_{ab} from group a to group b were calculated as in (5.6). 100 data sets were simulated following the procedure, 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 the sine function in $[-\pi, \pi]$ to obtain a wider range of average sampling differences between inland and fishing communities from 5% to 25% while retaining the structured proportions. Then, 100 simulations were generated under each of the sampling scenarios.

From the simulated data, we re-estimated the 576-dimensional transmission flows while ignoring the sampling heterogeneity, and then while accounting for the sampling heterogeneity as described in Section 5.4.1. To facilitate the comparison, we aggregated the true

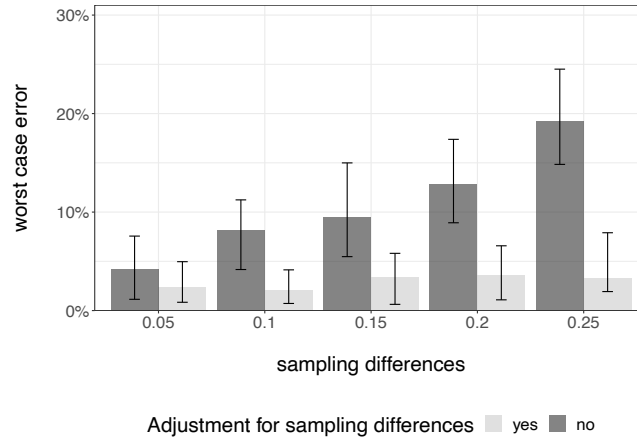


Figure 5.7 Worst case error in estimated transmission flows in more complex simulations of transmission flow between 24 population groups. We simulated transmission flows between men and women in 24 population groups as described in Section 5.4.3. Simulation scenarios varied in average sampling differences 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, WCE was reported, after aggregating flow estimates to the 4 flow combinations 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 WCE were similar to those found in the more simple ODE-based simulation experiments.

and re-estimated flows into the 4-dimensional vector of flows within and between inland and fishing communities and calculated the WCEs. Figure 5.6 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 5.7 shows that in these more complex simulation scenarios, the trend of WCEs was overall similar compared to the more simple ODE-based simulation experiments in Section 5.4.1. When ignoring sampling heterogeneity, the WCE increased considerably with the average sampling differences between inland and fishing communities. When accounting for the sampling heterogeneity, the WCE remained generally unaffected by the increasing average sampling differences, though overall it was slightly higher compared to that in the ODE-based simulation experiments.

5.4.4 Inference method comparison

We made a comparison of two inference methods in Section 5.3.2 and Section 5.3.3, HMC via Stan and the custom MCMC. We first took a ODE-based simulation under typical

Table 5.3 Performance of MCMC and HMC under the Gamma prior.

Method	Runtime (minutes)	Minimum ESSs
MCMC	6.15 (5.23 - 7.26)	23212.69 (17255.2 - 26016.33)
HMC	1.5 (1.21 - 1.87)	37204.04 (35541.39 - 39032.21)
MAE	WCE	Data in 95% PPCrI
0.01 (0 - 0.02)	0.02 (0 - 0.04)	1 (1 - 1)
0.01 (0 - 0.02)	0.02 (0 - 0.04)	1 (1 - 1)

sampling differences in real data, i.e. 10%, from Section 5.4.1, estimated the flows with adjustment for sampling heterogeneity, and compared the sampler efficiency and accuracy measures in Table 5.3. In both samplers, we implemented for 10^5 iterations and removed the first 2×10^4 warm-up steps. Table 5.3 illustrates Stan is faster and achieves higher effective sample sizes (ESS) for a fixed length of chains, though no dependence of accuracy on the inference approach is observed.

However, we will keep the MCMC algorithm as a part of `phyloflows`, as it appears more robust. For instance, Stan generated the chain that does not converge in the application to the RCCS data (Section 5.4.7) and the convergence diagnostics $\hat{R}$ is 2.36. Therefore, we used the custom MCMC sampler in the RCCS application, and did not experience the convergence and mixing issues. In practice, we would recommend Stan implementation first due to its speed and efficacy, and one may adopt the custom MCMC sampler whenever the sampler in Stan generates the chains that do not converge.

5.4.5 Modelling the sampling cascade of transmission events

In this Section, we briefly introduce the overall sampling patterns and how we characterised the sampling proportions in RCCS. 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 (Appendix A). 11404 (96.9%) of non-participants were absent for school or work. Next, infected individuals who did not report Antiretroviral Therapy (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 sites (Appendix A).

The estimated infection times of recipients in the phylogenetically reconstructed source-recipient pairs are between 2009-10-01 and 2015-01-30, which first defined the study period $\mathcal{T}$ during which we estimated transmission flows. To characterise these observations, we defined as denominator transmission events to census-eligible individuals in RCCS

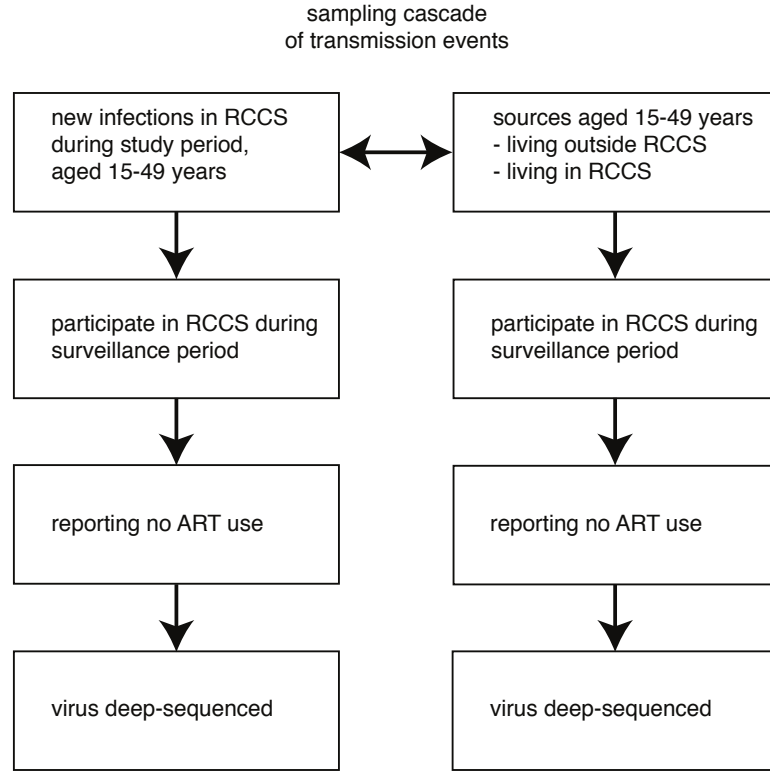


Figure 5.8 **Sampling cascade of transmission events.** The sampling cascade formalises the steps involved in sampling sources and recipients of transmission events. Recipients were defined as individuals aged 15-49 years who acquired infection in one of the RCCS communities during the study period. 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.

communities who were infected during the study period $\mathcal{T}$, and formalised the individual steps in the sampling cascade of transmission events in Figure 5.8. As aforementioned in Chapter 2, each survey in the RCCS was preceded 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 , $\forall a, b$. The detailed procedure is in Appendix A. The census data are available between 2011-08-10 and 2015-01-30, and we assume sampling keeps the same in a slightly longer period between 2009-10-01 and 2015-01-30.

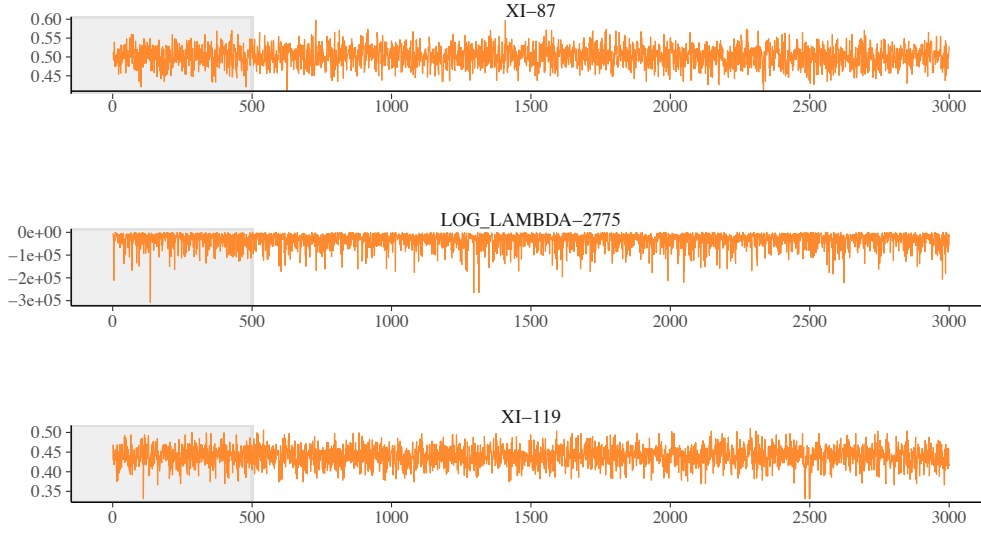


Figure 5.9 Trace plots for parameter inference on the Rakai data set, Poisson flow model with Gamma prior on transmission intensities. Stan’s dynamic HMC algorithm was run using the code for 10^4 iterations. Shown are trace plots for the first 3×10^3 iterations of the three parameters with the lowest ESSs.

5.4.6 Numerical performance on Rakai data

We describe the numerical performance of the Bayesian Poisson flow model as applied to the Rakai data. When using the Gamma prior (5.16), inference took 27 hours for 10^4 iterations of MCMC algorithm on a typical laptop. All ESSs were $\gg 5000$. Trace plots of the three parameters with the worst performance in terms of ESS are shown in Figure 5.9. They suggested no mixing and convergence problem.

5.4.7 Application to estimate transmission flows between areas with high and low disease prevalence in Rakai

We then used the source-recipient data to address problem (5.1) and estimate transmission HIV flows within and between high- and low-prevalence RCCS communities. The high-prevalence communities comprised the four fishing communities, and the low-prevalence communities comprised the remaining 36 inland communities. Here we focus on illustrating how the 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

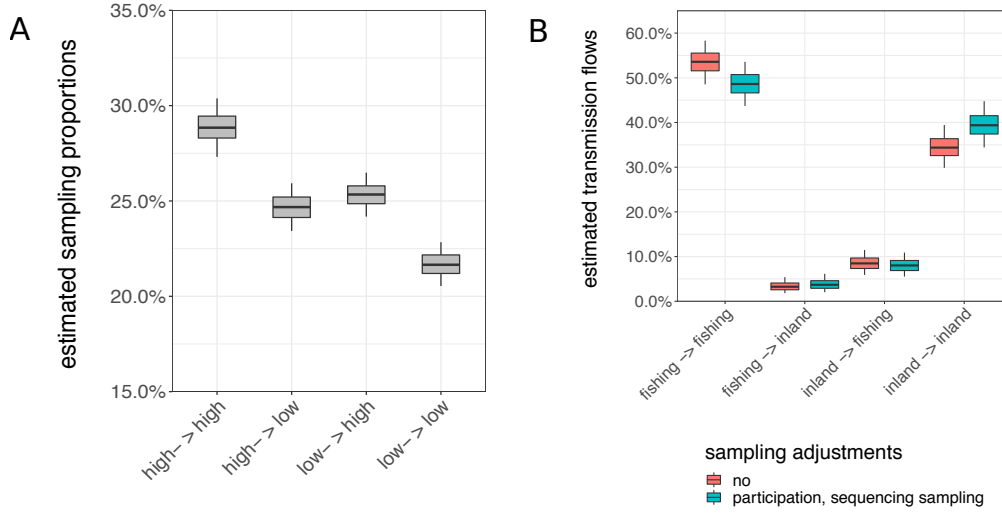


Figure 5.10 Estimated sampling probabilities of transmission events, and impact on flow estimates. (A) Boxplots of estimated pairwise sampling probabilities of transmission events between low-prevalence inland and high-prevalence fishing communities of the RCCS. The sampling probabilities were obtained 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.

that underlies the Poisson flow model (Section 5.3.1). Specifically, we stratified populations by gender, 1-year age bands (between 15 and 49 years), and resident locations (low or high prevalence), which resulted in 140 sampling groups. In fact, the age stratification is not necessarily by one-year as long as the SARWS assumption holds. We then sought to estimate transmission flows between the 2×35^2 age- and gender-specific transmission flows for each of the 4 combinations of geographic source and recipient locations, through the joint posterior distribution (5.14). We further accounted for geographic in-migration, resulting in 29,400 flow variables. On this high-dimensional flow space, we were able to directly apply the estimated, structured sampling probabilities that were calculated in Section 5.4.5. This shows that accounting for the observed heterogeneity in how the census population was sampled resulted in a substantially more complex inferential problem than (5.1) suggests.

Figure 5.10B shows the marginal posterior estimates of the aggregated flow vector $\pi = (\pi_{hh}, \pi_{hl}, \pi_{lh}, \pi_{ll})$, vectorised via Equation (5.1), when the 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 estimated flow ratio (inland→fishing / fishing→inland) was 2.15 (1.02-4.97) when the sampling heterogeneity was accounted for, and 2.58 (1.23-5.86) when the sampling heterogeneity was not accounted for. Contrary to what is commonly assumed (Uganda AIDS Commission, 2014), these findings indicate that in Rakai, Uganda, high-prevalence fishing communities were net sinks, and not sources, of local infection flows.

Next, we compare this analysis with an earlier version which has been reported in Ratmann et al. (2020). The earlier version focused on the relatively low dimensional transmission flows where population is stratified by 3 age bands, genders and 3 types of communities and immigration status, which avoids the low count per stratum and may be more suitable for estimating flows between inland and fishing communities. In addition, we did not distinguish the sampling fractions between sources and recipients, and considered all the viral unsuppressed individuals rather than the newly infected individuals as possible recipients. The earlier version reports 80.1% (75.1% – 84.6%) transmissions within lakeside or inland areas, 3.5% (1.7%–6.3%) from lakeside to inland and 7.8% (5.1%–11.3%) from inland to lakeside, 5.8% (3.5%–8.9%) from outside the Rakai region to fishing communities and 2.5% (1.0%–5.0%) from outside to inland communities. The thesis reports 46.5% (41.2%–52.2%) transmissions within lakeside areas, 34.7% (29.6%–39.9%) within inland areas, 3.3% (1.6%–5.9%) from lakeside to inland, 7.4% (4.7%–10.4%) from inland to lakeside, 5.4% (3.2%–8.3%) from outside to fishing, and 2.1% (0.8%–4.4%) from outside to inland. The estimated adjusted flow ratio changes from 2.25 (1.04 – 5.23) to 2.18 (1.06 - 4.71). Overall, no substantial differences are observed in these two discrete location analyses, and the ratio keeps significantly larger than 1.

5.5 Comparison with flow inference from consensus sequences

In this Section, we compare our approach using deep sequence data and the standard phylodynamics approach using the consensus sequence data, and discuss the benefits and drawbacks of our approach. A classical approach for inferring viral spatial evolution through Sanger sequences is "migration" (Lemey et al., 2009). The approach underpins the phylogeographical inference, but sampling heterogeneity is not usually incorporated (see Section 1.2). The approach has been applied to a neighbouring study operated by Medical Research

Council/ UK Research and Innovation (MRC/UKRI) Uganda Research Unit on Acquired immunodeficiency syndrome (AIDS) (Bbosa et al., 2019). We apply the similar migration model to the RCCS data and compare it with the deep sequences based method when sampling is not adjusted for. In addition, a recent success in addressing spatial sampling heterogeneity in phylodynamics was achieved by Kühnert et al. (2016) (see Section 5.2). We also attempt this approach and compare with our methods when sampling is taken into account.

To implement the standard phylodynamic methods, we build a phylogeny $\mathbf{F}$ (see Section 3.3.5) using the consensus sequences that were collected before January 2015 in Section 3.2, and fit the mathematical model to the phylogeny with the purpose of estimating the viral transmission rates between the demes including inland and fishing communities.

The varying evolutionary rates between subtypes motivate us to analyse them separately. In particular, we defined the subtype of a consensus sequence by the subtype of the closest reference sequence. We identified 1029 sequences of subtype A1 and 647 of subtype D. The previous analyses on Sanger sequence typically use partial fragments, and we focus on *gag* p24 Capsid between positions 1186 and 1881. To build a sensible phylogenetic tree, we only included the sequences that cover at least 80% positions, and kept 468 and 724 sequences of subtype A1 and D respectively. 170 of subtype A1 sequences were in inland and 223 were in fishing, while the numbers are 267 and 344 among subtype D. As a common practice in phylogenetics, we added the background sequences (see Section 3.3.3). FastTree (Price et al., 2010) is a similar software to IQTREE and RAxML in Section 3.3.5, and it combines several algorithms including neighbour-joining, minimum evolution and maximum likelihood to approximate the maximum likelihood phylogeny. In contrast to IQTREE and RAxML, FastTree is faster and works for large alignments. We applied FastTree version 2.1.10 to first construct a phylogeny using the subtype-specific RCCS consensus sequences and background sequences, without dividing them into potential transmission networks. To understand ancestral relationship, we rooted the phylogeny by the representative of subtype H which was limited to central Africa where HIV originated. In the subsequent analyses, we regard the phylogeny as fixed on which the evolutionary process is investigated.

5.5.1 Migration model

The geography is partitioned into J states. For example, we may be interested in {inland, fishing, unknown}, abbreviated as $\{f, i, u\}$. Assume the states of tips $\mathbf{S}$ are observed, the purpose is to infer the process along the tree branches that leads to the observed

tip states. Migration model describes the states along the branches by independent time-homogeneous Continuous-Time Markov Chain (CTMC) that is characterised by infinitesimal rate matrix $\mathbf{G}$ (Lemey et al., 2009). One may make inference on the posterior distribution $p(\mathbf{G}|\mathbf{X})$. The transition matrix $\mathbf{G}$ assumes $J(J-1)$ possible transitions, but indeed most are rare. To address this challenge, the authors integrated Bayesian stochastic search variable selection (BSSVS) to simultaneously determine the non-zero entries in $\mathbf{G}$ by the associated binary indicator δ .

We adopted most of the prior choices in Bbosa et al. (2019). In particular, we used the strict molecular clock model which assumed the evolutionary rates to be constant, and the prior on the clock rate was an uninformative uniform distribution between 0 and 10^{100} . We used Gamma(1, 1) prior on the asymmetric rate matrix $\mathbf{G}$, and Poisson(2) on the count of non-zero rates.

We inferred the posterior distribution in BEAST version 1.10.4, a platform for Bayesian analysis of molecular sequences using MCMC (Drummond and Rambaut, 2007). We ran MCMC for 10^8 iterations that were thinned to every 10^4 iterations. The resulting ESSs are over 7,000 for subtype A1 and over 1,000 for subtype D. If the states are one-year age bands by sex, which will be described in the next Chapter, neither of the analyses is possible within three days.

We summarise the posterior samples from the MCMC, and visualise the dispersion rates between inland and fishing communities in Figure 5.11, i. e. the entries of the generator matrix $\mathbf{G}$. The transition probability matrix from 0 to T , $\mathbf{P}(0, T) = \mathbf{P}(T)$, can be derived from Kolmogorov forward equation. We are thus able to evaluate the conditional flow ratio via the entries of $\mathbf{P}(T)$, and it differs from the flow ratio (5.2c). It is challenging to interpret this ratio because it does not account for the number of infectious individuals in each location, and it depends on a time variable that is not easily specified. We then compare this conditional flow ratio between "migration" model and deep sequence based method, when sampling heterogeneity is not accounted for. The conditional flow ratio is 2.4443 (1.1027 - 5.6778) in deep sequence based method, and in contrast it depends on the time in "migration" model. An ideal way is to track transition events backward in time and identify the time period when the transition occurred. We did not explore this option in the thesis but attempted several candidates of the time, including the observational period between October 2019 and January 2015, $T = 5 + 3/12$, and a longer period 10 years. Table 5.4 shows the median estimates of conditional flow ratios are > 1 but the 95% credibility intervals (CI) are wide, and the conditional ratio turns insignificantly different from 1. We are not surprised by this incongruence. Our estimates reflect the epidemic behaviour within a well-defined sample and

Table 5.4 Comparison of conditional flow ratios in "mugration" model and deep sequence based method.

Method	Time	Subtype A1	Subtype D
Deep sequence based	5.25	2.4443 (1.1027 - 5.6778)	
Mugration	5.25	1.3611 (0.2705 - 4.5116)	1.5731 (0.3264 - 4.091)
Mugration	10	1.3619 (0.2705 - 4.5144)	1.571 (0.3269 - 4.0858)

time period, that are directly supported by observed counts after resolving recent migration events, whereby individuals were labelled by their location of migration origin in a case of a recent migration event (Ratmann et al., 2020). The CTMC model measures transition rates over larger time periods through the shape of the phylogeny, and we were not able to resolve recent migration events to the same quality standard for all sequenced individuals, while at the same time we know that HIV infected individuals preferentially move to fishing sites (Grabowski et al., 2020; Ratmann et al., 2020). Other possible reasons may include the investigation of a partial genome segment as well as evidence on state transitions gathered from the entire phylogenetic tree.

Apart from the conditional flow ratio, we will then investigate standard quantities that are usually reported in the "mugration" model. We calculated the Bayes factor (BF) to test the non-zero transition rates with spread3 version 0.9.6, i.e. the competing hypothesis $H_0: G_{jk} \neq 0$ and $H_1: G_{jk} = 0$. For subtype A1 we found a non-zero transition rate from fishing to inland is decisively supported ($\text{BF} = 247 > 100$), and a non-zero transition rate from inland to fishing is strongly supported ($\text{BF} = 18.1620 \in (10, 100]$), based on the criteria in Kass and Raftery (1995). For subtype D we found the opposite, a non-zero transition rate from fishing to inland is not significantly supported ($\text{BF} = 2.2 < 3$), and a non-zero transition rate from inland to fishing is decisively supported ($\text{BF} = 828 > 100$). By comparison, Bbosa et al. (2019) focused on 385 sequences of pure subtype A1 and D, and found > 10 BF for general population to fisherfolk viral migration and < 3 BF for the opposite direction. This coincides with our outcome based on subtype D that are more prevalent.

A suspicious observation in this analysis is the inconsistency between transition rates and BF on subtype A1. The median rate of transition from inland to fishing is greater than the opposite in Figure 5.11. However, the BF suggests the decisively supported non-zero transition rates from fishing to inland and strongly supported from inland to fishing. Indeed the BF calculation is based on the indicators δ that are summarised in Figure 5.12 where the indicator for non-zero fishing-to-inland rate is 1 (1 - 1), and that for inland-to-fishing is 1

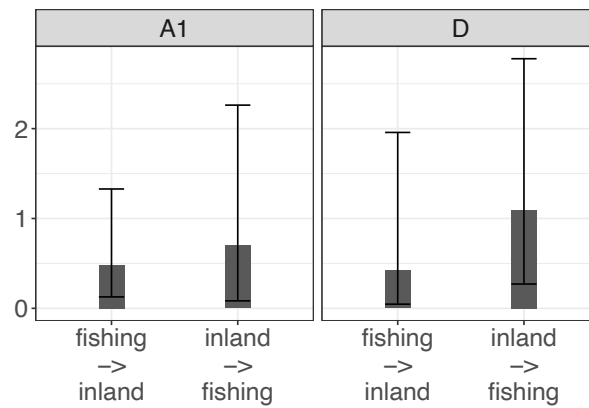


Figure 5.11 **Transition rates.** We selected high coverage nucleotide sequences of subtypes A1 and D to build phylogenies. We used BEAST to reconstruct discrete ancestral states for each phylogeny and inferred transition rates between inland and fishing communities. The error bar plot shows the transition rates from fishing to inland and from inland to fishing, based on the subtype A1 phylogeny (left panel) and the subtype D phylogeny (right panel).

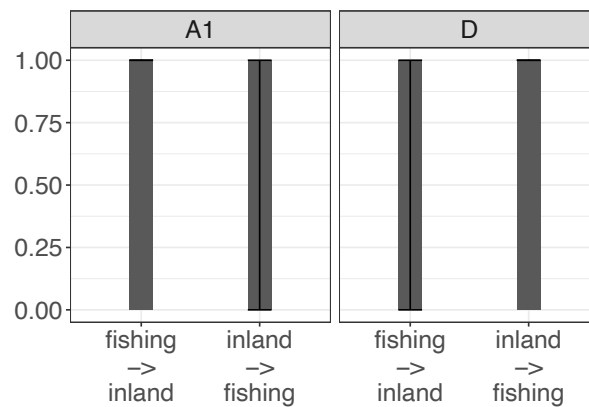


Figure 5.12 **Indicator for non-zero transition rates.** We used BEAST to reconstruct discrete ancestral states for each phylogeny and inferred transition rates between inland and fishing communities. Meanwhile, we applied BSSVS to select the nonzero rates. The error bar plot shows the indicator for non-zero transition rates from fishing to inland and from inland to fishing, based on the subtype A1 phylogeny (left panel) and the subtype D phylogeny (right panel).

(0 - 1). We were not able to draw a firm conclusion in the migration model because of the inconsistent outcomes by subtypes and metrics (condition flow ratio and BF).

We summarise the caveats of the CTMC through the experiment. First, the conditional flow ratio instead of (5.2c) can be evaluated from the CTMC model, and it does not take the infectious population into account. The deep sequence based approach relies on the source-recipient relationships aggregated from deep sequence phylogenies, and thus provide more than the conditional flows. Second, comparison of the conditional flow ratios between two approaches indicates the propagation of uncertainties across the phylogeny generates more variable estimates, from which we are unable to draw the conclusion. However, it may be attributed to the partial fragments, rather than whole genome, that are typically utilised in Sanger sequences based approach. Third, the CTMC requires the specification of a time interval of interest, which is challenging and arbitrary, while the deep sequence based approach works on a well-defined duration. Fourth, it is not straightforward to combine the outcomes based on varying subtypes in the CTMC.

5.5.2 Birth-Death Model

Next, we compared the deep sequence based approach with sampling adjustment and a Sanger sequence based approach (Kühnert et al., 2016). This method relaxed the assumption of independence between population structure and the evolution process in Lemey et al. (2009). Out of our interests, the paper accounts for the spatial sampling heterogeneity.

Kühnert et al. (2016) models the phylogeny by a multi-type birth-death process where types are similar as states in Section 5.5.1, e.g. $\{i, f, u\}$. The process is specified by birth rates b_k , death rates d_k , samplings rate ψ_k and type change rates to j given by m_{kj} where $k \in \{i, f, u\}$. The authors derived the tree likelihood under this process. The posterior distribution of epidemiological parameters is inferred in BEAST package `bdmm`. For the sake of simplicity in specifying the prior, the model is reparameterised by the rate of an individual of type k becoming noninfectious δ_k , the effective reproductive number R_k , and the rate of being serially sampled s_k ,

$$\delta_k = d_k + r\psi_k, \quad R_k = \frac{b_k}{\delta_k}, \quad s_k = \frac{\psi_k}{\psi_k + d_k}, \quad (5.30)$$

where r is the removal probability upon sampling and takes the value 1 for simplicity. We chose priors according to either previous studies, empirical data or the CTMC model in Section 5.5.1. The choices are described in Table 5.5.

Table 5.5 Prior choices for the multi-type birth and death model.

Parameters	Priors	Interquartile ranges	Reasons
s_{inland}	Beta(46, 54)	0.46 (0.43 - 0.49)	Empirical
s_{fishing}	Beta(54, 46)	0.54 (0.51 - 0.57)	Empirical
s_{unknown}	Beta(50, 50)	0.50 (0.47 - 0.53)	Average of inland and fishing
Clock rate	Uniform (0, 1e100)	NA	CTMC
$m_{k,j}$	LogNormal (0, 1)	1yr (51d - 7.1 yr)	Nadeau et al. (2021)
R_k	LogNormal (0, 0.75)	1 (0.60 - 1.66)	Ratmann et al. (2017)
δ_k	LogNormal (-1, 0.5)	0.37 (0.26 - 0.52)	Ratmann et al. (2017)

We ran 10 MCMC chains to infer the epidemiological and migration parameters on a fixed phylogeny, each for 10^7 iterations, took the first 2×10^6 iterations as warm-up, and was thinned every 10^4 iterations. It takes 0.77-1.51 days using subtype A1, and more than 3 days for subtype D. We aborted the runs after 3 days due to the time constraints on High Performance Computing, and focused on subtype A1. Effective sample sizes per chain are > 500 . However, we observed the combined chain does not converge to the same stationary distribution in Figure 5.13, and the ESSs are < 50 , though strongly informative priors are imposed on sampling fractions in Table 5.5.

Indeed, the previous studies have highlighted the difficulty in applying bdmm at the scale of our problem. For instance, the authors of the bdmm model reported analyses on 500 influenza sequences distributed over 10 MCMC chains with each 90 million iterations, which in total resulted in ESSs of over 200 (Scire et al., 2020). We further searched the citations of the two bdmm papers for applications on HIV. Kühnert et al. (2018) analysed infections from the drug resistant strain and the drug sensitive strain using two-type birth-death-sampling process. They analysed 15 resistance mutation related data sets, and estimated tree topologies for each cluster, with cluster sizes ranging from 2-184. bdmm was run for at least 250 million combined iterations after burn-in, and > 200 ESS were obtained on these much smaller data sets. Jin et al. (2021) studied the spatial spread of HIV among men having sex with men populations diagnosed during 2009-2013 in 13 Chinese provinces. The analysis included 1205 *pol* sequences of 1000bp from 5 subtypes, so again smaller sample sizes per subtype. bdmm was applied to two subtypes for 80 million iterations, and the ESS was > 200 . In two other studies, runtime, convergence and mixing metrics were not reported (Hanke et al., 2019; Ratmann et al., 2017). These results reinforce the formidable, computational challenges

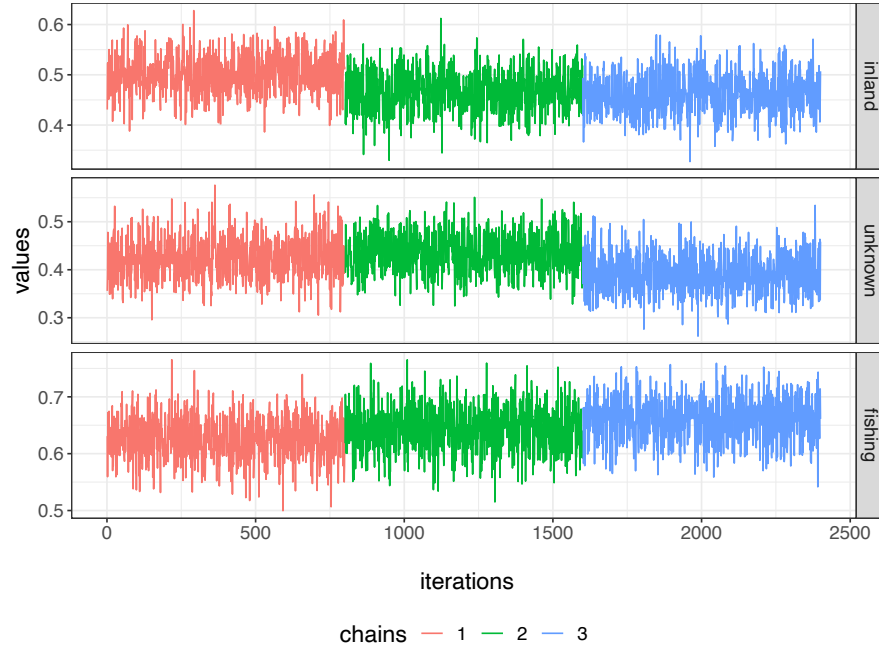


Figure 5.13 **Trace plots of the birth-death model.** Three thinned chains (in colour) of sampling fractions without the burn-in are combined and plotted in lines, showing these chains converge to different stationary distributions.

met when applying existing tools to consensus sequence data at scale, which we outlined in Chapter 1.

Last, we explore whether we are able to evaluate the flow ratio (5.2c) in the multi-type birth-death model. Ratmann et al. (2017) suggested the proportion of transmissions that originated from individuals of type k is

$$\pi_{k\cdot} = \sum_j \pi_{kj} = \frac{f_k b_k}{\sum_{k'} f_{k'} b_{k'}} = \frac{f_k R_k \delta_j}{\sum_{k'} f_{k'} R_{k'} \delta_{k'}}, \quad (5.31)$$

where f_k is the expected fractions of individuals in the type k . Therefore, we confront the same problem as Section 5.5.1 and (5.31) requires additional information on the infectious individuals during the time of interest. In addition, (5.31) only provides the transmissions from the deme through the type-specific birth rate. To quantify the transmission proportions between sub-populations, we would allow the birth rate among demes, i. e. b_{kj} . This was not implemented in Kühnert et al. (2016), but is recently possible in Scire et al. (2020). However, the children dependent birth rates have not been applied to HIV data through `bdmm` to our knowledge.

In summary, *bdmm* suffers from the drawback in the CTMC that the inference of (5.2c) requires additional information on infectious population in a well-defined time period. In addition, it is computational challenging to implement *bdmm* and achieve the convergence of MCMC.

5.5.3 Discussion

We compare the 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) and our approach tailored for pathogen deep-sequence data based on the experiments in Section 5.4.7, Section 5.5.1 and Section 5.5.2, and highlight the pros and cons of our approach.

The main advantages are first, that population-level spread can be directly estimated from individual source-recipient relationships extracted from deep-sequence phylogenies, and modelled in terms of associated individual covariates. Second, relatively little computational effort is needed to fit the Poisson flow model (5.12-5.14) to deep-sequence data, because it falls within the class of Bayesian hierarchical models for binary data, for which efficient fitting exists (Carpenter et al., 2017). Third, differences in how the phylogenetic data were sampled for each stratum can be explicitly accounted for in the model, which 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 5.2B). 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 3km 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 a Bayesian flow model provides a useful tool in this endeavour. Fourth,

the computational efficacy makes it feasible to investigate complex aspects of disease spread, such as mobility. Grabowski et al. (2020) found the evidence of HIV infected individuals migrating into fishing sites, and it is straightforward to incorporate the additional known migration data in the Bayesian hierarchical model framework. However, migration data are usually unavailable for all the taxa, and thus challenging in existing phylodynamics methods.

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 and 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 (6.2), the resulting flow estimates will in general not be consistent with the constraints imposed by standard assumptions on disease spread, as in Equations (5.3-5.5), and other techniques such as the structured coalescent may be better suited (Volz et al., 2009).

5.6 Summary

This Chapter proposes a basic modelling framework to estimate transmission flows with a likelihood model described by a Poisson system. This model is very different to typical phylodynamic modelling approaches (Lemey et al., 2009; Scire et al., 2020; Vaughan et al., 2014; Volz et al., 2009), and the proposed approach is possible because the data can be brought into the form of a flow matrix. This model also has a lot in common with flow models in other disciplines. In particular, flow matrices of the form (5.1) are also of central interest to characterise human migration flows between countries (Raymer et al., 2013), transport

flows between locations (Tebaldi and West, 1998), bacterial migration in humans (Ailloud et al., 2019), or human contact intensities (van de Kastele et al., 2017), and are alternatively referred to as origin-destination matrices (Hazelton, 2001; Miller et al., 2019). We find this convergence of Bayesian methods very attractive. It was challenging to account for structured sampling probabilities that differ by gender, age, and across locations, which ultimately led us to consider a much higher dimensional flow problem than we are actually interested in. To ensure that the data are not overwhelmed by the prior in this high dimensional space, we opted for objective priors on the transmission intensities. Simulations suggest that the impact of sampling heterogeneity on flow estimates can be large, and give a general idea of the bias reductions that can be achieved with the basic Poisson model. Application to data from the RCCS demonstrates the danger of overestimating flow ratios (inland $\rightarrow$ fishing/fishing $\rightarrow$ inland) if sampling heterogeneities are neglected. However, both the proposed model which adjusted for sampling heterogeneity and naïve flow model without sampling adjustment indicate that fishing sites, the high prevalence area (transmission hotspot), are the sink instead of the source of transmissions. This contradicts the common belief of fishing sites and may reduce stigma toward fisherfolk. From a prevention perspective, prioritising prevention portfolios to fishing sites would directly benefit residents, but would not have a massive indirect influence on the surrounding general population. We finally showed it is computational challenging to consider the sampling differences between population strata in standard phylodynamic methods (Le Vu et al., 2019; Scire et al., 2020; Volz et al., 2009), and only a small number of strata were usually allowed for the implementation.

Chapter 6

The semi-parametric Poisson flow model

6.1 Introduction

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. If true, combination prevention interventions that seek to reduce HIV acquisition in young women from older men, and seek to reduce transmission from older men and women could break the transmission cycle (De Oliveira et al., 2017). In this application, the full population is divided into sex-specific age groups (here we consider 1-year age groups between 15 to 49 years), and then transmission flows between and within age groups are estimated,

$$\pi = \begin{pmatrix} \pi^{\text{mf}} & 0 \\ 0 & \pi^{\text{fm}} \end{pmatrix}, \quad \pi^{\text{mf}} = \begin{pmatrix} \pi_{11}^{\text{mf}} & \dots & \pi_{1K}^{\text{mf}} \\ \vdots & \ddots & \vdots \\ \pi_{K1}^{\text{mf}} & \dots & \pi_{KK}^{\text{mf}} \end{pmatrix} \quad \pi^{\text{fm}} = \begin{pmatrix} \pi_{11}^{\text{fm}} & \dots & \pi_{1K}^{\text{fm}} \\ \vdots & \ddots & \vdots \\ \pi_{K1}^{\text{fm}} & \dots & \pi_{KK}^{\text{fm}} \end{pmatrix}, \quad (6.1)$$

where π_{ab}^{mf} denotes the proportion of transmissions from men in age band a to women in age band b , and similarly for π_{ab}^{fm} . The flow matrix (6.1) has $2K^2$ non-zero entries to estimate, which for 1-year age bands amounts to 2450 variables.

The model developed in Chapter 5 starts to break down in this application when the interest is in transmission flows between many population strata and data are sparse. In this Chapter, we will extend the model framework and exploit the fact that in many settings population strata are naturally ordered, for example, K 1-year age bands in (6.1), and make use of smoothing, regularising prior densities to facilitate flow inferences. Before delving

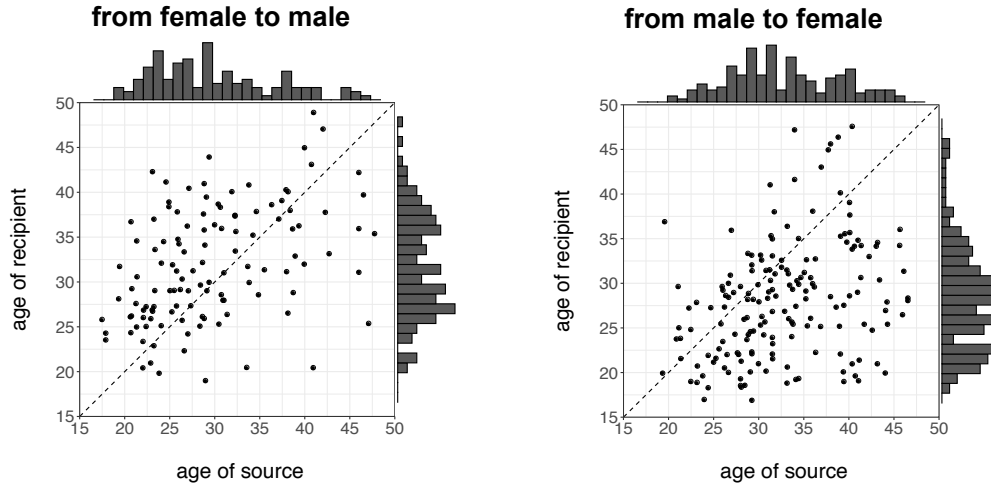


Figure 6.1 **Age-specific transmission data.** The study did a household census, and all individuals aged 15-49 years capable to provide informed consent and resident for at least 1 month with the intention to stay were invited to participate. Viral deep-sequencing was performed on plasma blood samples from HIV infected participants who reported no ART use. Deep-sequence phylogenetic analysis returned phylogenetic transmission scores between individuals, and 293 pairs had the 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.

into the model, we visualise the phylogenetically reconstructed source-recipient pairs by the ages of source and recipients in Figure 6.1. It illustrates the sparse data in each cell, if the stratification is by one-year age bands. Regularising prior utilises the smoothing properties and allows borrowing information across the cells, which might be beneficial for more accurate flow estimates.

6.2 Development of the semi-parametric Poisson model

6.2.1 Model framework

We utilise the likelihood (5.12), and replace the objective prior density (5.16) with a regularising prior density in multiple dimensions, such as bivariate Gaussian process (GP), for regularising transmission dynamics by age of source and recipient. This penalises against the large deviation in transmission intensities between the source and recipient subpopulations of similar ages. Specifically, we opted for (stacked) two-dimensional Gaussian-process priors

on the entries of Poisson intensities λ in (5.12) to estimate age-specific transmission flows from men to women and from men to women,

$$\begin{aligned}
\log \lambda &= \mu \mathbf{1} + \mathbf{v} \mathbf{1}_{\text{mf}} + f, \\
f &= (f_{\text{mf}}^T, f_{\text{fm}}^T)^T, \\
f_{\text{mf}} &\sim \text{GP}(0, k_{\text{mf}}), \quad f_{\text{fm}} \sim \text{GP}(0, k_{\text{fm}}), \\
k_{\text{mf}}((a_1, b_1), (a_2, b_2)) &= \sigma_{\text{mf}}^2 \exp \left(- \left[\frac{(a_2 - a_1)^2}{2\ell_{\text{mf},a}^2} + \frac{(b_2 - b_1)^2}{2\ell_{\text{mf},b}^2} \right] \right) \\
k_{\text{fm}}((a_1, b_1), (a_2, b_2)) &= \sigma_{\text{fm}}^2 \exp \left(- \left[\frac{(a_2 - a_1)^2}{2\ell_{\text{fm},a}^2} + \frac{(b_2 - b_1)^2}{2\ell_{\text{fm},b}^2} \right] \right) \\
\mu, \mathbf{v} &\sim \text{Normal}(0, 10) \\
\sigma_{\text{mf}}^2, \sigma_{\text{fm}}^2 &\sim \text{Half-Normal}(0, 10) \\
\ell_{d,i} &\sim \text{Inv-Gamma}(\alpha_{d,i}, \beta_{d,i}),
\end{aligned} \tag{6.2}$$

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 for female-to-male, $\mathbf{v}$ is a scalar on the elements of λ that represent the difference from the male-female direction to the baseline, and $k_{\text{mf}}, k_{\text{fm}}$ are gender-specific squared exponential kernels with variance parameters $\sigma_{\text{mf}}^2, \sigma_{\text{fm}}^2$ and length scales $\ell_{mf,a}, \ell_{mf,b}, \ell_{fm,a}, \ell_{fm,b}$ (Rasmussen and Williams, 2006). We chose the above structure because we hypothesised that the structure of age-specific flows from men to women is completely different to the structure of age-specific flows from women to men. Alternative models will be investigated in the next Chapter, where baselines and age-specific patterns may depend on locations of sources and recipients. We next describe inference of the posterior distribution

$$\begin{aligned}
p(\mu, \mathbf{v}, f_{\text{mf}}, f_{\text{fm}}, \sigma_{\text{mf}}^2, \sigma_{\text{fm}}^2, \ell_{\text{mf},a}, \ell_{\text{fm},a}, \ell_{\text{mf},b}, \ell_{\text{fm},b}, \xi | \mathbf{n}) &\propto \\
p(\mathbf{n} | \mu, \mathbf{v}, f_{\text{mf}}, f_{\text{fm}}, \xi) p(\mu) p(\mathbf{v}) p(f_{\text{mf}} | \sigma_{\text{mf}}^2, \ell_{\text{mf},a}, \ell_{\text{mf},b}) p(f_{\text{fm}} | \sigma_{\text{fm}}^2, \ell_{\text{fm},a}, \ell_{\text{fm},b}) p(\xi).
\end{aligned} \tag{6.3}$$

6.2.2 Numerical inference using 2D Hilbert Space Gaussian Process Approximations in Stan

It is computationally infeasible to infer the hyperparameters under the GP prior (6.2) using the codes in https://github.com/BDI-pathogens/phyloscanner/blob/master/phyloflows/inst/misc/gp_prior.stan and Appendix E.3.1. The main challenge in the inference lies in the inversion of the large covariance matrices as in Section 4.3.2. We found that the transmission

flows and model parameters can be efficiently estimated with the dynamic Hamiltonian Monte Carlo (HMC) sampler of the Stan probabilistic programming framework when the 2D GPs are approximated with a Hilbert Space Gaussian process approximation (HSGP) that was originally developed by Solin and Särkkä (2020).

In this approximation, the squared exponential covariance kernel k in (6.2) is approximated through a series expansion of eigenvalues and eigenfunctions of the Laplacian differential operator on a compact domain Ω of the input space. To prove this theorem, Solin and Särkkä (2020) defined a covariance operator associated with an isotropic covariance function, and expressed it by a series expansion of Laplacian operators. Second, the authors considered the eigenvalue problems for the Laplacian operator with Dirichlet boundary conditions (or other boundary conditions), and the Laplacian operator can be approximated with a Hilbert space method. As a result, the approximation only holds on the compact domain where a particular boundary condition holds. In our two-dimensional case, we consider the compact domain $\Omega = [-B_1, B_1] \times [-B_2, B_2]$, and the boundary points play an important role in the approximation, and need to be specified appropriately.

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

$$S_{\theta}(\omega) = 2\pi\sigma^2 \prod_{d=1}^2 \ell_d \exp\left(-\frac{1}{2} \sum_{d=1}^2 \ell_d^2 \omega_d^2\right), \quad (6.4)$$

where $\omega = (\omega_1, \omega_2)$ denote the frequencies, and $\theta = (\sigma, \ell_1, \ell_2)$ are the kernel parameters (Rasmussen and Williams, 2006). The approximation further involves the eigenvalues and eigenfunctions of the Laplacian differential operator. On the compact domain Ω , the j th univariate eigenvalues and eigenfunctions in dimension $d = 1, 2$ under the Dirichlet boundary condition are

$$\begin{aligned} \Lambda_{dj} &= \left(\frac{j\pi}{2B_d}\right)^2 \\ \phi_{dj}(x_d) &= \sqrt{1/B_d} \sin(\sqrt{\Lambda_{dj}}(x_d + B_d)), \end{aligned} \quad (6.5)$$

where x_d refers to age of sources ($d = 1$) or recipients ($d = 2$) (Solin and Särkkä, 2020). 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 $Y \in \mathbb{N}^{2 \times m}$. For

example, if $m_1 = 3$ and $m_2 = 4$, then

$$Y = \begin{pmatrix} 1 & 1 & 1 & 1 & 2 & 2 & 2 & 2 & 3 & 3 & 3 & 3 \\ 1 & 2 & 3 & 4 & 1 & 2 & 3 & 4 & 1 & 2 & 3 & 4 \end{pmatrix}. \quad (6.6)$$

For 2D inputs, the $j = 1, \dots, m$ eigenvalues and eigenfunctions are the combinations of the univariate eigenvalues and eigenfunctions, $\Lambda_j = (\Lambda_{Y_{1j}}, \Lambda_{Y_{2j}})$, and $\phi_j(a, b) = \phi_{1Y_{1j}}(a)\phi_{2Y_{2j}}(b)$. The derivation of (6.4) and verification of (6.5) are in Appendix C.

This fully specifies the HSGP approximation to GPs with 2D inputs,

$$f \sim \text{HSGP}(0, k^{\text{HSGP}}) \\ k^{\text{HSGP}}((a, b), (a', b')) = \sum_{j=1}^m S_{\theta}(\sqrt{\Lambda_j}) \phi_j(a, b) \phi_j(a', b'), \quad (6.7)$$

which depends on the choice of m_1, m_2 and B_1, B_2 . Therefore, Equation (6.7) can be more compactly written as

$$k^{\text{HSGP}}((a, b), (a', b')) = \phi(a, b)^T \Delta \phi(a', b'), \quad (6.8)$$

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

$$\mathbf{f} \sim \text{Multivariate Normal}(\mathbf{0}, \mathbf{K}^{\text{HSGP}}) \\ \mathbf{K}^{\text{HSGP}} = \Phi \Delta \Phi^T, \quad (6.9)$$

where

$$\Phi = \begin{pmatrix} \phi_1(a_1, b_1) & \cdots & \phi_m(a_1, b_1) \\ \vdots & \ddots & \vdots \\ \phi_1(a_n, b_n) & \cdots & \phi_m(a_n, b_n) \end{pmatrix} \in \mathbb{R}^{n \times m}. \quad (6.10)$$

The two-dimensional zero-mean HSGP model is fully defined in terms of the kernel (6.7), 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 (6.7), and this is vectorised as (6.9). Like Section 4.3.2, we reparameterise in the non-centred form,

$$f(a, b) = \sum_{j=1}^m \sqrt{S_{\theta}(\sqrt{\Lambda_j})} \phi_j(a, b) \beta_j, \quad (6.11)$$

where $\beta_j \sim \text{Normal}(0, 1)$ for $j = 1, \dots, m$. (6.11) does not require the Cholesky decomposition as the standard GP, and thus eases the computation. The implementation of the HSGP approximation in Stan can be found in https://github.com/BDI-pathogens/phyloscanner/blob/master/phyloflows/inst/misc/gpappr_prior.stan or Appendix E.3.2.

To see why the HSGP approximation is computationally faster than a typical GP model, consider first the eigenvectors and eigenvalues described in (6.5). Clearly, they do not depend on any of the hyperparameters ℓ, σ , and thus can be efficiently precalculated. The only additional cost of the HSGP approximation involves the calculation of the spectral density (6.4) at each HMC iteration. These computations are manageable because it involves no large scale matrix computations. Importantly, the HSGP approximation in Equation (6.11) avoids the Cholesky decompositions that are required in the numerical implementation of the 2D GP prior, and this makes the HSGP approximation computationally more efficient than a standard implementation of 2D GP priors.

In our work, we applied the approximation (6.7) to each of the stacked GP prior components in (6.2). To our knowledge, this is one of the first applications of the HSGP approximation.

6.3 Results

We first compare the performance of GP and HSGP through a simulated experiment in Section 6.3.1. Then we apply the approach to Rakai Community Cohort Study (RCCS) and analyse the age dynamics by genders in the rest of the Section. Apart from the epidemiological question of interests, we explore the impact of regularisation in Section 6.3.3, age stratification in Section 6.3.4 and sampling heterogeneity in Section 6.3.8.

6.3.1 Simulations

We first assess on simulations the impact of different regularising prior densities on estimating transmission flows across orderable population strata, which is typically not considered in standard phylodynamics 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 transmission pairs from a similar GP model as (6.2). The model is the same as (6.2) except we used the baselines dependent on genders and locations, $\mu_{x,y}$ where $x = \text{mf}, \text{fm}$ and $y = \text{hh}, \text{hl}, \text{lh}, \text{ll}$ represent genders and locations

Table 6.1 Parameter values used to simulate transmission flows in populations stratified by 1-year age bands under a variation of GP model (6.2).

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	ℓ_{fm}	(4.1,2.3)
male to female	ℓ_{mf}	(2.3,4.6)
marginal standard deviation		
female to male	σ_{fm}	1.8
male to female	σ_{mf}	1.5

respectively. Here the ground-truth parameters were motivated by analyses of the Rakai data set, and specified in Table 6.1. All the parameters were in Table 6.1 except sampling fractions. We re-used the estimates from Appendix A and took the stratum averages as the ground-truth sampling intensities. 20 replicates of transmission intensities and counts were then generated. Figure 6.2A illustrates the simulated transmission pairs from men to women (in the red points), and the corresponding underlying flows (in the black contours) in two randomly selected simulations out of 20 generated simulations.

Transmission flows were then inferred using the HSGP approximation (6.7) in our Poisson flow model. We varied the number m of basis functions from 100 to 3600 by setting $m_1 = m_2 = 10, 20, \dots, 60$, and chose an expanded version of the input domain as HSGP domain Ω , $[15/B, 25 \times B] \times [15/B, 25 \times B]$, with boundary 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 from which the data were simulated as in (5.9) and (6.2), where we used prior distributions of sampling fractions in the RCCS analysis. Then we estimated the hyperparameters and flows using both the GP model and the HSGP model based on the

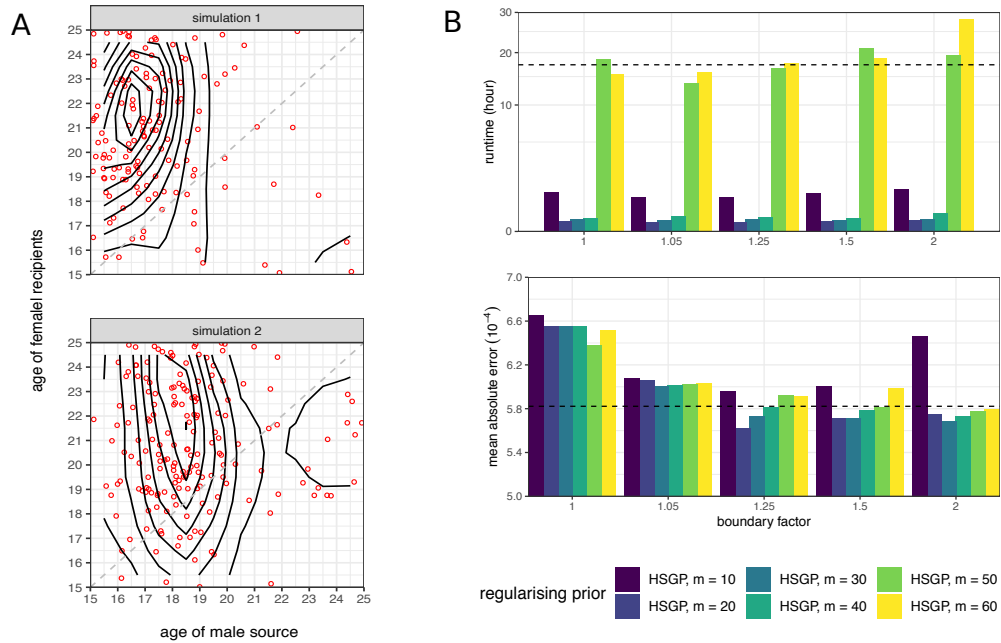


Figure 6.2 Simulation experiments to assess the accuracy of flow estimates under different smoothing priors. (A) Age- and gender-specific transmission pairs were simulated from the GP model (6.2), and shown are the simulated male to female transmission pairs for two simulations (red), along with contours of the simulated ground-truth transmission flow density. (B) Average runtimes of transmission flow inferences across 20 simulated data sets using different 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.

simulated observations. Figure 6.2B 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 6.2B 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 GP 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 $B = 1.25, m = 30$. Here, we illustrate further the performance of the HSGP approximation under the final tuning parameters, $B = 1.25, m = 30$. Figure 6.3 compares the marginal posterior distribution of the sources of infections in women when using the chosen HSGP approximation, compared to

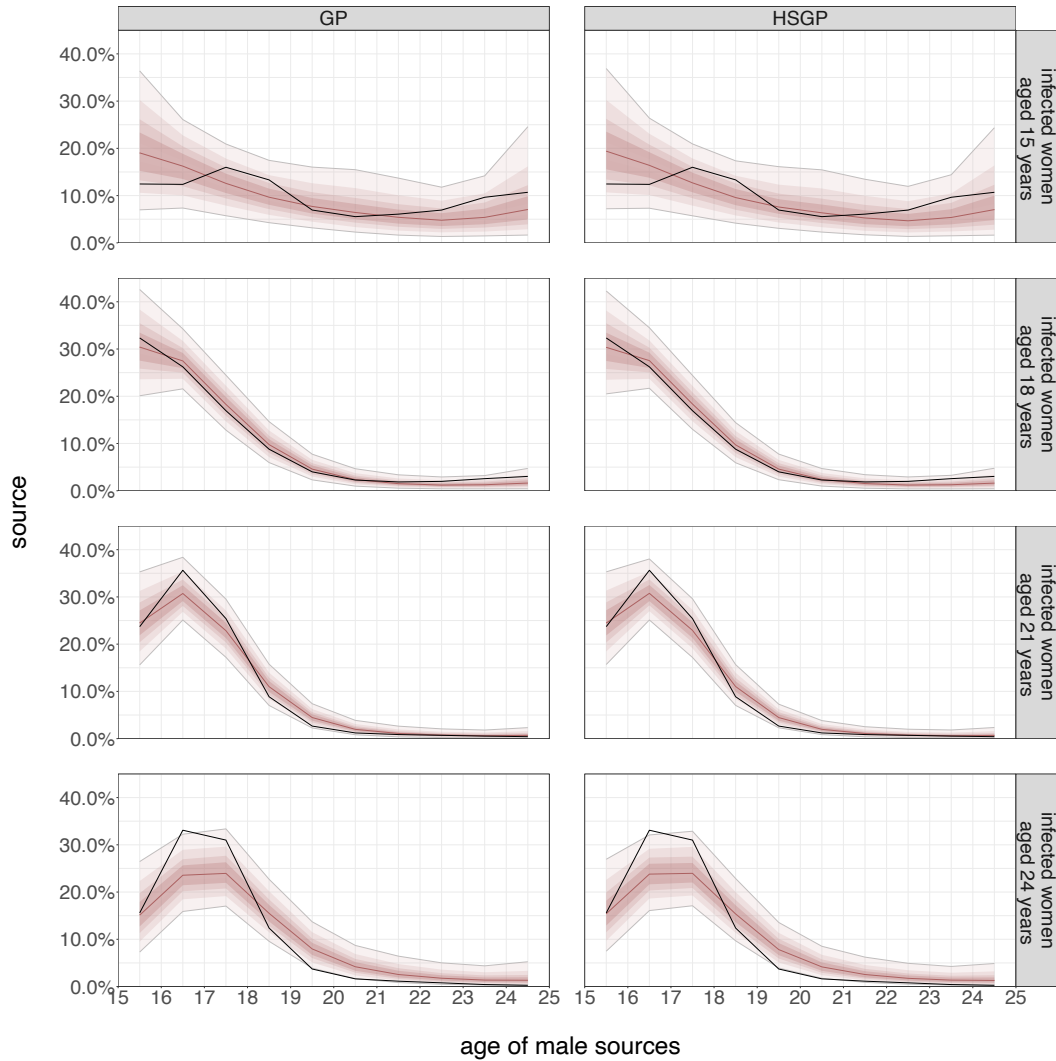


Figure 6.3 Estimated sources of infection in women under the final HSGP approximation. We simulated age-specific transmission flows under the kernel parameters shown in Table 6.1, and re-estimated transmission flows along with all other parameters under HSGP approximations to (6.2). The subfigures on 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 (CI), showing 40% CIs in dark red to 80% CIs in light red. Posterior medians are shown as red lines. True values are shown in black. We found no systematic differences in source estimates when using the final HSGP approximation over the GP prior.

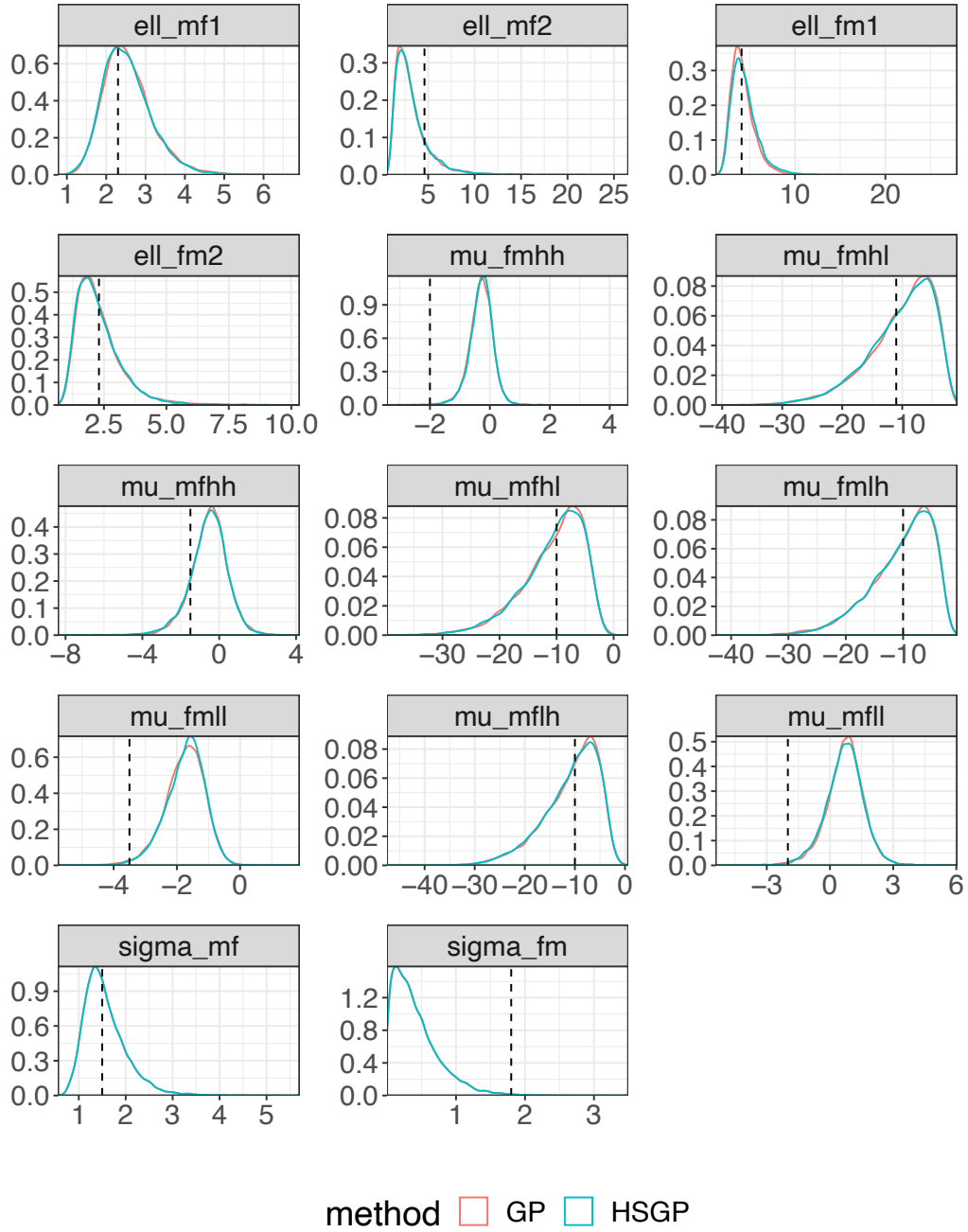


Figure 6.4 **Estimated covariance kernel parameters under the final HSGP approximation.** We simulated age-specific transmission flows under the kernel parameters shown in Table 6.1, and re-estimated transmission flows along with all other parameters under HSGP approximations to (6.2). The subfigures illustrate the marginal posterior densities of the kernel parameters under the GP model (in red) and the HSGP approximation (in blue) based on $B = 1.25, m = 30$ (right panels).

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 6.4 compares 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.

6.3.2 Numerical performance on Rakai data

We estimated transmission flows by age and gender from the source-recipient data shown in Figure 6.1. We will follow the same stratification of the population as in Chapter 5, i. e. stratify populations by gender, 1-year age bands, and fishing or inland location. To regularise inferences, we used the HSGP approximation (6.7) to the stacked GP prior (6.2). 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.

When using the HSGP approximation to the GP prior with $B = 1.25$ and $m = 30$, Stan's dynamic HMC algorithm took 89 hours for 10,000 iterations to infer the joint posterior density (5.14). We initially observed a larger number of divergent transitions as in Chapter 4.3, which concentrated on large length scale values. Using informative priors on the length scale hyper-parameters removed the problem. Specifically, we chose informative inverse Gamma distributions with hyperparameters set such that the prior 90% CIs match the 90% interquartile range of the inputs. All effective sample sizes were $\gg 5000$. Trace plots of the three parameters with the worst performance in terms of effective sample size are in Figure 6.5. We do not observe the mixing issues in the three chains. This shows that phylodynamic inferences remained computationally manageable even in the large flow space that was necessary to consider in order to account for sampling biases.

6.3.3 Transmission flows between age groups defined by 1-year age bands

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 non-parametric inferences by 1-year age bands.

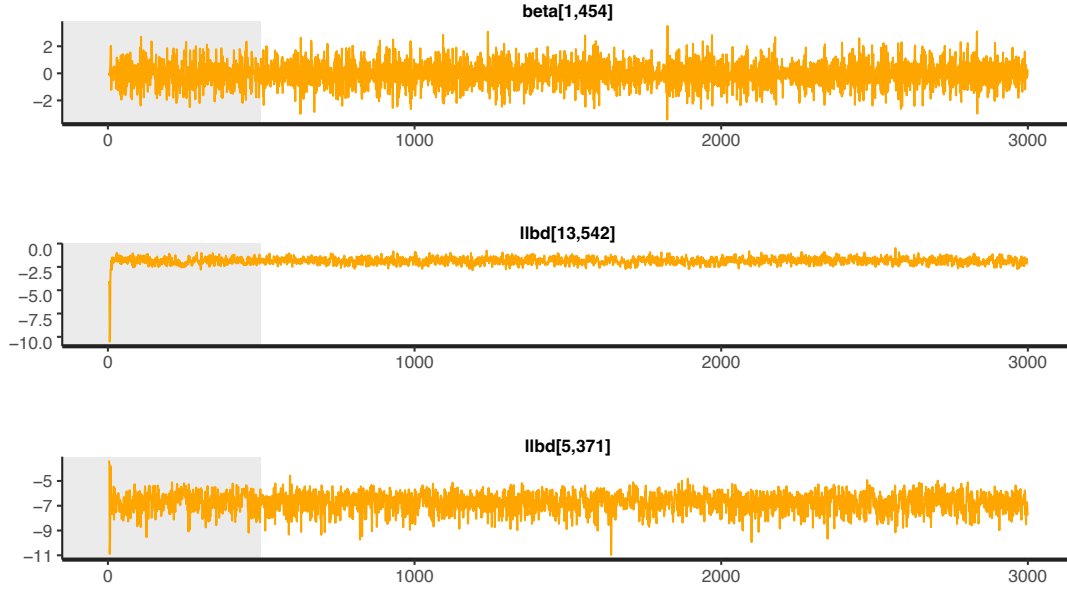


Figure 6.5 Trace plots for parameter inference on the Rakai data set, Poisson flow model with HSGP prior on the transmission intensities. Stan’s dynamic HMC algorithm was run for 10^4 iterations. Shown are trace plots for the first 3×10^3 iterations of the three parameters with the lowest effective sample sizes.

We compared estimates of the source vectors δ , defined in (5.2a), when we used the independent Gamma prior density (5.16) (no regularisation) to those when we used the HSGP prior density (6.7) (with regularisation). We focused the comparison on the age- and gender-specific sources of infection regardless of location, i.e. the source vector that corresponds to Equation (6.1), which was obtained by aggregating over the high- and low-prevalence locations of the source and recipient population groups. Figures 6.6A-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 6.6C-D show the corresponding posterior coefficients of variation. The estimated coefficients of variation were larger without regularisation, and well below 1 with regularisation, except 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.

6.3.4 Comparison to flows estimated using broader age strata

In this Section, we present results on analyses using data that is stratified by 2-year age bands, and 5-year age bands. Figure 6.7A below shows the estimated sources of infection

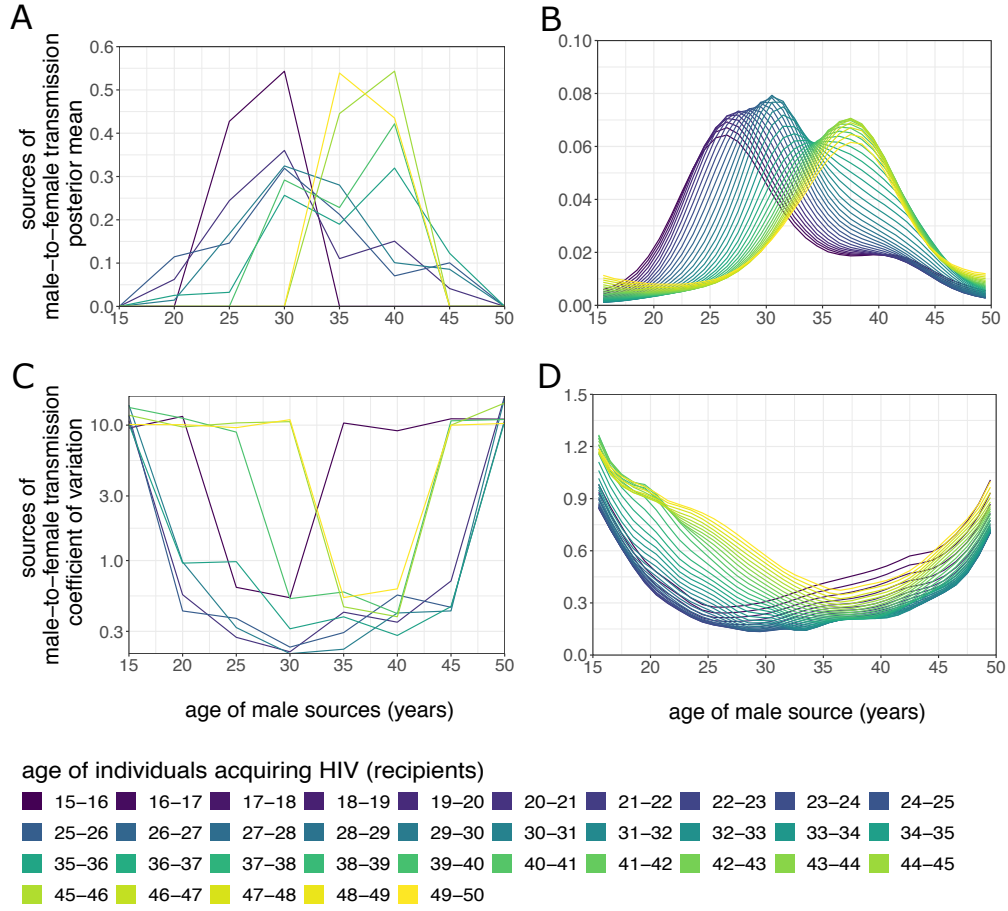


Figure 6.6 Impact of using regularising prior densities on the estimation of age- and gender-specific sources of infection. We compared estimates of the sources of HIV infection in women without regularisation, using the Gamma prior densities (5.16), to estimates obtained under the regularising HSGP prior on the log transmission intensities (6.7). 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.

in female recipients obtained without prior regularisation when 2-year age bands are used, and Figure 6.7B shows the same when 5-year age bands are used. It is clear that one would have to resort to 5-year age strata or wider in order to arrive at less pronounced regularisation effects, and so we compared inferences using 1-year, 2-year, and 5-year age strata in detail.

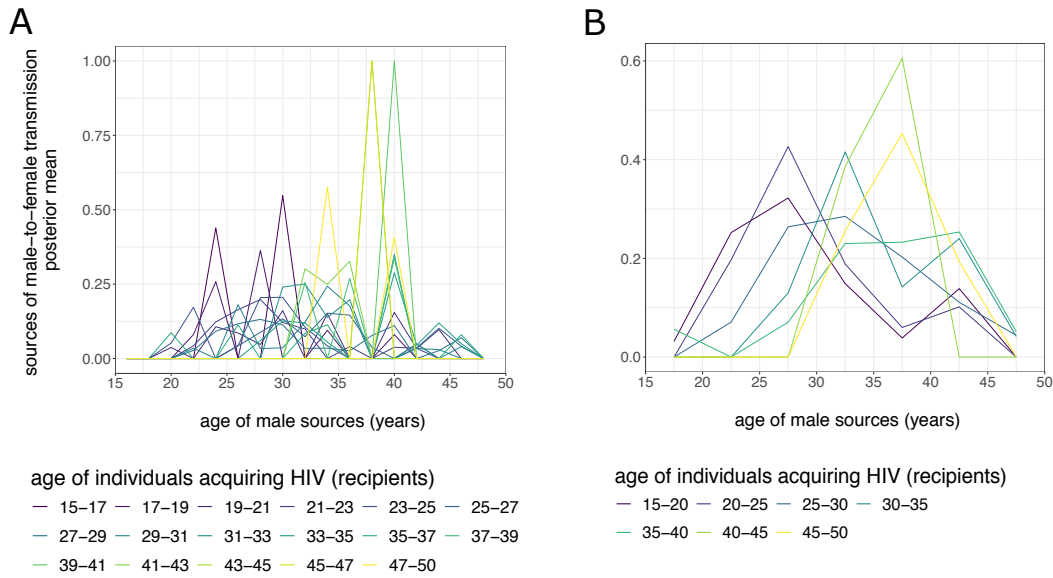


Figure 6.7 Unregularised flow estimates by age bands and genders. We compared estimates of the sources of HIV infection in women without regularisation, using the Gamma prior densities (5.16), when 2-year age bands and 5-year age bands were applied. Each colour corresponds to infection recipients of a particular age group. **(A)** Each line shows the posterior median estimates of the probability of infections attributed to the 2-year increment age group of individuals of the opposite gender, i.e. the sources of infection, obtained without regularisation. **(B)** Same as A, using 5-year increment age groups.

For the 5-year age band analysis, we consider two stratifications, $(15 - 19]$, $(20 - 24]$, ... and $(12.5 - 17.5]$, $(17.5 - 22.5]$, ... We used the same methodological approach as for 1-year age bands, except that based on the fitted GP model in Equation (6.2), 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 6.8A-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 evaluated 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 6.8. Intuitively, it is clear analyses stratified by 5-year age strata are very coarse to address the primary analytic aims.

Figure 6.9A illustrates the estimated age-specific transmission flows at 1-year age stratification (posterior median) from data stratified by 1-year age bands, and Figures 6.9B-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 6.10 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.

6.3.5 Sources of transmission to young and adolescent women

Across sub-Saharan Africa, HIV prevalence rises rapidly among young and adolescent women aged < 25 years (UNAIDS, 2018b), which has prompted efforts to prevent infection among adolescent girls and young women, most notably the Determined, Resilient, Empowered, AIDS-free, Mentored and Safe (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.

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

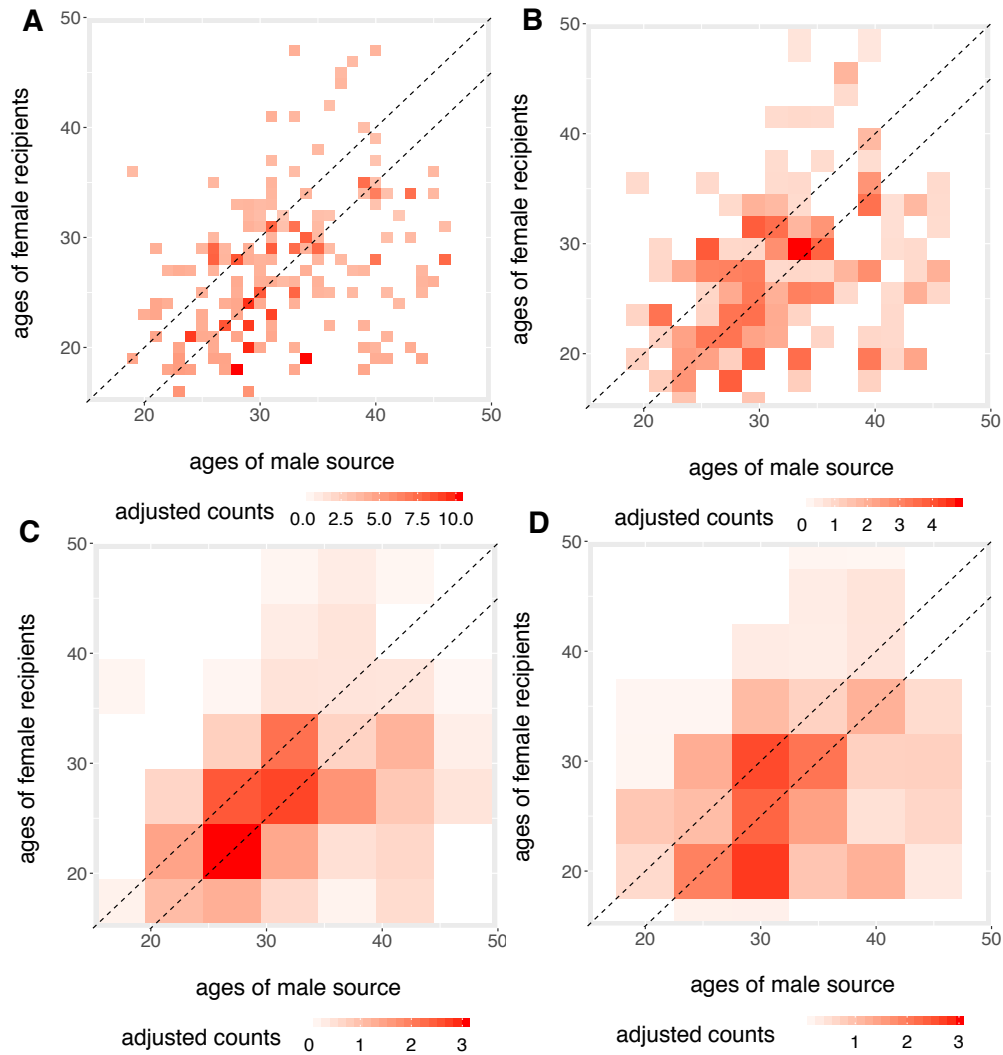


Figure 6.8 **Aggregated male-to-female transmission counts.** We compare male-to-female transmission counts (in colour) aggregated by 1-year (Panel A), 2-year (Panel B) and 5-year age bands (Panel C-D). These counts were adjusted by average sampling fractions per stratum. The two black dashed lines are reference lines, representing sources that are the same age as recipients (top) and 5 years older than recipients.

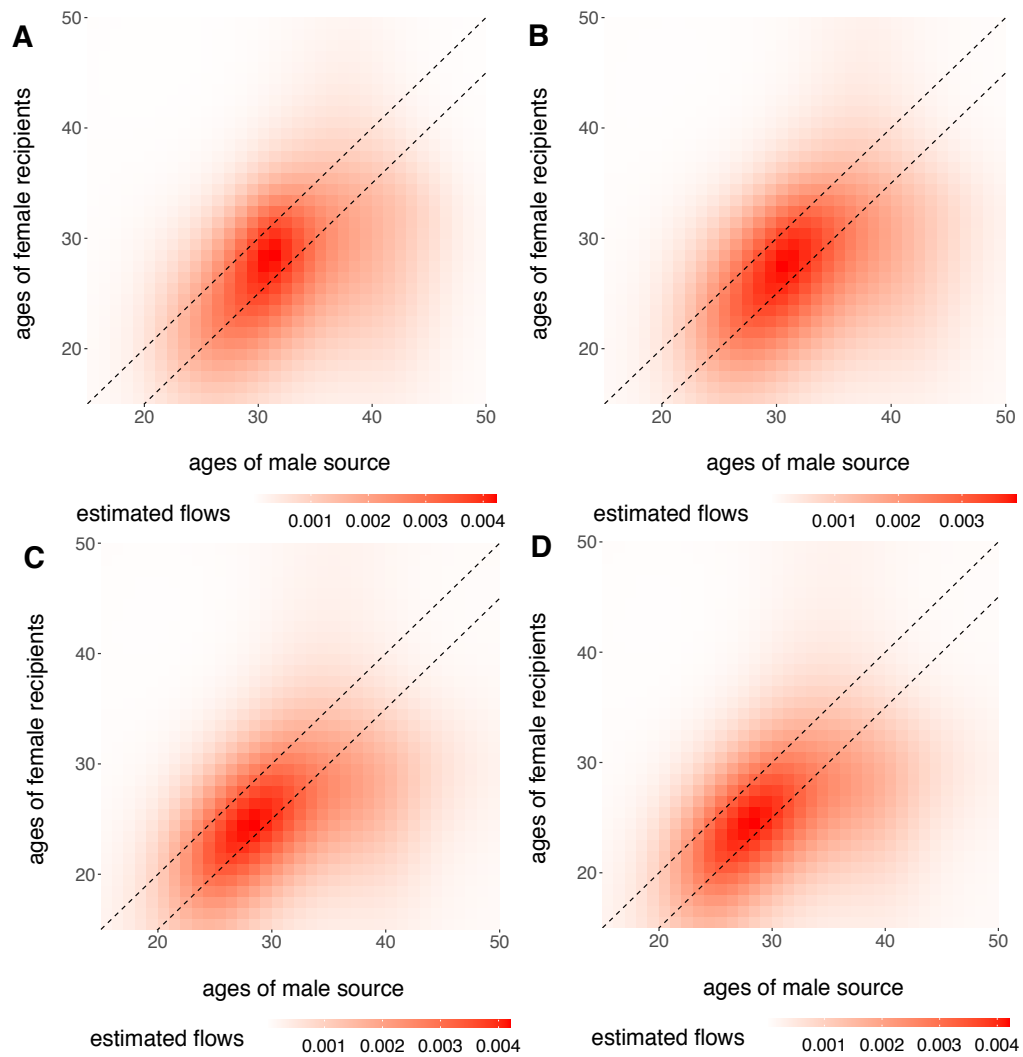


Figure 6.9 Estimated male-to-female transmission flows. Using the aggregated counts in Figure 6.8, we modelled the transmission intensities using the GP prior densities (6.2), and predicted flows by 1-year age bands. Panels A-D show the estimated transmission flows (in colour) from men (x-axis) to women (y-axis) using corresponding data in Figure 6.8. The two black dashed lines are reference lines, representing sources that are the same age as recipients (top) and 5 years older than recipients.

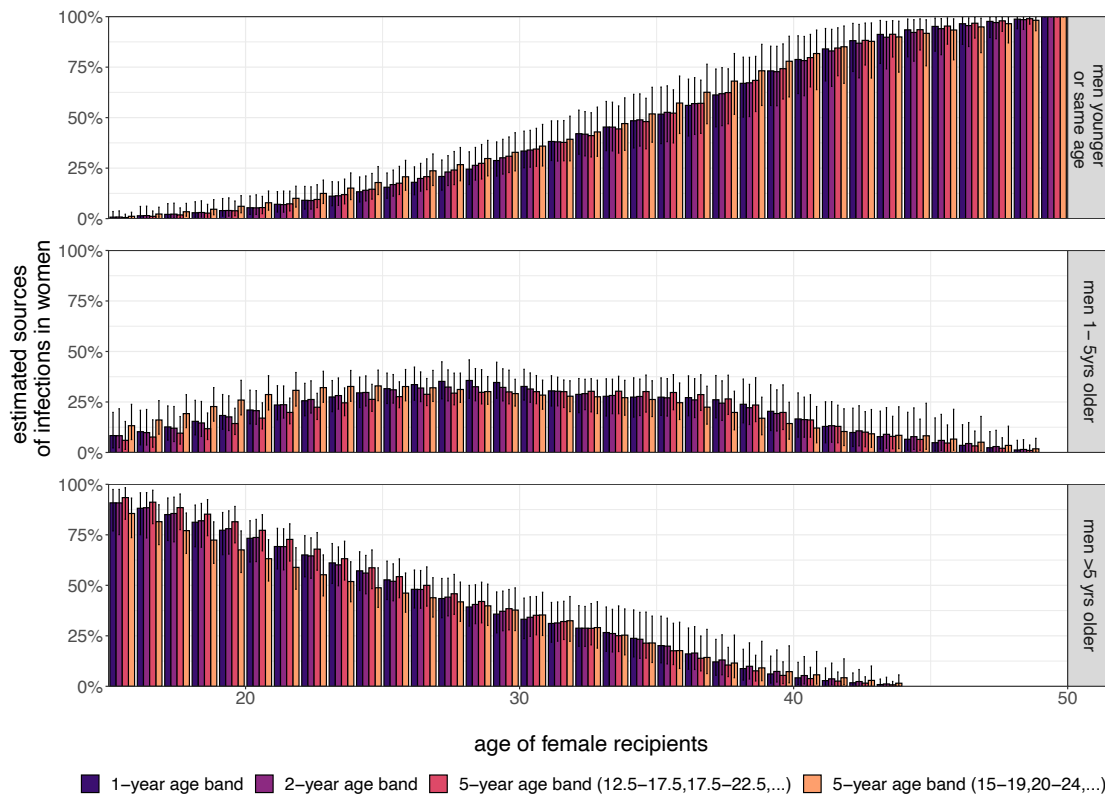


Figure 6.10 **Source of infections in women under 4 types of aggregation.** Using the aggregated counts in Figure 6.8, we modelled the transmission intensities using the GP prior densities (6.2), predicted flows by 1-year age bands and evaluated the proportion of infections from men in three age categories (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) are shown against ages of female recipients in the bar plots by four types of aggregations (in colour).

HSGP prior density (6.7), our gender- and age-specific flow estimates also borrow information from the other source-recipient pairs that involved older women. We report in Figure 6.11 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 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 2010-2015, and they show that the contribution of these men on infection in women declines rapidly with the age of the women.

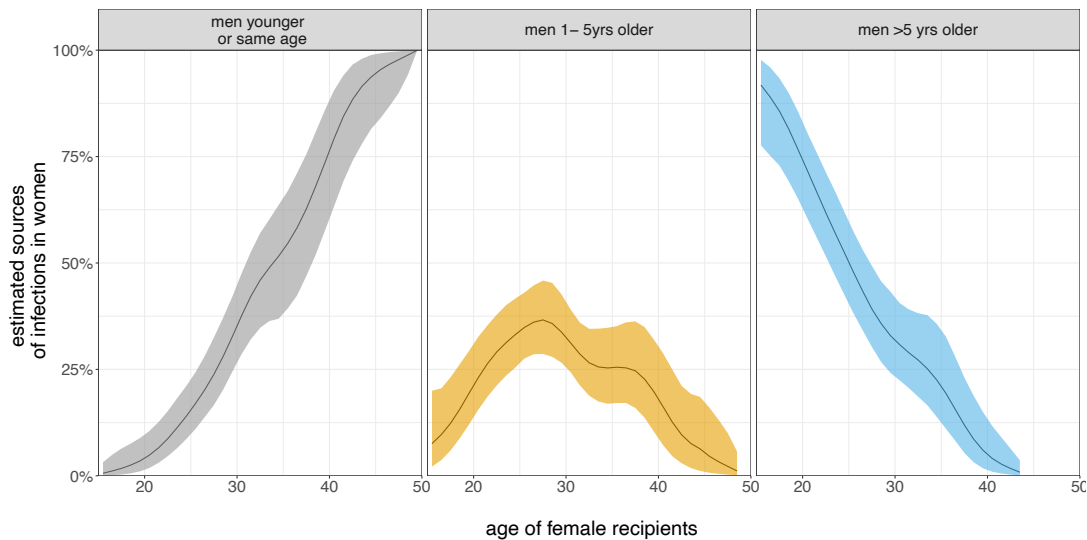


Figure 6.11 Estimated sources of infection in women in Rakai, Uganda, 2009-2015. We estimated the sources of HIV infection in women of increasing age. Sources were defined as men that were younger or the same age as the women (grey), 1-5 years older (orange), and over 5 years older (blue), and sum to 100% for each age of infected women on the x-axis. Posterior median source estimates are shown along with 95% marginal CIs.

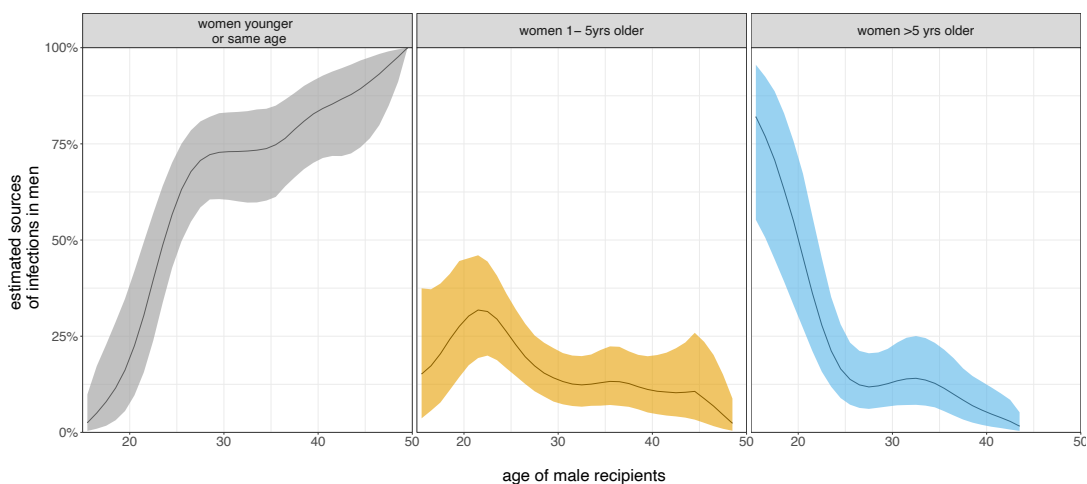


Figure 6.12 Estimated sources of infection in men in Rakai, Uganda, 2009-2015. We estimated the sources of HIV infection in men of increasing age. Sources were defined as women that were younger or the same age as the men (grey), 1-5 years older (orange), and over 5 years older (blue), and sum to 100% for each age of infected men on the x-axis. Posterior median source estimates are shown along with 95% marginal CIs.

Reducing HIV incidence among adolescents and young women is a key priority for public health programs across sub-Saharan Africa to achieve epidemic control milestones (Dellar

et al., 2015). The DREAMS intervention aims to promote determined, resilient, empowered, AIDS-free, mentored, and safe adolescent girls and young women, and includes educational programs that aim to address the socio-behavioural factors that underlie vulnerability and infection risk (Saul et al., 2018). In this context, two recent studies by Harling et al. (2014) and Balkus et al. (2015) found no association between HIV acquisition and being in an age-disparate sexual relationship with a partner more than 5 years older, suggesting that age-disparate relationships may not be a key determinant of current high infection risk among young women. In contrast, Akullian et al. (2017) and others found such associations, pointing in particular to non-linear changes in infection risk from age-disparate relationships by age of the newly infected partner. These analyses were based on self-reported sexual contact data and are subject to reporting bias (Akullian et al., 2017) and social desirability bias (Van de Mortel et al., 2008). For this reason, there is the increasing focus on viral phylogenetic analyses that are not based on self-reported data on sexual partners. Analysing consensus sequences of a household-based HIV survey in KwaZulu-Natal, South Africa, De Oliveira et al. (2017) identified 60 women aged <25 years with a phylogenetic linkage, of whom 42 (70.0%) had a probable male partner aged >25 years. Using consensus sequences, this study was unable to resolve the direction of transmission in phylogenetically linked transmission pairs. Probert et al. (2019) compared results obtained from deep sequence data to those from a fitted individual-based mathematical model of HIV transmission, both using data from the HIV Prevention Trials Network 071 (PopART) trial in Zambia (Hayes et al., 2019). The analyses suggested that the male partner in transmissions was on average older than the female, with some evidence that this gap was wider in male-to-female transmissions. Male sources were most frequently in their late twenties or early thirties while female sources were most commonly in their early twenties. Bbosa et al. (2020) analysed HIV deep-sequence data obtained from patients in health-care facilities and during prevention campaigns in south-eastern Uganda to the north of Rakai District. There were 22 women aged <25 years with a phylogenetically probable male source case, of whom 13 (59.0%) were aged >25 years and 10 (45%) had a phylogenetically probable male source who was at least 5 years older. Our findings, which draw from a large and unusually well-characterised population-based sample of sequence data, support these previous analyses and suggest that the large majority of infections in young women originates from age-disparate sexual relationships. In particular, we found that the contribution of age-disparate sexual relationships to infections in women depends strongly on the age of the woman, declining from an estimated 90.7% in women aged 15 years to 52.7% by age 25. This age dependence is obscured when considering larger age brackets. Taken together, the data from this study and other phylogenetic studies

from Uganda, Zambia, and South Africa suggest that the rapid increase in HIV prevalence among adolescent and young women may be driven by the same source populations across sub-Saharan Africa.

6.3.6 Role of age discordance in transmissions

Apart from adolescent girls and young women, we further explore the role of age discordancy among other age groups of women and men. In particular, Figure 6.11 also provides estimates of the age discordancy of male sources for older women. Over the entire age range of infected women, we found rapid decreases in the role of men more than 5 years older. For example, in women aged 35, the proportion of transmissions attributable to men at least 5 years older has fallen sharply to approximately 25%. Likewise, in women aged 35, the proportion of transmissions attributable to men 1-5 years older has also fallen to approximately 25%. This suggests men younger than 35 are the primary source of infection in women, both to younger women and to older women. In contrast, Figure 6.12 provides little evidence for age discordant female-to-male transmissions. In particular, the proportion of infections from women aged > 5 years older dropped dramatically with male ages. Approximately 50% of the infections in men were attributed to over 5 years older women in their early 20s. The proportion reduced to less than 20% at age 25.

6.3.7 Gender differences in HIV transmission flow

Section 6.3.6 already hinted at the asymmetry in age-specific transmission dynamics from men to women versus women to men. Our data suggest substantial differences. Figure 6.13 shows posterior median estimates of transmission flows from men to women on the left, and from women to men on the right. The observed data points are shown as dots in red, and posterior quantiles are shown as quantiles. Flow distributions were shifted downward for male-to-female transmissions and upward for female-to-male transmissions when compared with the diagonal line. Overall, transmissions from men to women appeared much more concentrated than transmissions from women to men.

We can again focus on the sources of HIV infections, and compare the age profile of male sources to the age profile of female sources. Figure 6.14 illustrates the proportion of transmissions from men in increasing age bands (in x-axis) to women of all ages, and similarly the proportion of transmissions from women in increasing age bands to men of all ages. Particularly, we estimate a peak (posterior mode) at 30 in the left and at 25 in the right.

About 50% of infections in women were attributed to men aged 26-35, and approximately half of transmission in men was attributed to women aged 21-30.

We then break down this analysis of sources even further, and consider the conditional age profile of male sources conditional on the age of infected women, and similarly for female sources in male infections. Figure 6.15 indicates substantial differences in the ages of male and female sources. In particular, the infections among young women were mainly acquired from men aged 25-30, and those among elder women from men between 35 and 40. In contrast, infections among young men and elder men were acquired from less concentrated age groups. For instance, there were more evidence that young men acquired infections from older women, and the opposite held for older men.

A simpler way to present these findings is to plot posterior interquartile ranges of the sources on the y-axis and the age of recipients on the x-axis. In Figure 6.16, we found evidence for a reversing role of partner age on transmission dynamics. While the adolescent girl and young women were infected by older men, women above 35 years old tended to be infected by younger men. The shifting role of partner age was observed in both directions, but the transition occurred earlier among transmissions to men.

6.3.8 Impact of adjustments for sampling heterogeneity

To explore the influence of the sampling adjustments presented in Chapter 5, we compared the smoothed flow estimates when the sampling fractions were and were not incorporated in the model. Recall that we denote the estimated flow from a to b when sampling heterogeneity is adjusted as $\hat{\pi}_{ab}$ and that without adjustment for sampling heterogeneity as $\hat{\pi}_{ab}^n$. The relative difference brought by sampling adjustment is then defined as

$$rd = \frac{\hat{\pi}_{ab} - \hat{\pi}_{ab}^n}{\hat{\pi}_{ab}^n}. \quad (6.12)$$

The quantity is visualised in Figure 6.17. The figure illustrates that transmissions from men aged 20-30 to women aged 15-35 and transmissions from women aged 20-30 would be over-estimated, if differential sampling probabilities had not been considered. In addition, failure to incorporate the sampling probabilities may have led to underestimation of transmissions from 30-45 year-old men to 30-40 year-old women and from 30-40 year-old women to 35-45 year-old men. The large relative differences on the edges do not reflect the considerable impacts of sampling adjustment but the small scale of flows.

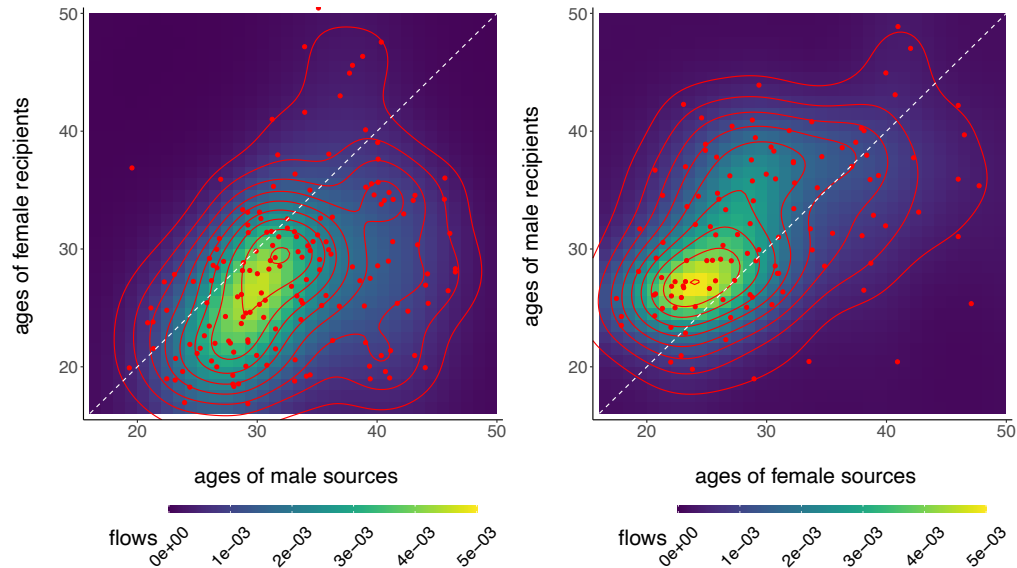


Figure 6.13 **Comparison of male to female and female to male transmission flows.** We visualise posterior median flow estimates between age groups for male-to-female (left) and female-to-male (right) transmissions through a heat map. Red dots show the observed phylogenetically likely source-recipient pairs and contours are 2D kernel density estimates. The white dashed lines are diagonal and provide a reference for transmission between individuals of the same ages.

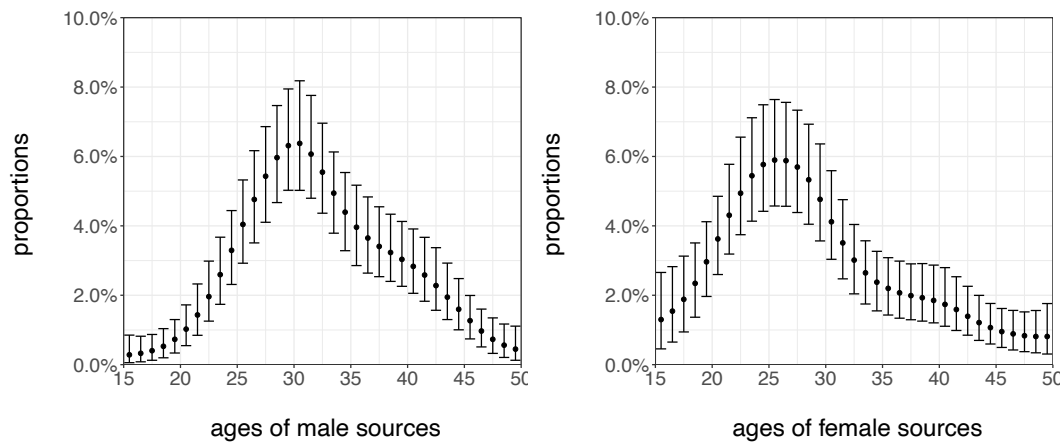


Figure 6.14 **Estimated age profile of male and female sources.** Shown is the contribution of age groups to the infections in women of all ages (left) and infections in men of all ages (right). Posterior median estimates are shown as points, and associated 95% CIs are plotted by error bars.

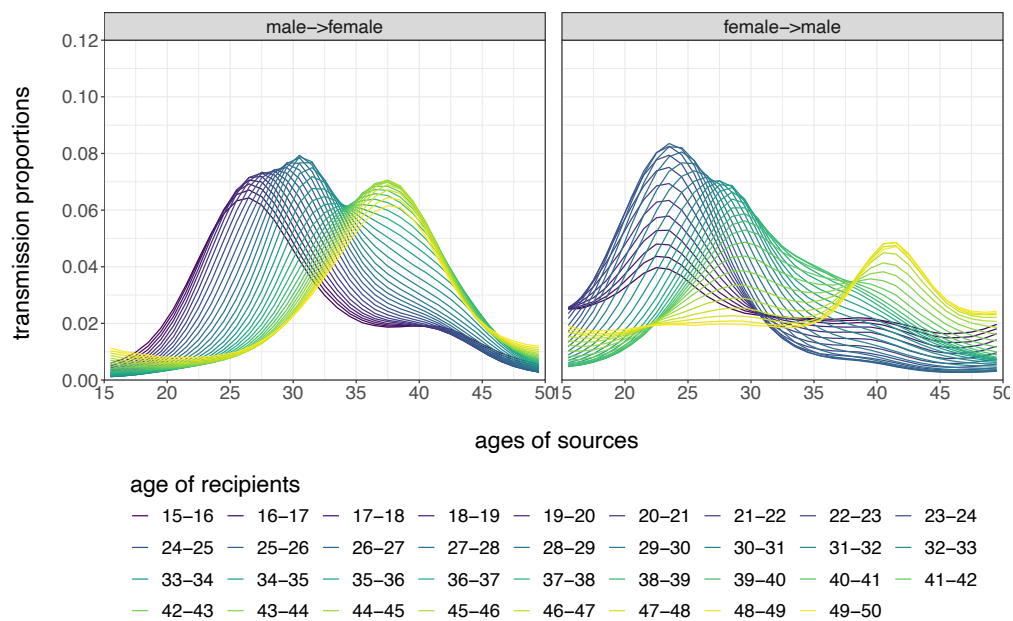


Figure 6.15 **Estimated age profile of infections by age of the infected partner.** We estimated the flows between one-year age bands, and normalised the flows to one age group (in the x-axis) among women (in the left) and men (in the right). Posterior median estimates are plotted by lines, and the colours of the lines indicate the given age of the recipients.



Figure 6.16 **Estimated age profile of infections by age of infected partner, second version.** The left panel shows posterior interquartile ranges of male sources to infection in women, and the right panel shows posterior interquartile ranges of female sources to infection in men. Posterior medians are shown as lines, and the 50% highest posterior density interval (HPI) and 90% HPI are shown in darker and lighter shaded areas. The diagonal dashed lines act as the reference and indicate the same source and recipient ages.

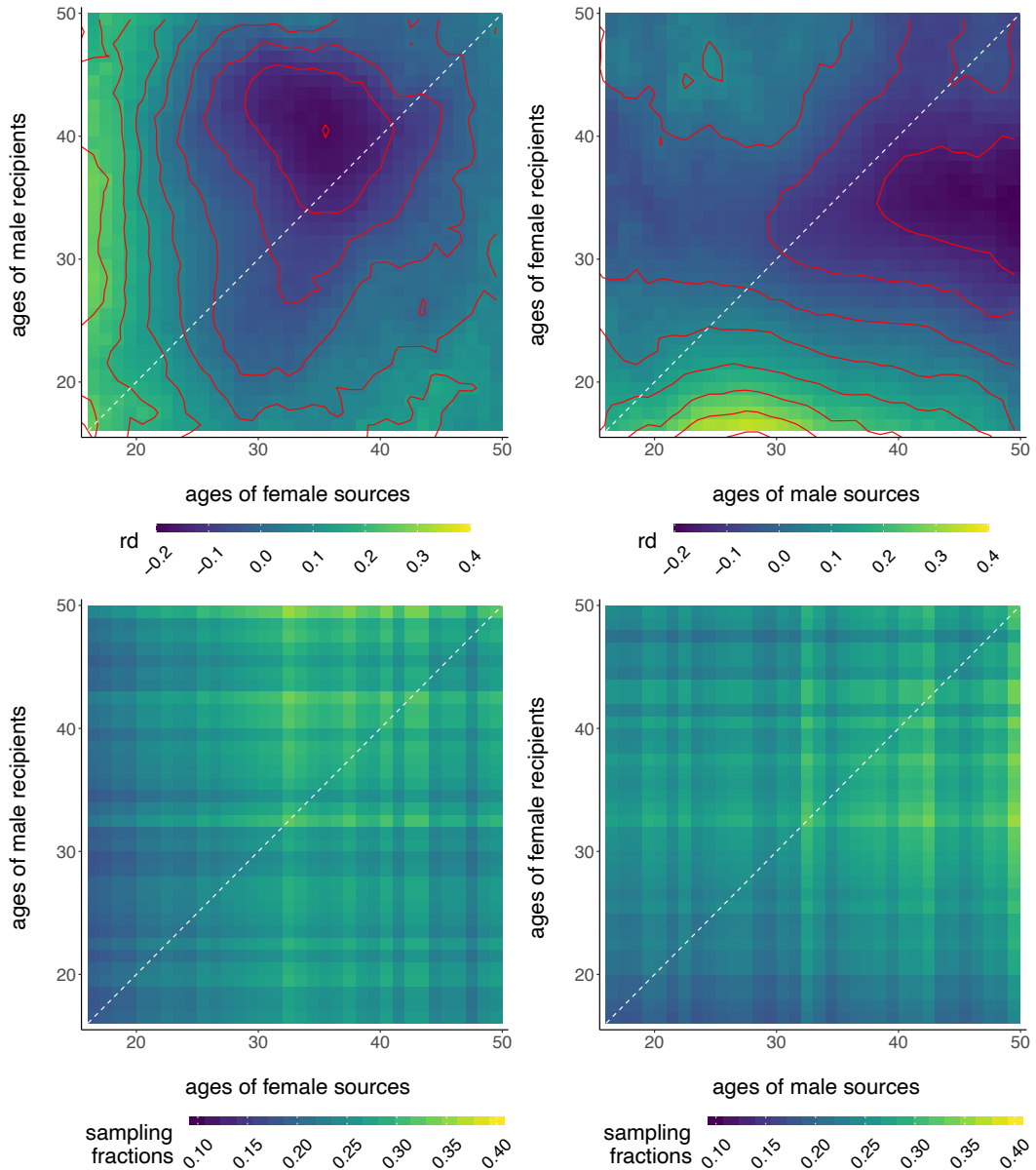


Figure 6.17 **Comparison of sampling adjusted and sampling unadjusted transmission flows.** The two figures on the top show the relative differences (abbreviated by rd) of flow estimates (by posterior medians) with and without sampling adjustment, for male-to-female transmissions (left) and female-to-male transmissions (right). The two figures on the bottom show sampling fractions of transmissions by one-year age band and gender.

6.4 Summary

This Chapter proposed an extension to Chapter 5, and inferred age-specific transmissions at a high resolution. This is challenging in standard phylodynamic analysis, because of the growing complexity in modelling and inference brought by small age bands. Deep sequence analysis, combined with statistical methods addressed the problem. HSGP approximation makes the computation scalable, but includes extra parameters to make approximation work. Choices of these parameters were discussed in the simulation. Applying the model to RCCS allows investigating the role of age-discordant relationship in HIV epidemics. We found age-disparate relationship drives the transmissions to adolescent girls and young women.

Chapter 7

Statistical methods for comparing flow models

7.1 Introduction

It is clear that there are multiple alternatives for modelling transmission flows based on the phylogenetic source-recipient data. For example, transmission flow models of age-specific transmission dynamics may or may not depend on gender, the location of the phylogenetically likely recipients, or the location of the phylogenetically source. Such alternatives can be easily formulated in the linear predictor of model (6.2). In this Chapter, we aim to develop methods to compare such models, and we will apply these methods to compare seven candidate models of age-specific transmission flows on the Rakai data set. The models have similar forms as (6.2), except that the baseline varies across (1) genders, (2) genders and source locations and recipient locations, and the age structure depends on (1) genders (2) genders and destination locations, (3) genders, source locations and recipient locations. The baseline model does not reply on both genders and locations.

We focus on using the Bayes factor for model comparison, which we explained in brief in (7.2). Recall that the Bayes factor is defined as the ratio of the marginal likelihood of two competing models. Suppose M_1 and M_2 denote the two models with parameters θ_1 and θ_2 respectively, and $\mathbf{y}$ represents the data. The marginal likelihood of M_1 is defined as

$$p(\mathbf{y}|M_1) = \int p(\mathbf{y}|\theta, M_1)p(\theta|M_1)d\theta \quad (7.1)$$

and similarly for M_2 . The Bayes factor (BF) is given by

$$\text{BF} = \frac{p(\mathbf{y}|M_1)}{p(\mathbf{y}|M_2)}. \quad (7.2)$$

The marginal likelihoods are not known analytically in most cases, and it is not straightforward to approximate the marginal likelihood with numerical Monte Carlo methods. Thus, it is often not straightforward to compute the Bayes factor, especially in hierarchical models which may contain a large number of parameters (Gronau et al., 2017a). Chib (1995) exploits a simple identity to obtain the marginal likelihood estimate from Gibbs output, and this is then extended to the outputs from Metropolis Hasting algorithm (Chib and Jeliazkov, 2001). However, they may not be suitable for multi-modal distributions. In addition, the marginal likelihood can be estimated through importance sampling (Gelfand and Dey, 1994; Touloupou et al., 2018). Meng and Wong (1996) generalises importance sampling by introducing a "bridge" function, i.e. bridge sampling. A further generalisation raises the likelihood to powers between 0 and 1, generates a series of power posteriors, and thus forms the infinite number of bridges, i.e. thermodynamic integration (Lartillot and Philippe, 2006; Ogata, 1989) or path sampling (Gelman and Meng, 1998). Xie et al. (2011) proposes steppingstone sampling which uses power posterior series in the importance sampling and yields faster convergence than path sampling. Other methods such as sequential Monte Carlo (Zhou et al., 2016) and nested sampling (Skilling, 2004) do not need Markov Chain Monte Carlo (MCMC). Despite of the superior performance and popularity of path sampling and steppingstone sampling in the phylogenetics literature, we will apply a relatively simple approach, bridge sampling, to compare our candidate models.

7.2 Summary of four methods for numerically approximating the marginal likelihood

7.2.1 Four methods to estimate the marginal likelihood

We will discuss four Monte Carlo methods for marginal likelihood estimations in this Section as introduced in Gronau et al. (2017a), with the emphasis on the last approach, bridge sampling. The simplest method is the direct Monte Carlo estimate (Hammersley, 2013)

based on the definition of the marginal likelihood (7.1),

$$\hat{p}^{\text{MC}} = \frac{1}{N} \sum_{n=1}^N p(\mathbf{y}|\tilde{\theta}_n, M), \quad \tilde{\theta}_n \sim p(\theta|M), \quad (7.3)$$

where $p(\theta|M)$ denotes the prior distribution and N is the number of Monte Carlo samples. This method provides an accurate estimate if the prior and the posterior are similar. Otherwise, most samples from the prior distribution contribute only slightly to the estimate and few samples dominate the estimate, implying a large variance of the estimator (7.3).

To tackle the high variance problem, we may utilise an importance density $g(\theta)$ that generates more samples at which the likelihood is high (Geweke, 1989). With this in mind, (7.1) can be rewritten as

$$\begin{aligned} p(\mathbf{y}|M) &= \int p(\mathbf{y}|\theta, M) p(\theta|M) d\theta \\ &= \int p(\mathbf{y}|\theta, M) \frac{p(\theta|M)}{g(\theta)} g(\theta) d\theta. \end{aligned} \quad (7.4)$$

In this form, the marginal likelihood is estimated by

$$\hat{p}^{\text{IS}} = \frac{1}{N} \sum_{n=1}^N \frac{p(\mathbf{y}|\tilde{\theta}_n, M) p(\theta_n|M)}{g(\tilde{\theta}_n)}, \quad \tilde{\theta}_n \sim g(\theta). \quad (7.5)$$

An ideal importance density has a heavier tail than the posterior distribution to avoid the high importance weights assigned to the samples generated from the tail. Despite of efficiency in variance reduction, (7.5) works best if samples are independent identically distributed. However, samples from MCMC are often strongly correlated.

Another estimate was developed to address problems when the importance density has thinner tails than the posterior distribution, namely the generalised harmonic mean estimator, often also called the reciprocal importance sampling estimator (Gelfand and Dey, 1994). The generalised harmonic mean estimator is based on the equation

$$\frac{1}{p(\mathbf{y})} = \int \frac{g(\theta)}{p(\mathbf{y})} d\theta = \int \frac{g(\theta)}{p(\mathbf{y}|\theta)p(\theta)} p(\theta|\mathbf{y}) d\theta. \quad (7.6)$$

Then, the generalised harmonic mean estimator is given by

$$\hat{p}^{\text{HM}} = \frac{1}{N} \sum_{n=1}^N \left(\frac{g(\tilde{\theta}_n)}{p(\mathbf{y}|\tilde{\theta}_n)p(\tilde{\theta}_n)} \right)^{-1}, \quad \tilde{\theta}_n \sim p(\theta|\mathbf{y}), \quad (7.7)$$

where $p(\theta|\mathbf{y})$ is the posterior distribution. In contrast to the importance sampler, the generalised harmonic mean estimator requires that g has thinner tails. Harmonic mean estimators lead to infinite variance even in simple problems and are not recommended in the statistical literature.

Both approaches (7.5) and (7.7) require particular tail behaviour of the importance density relative to the posterior distribution. This is an undesired property, and makes the sampler challenging to apply in high dimensional problems. Bridge sampling relaxes this constraint (Meng and Wong, 1996). Bridge sampling is based on the following identity,

$$\begin{aligned} p(\mathbf{y}) &= \frac{\int p(\mathbf{y}|\theta)p(\theta)h(\theta)g(\theta)d\theta}{\int p(\mathbf{y}|\theta)p(\theta)h(\theta)g(\theta)d\theta} p(\mathbf{y}) \\ &= \frac{\int p(\mathbf{y}|\theta)p(\theta)h(\theta)g(\theta)d\theta}{\int \frac{p(\mathbf{y}|\theta)p(\theta)}{p(\mathbf{y})}h(\theta)g(\theta)d\theta}, \end{aligned} \quad (7.8)$$

where $g(\theta)$ is the proposal function and $h(\theta)$ is the so-called bridge function. This gives the bridge sampler through

$$\hat{p}^{\text{BS}}(\mathbf{y}) = \frac{\frac{1}{N_2} \sum_{i=1}^{N_2} p(\mathbf{y}|\tilde{\theta}_i)p(\tilde{\theta}_i)h(\tilde{\theta}_i)}{\frac{1}{N_1} \sum_{j=1}^{N_1} h(\theta_j^*)g(\theta_j^*)}, \quad \tilde{\theta}_i \sim g(\theta), \theta_j^* \sim p(\theta|\mathbf{y}), \quad (7.9)$$

where N_1 and N_2 are the number of samples drawn from the posterior density and the proposal density respectively.

The proposal density has similar functionality as the importance density in (7.5). The only requirement is that its support should overlap with that of the posterior distribution, and otherwise, the denominator is zero and the estimate is undefined. A standard choice is to fit a multivariate normal distribution to the transformed parameters on the real line. For instance, we estimated the mean vector and covariance matrix through the maximum likelihood estimates (MLE).

For the bridge function, Meng and Wong (1996) suggested an optimal choice in terms of expected relative mean square error (7.18), which is given by

$$h(\theta) = C \frac{1}{s_1 p(\mathbf{y}|\theta)p(\theta) + s_2 p(\mathbf{y})g(\theta)}, \quad (7.10)$$

where $s_1 = \frac{N_1}{N_1+N_2}$ and $s_2 = \frac{N_2}{N_1+N_2}$. This choice is optimal when the samples from the posterior distribution are independent, the samples from the proposal distribution are independent, and the two sets of samples are independent (Meng and Wong, 1996). The optimal bridge

function makes no assumption of the marginal likelihood of interests. Substituting the bridge function (7.10) into (7.9) gives

$$\hat{p}^{\text{BS}}(\mathbf{y}) = \frac{\frac{1}{N_2} \sum_{i=1}^{N_2} \frac{p(\mathbf{y}|\tilde{\theta}_i)p(\tilde{\theta}_i)}{s_1 p(\mathbf{y}|\tilde{\theta}_i)p(\tilde{\theta}_i) + s_2 p(\mathbf{y})g(\tilde{\theta}_i)}}{\frac{1}{N_1} \sum_{j=1}^{N_1} \frac{g(\theta_j^*)}{s_1 p(\mathbf{y}|\theta_j^*)p(\theta_j^*) + s_2 p(\mathbf{y})g(\theta_j^*)}}, \quad \tilde{\theta}_i \sim g(\theta), \theta_j^* \sim p(\theta|\mathbf{y}). \quad (7.11)$$

The equation shows the tail behaviour of the proposal distribution relative to the posterior distribution is a less important issue. In particular, each term in the summation in both numerator and denominator has a ratio of which the denominator contains both the unnormalised posterior distribution and the proposal function. That makes none of the single terms in the summation dominate the estimate. The problem of (7.11) is its dependence on the marginal likelihood $p(\mathbf{y})$. To solve the problem, the marginal likelihood is numerically approximated through an iterative scheme. At the $(t+1)$ th iteration, the numerical estimates of the marginal likelihood is updated from that at the t th iteration based on

$$\hat{p}_{t+1}^{\text{BS}}(\mathbf{y}) = \frac{\frac{1}{N_2} \sum_{i=1}^{N_2} \frac{p(\mathbf{y}|\tilde{\theta}_i)p(\tilde{\theta}_i)}{s_1 p(\mathbf{y}|\tilde{\theta}_i)p(\tilde{\theta}_i) + s_2 \hat{p}_t^{\text{BS}}(\mathbf{y})g(\tilde{\theta}_i)}}{\frac{1}{N_1} \sum_{j=1}^{N_1} \frac{g(\theta_j^*)}{s_1 p(\mathbf{y}|\theta_j^*)p(\theta_j^*) + s_2 \hat{p}_t^{\text{BS}}(\mathbf{y})g(\theta_j^*)}}, \quad \tilde{\theta}_i \sim g(\theta), \theta_j^* \sim p(\theta|\mathbf{y}). \quad (7.12)$$

Let us define $l_{1,i} = \frac{p(\mathbf{y}|\theta_i^*)p(\theta_i^*)}{g(\theta_i^*)}$, $l_{2,i} = \frac{p(\mathbf{y}|\tilde{\theta}_i)p(\tilde{\theta}_i)}{g(\tilde{\theta}_i)}$. Then (7.12) can be simplified and expressed by the logarithm of the marginal likelihood,

$$\begin{aligned} \hat{p}_{t+1}^{\text{BS}}(\mathbf{y}) &= \frac{\frac{1}{N_2} \sum_{i=1}^{N_2} \frac{l_{2,i}}{s_1 l_{2,i} + s_2 \hat{p}_t^{\text{BS}}(\mathbf{y})}}{\frac{1}{N_1} \sum_{j=1}^{N_1} \frac{1}{s_1 l_{1,j} + s_2 \hat{p}_t^{\text{BS}}(\mathbf{y})}} \\ &= \frac{\frac{1}{N_2} \sum_{i=1}^{N_2} \frac{\exp(\log(l_{2,i}) - \bar{l}_2)}{s_1 \exp(\log(l_{2,i}) - \bar{l}_2) + s_2 \exp(\log(\hat{p}_t^{\text{BS}}(\mathbf{y}) - \bar{l}_2))}}{\frac{1}{N_1} \sum_{j=1}^{N_1} \frac{-\exp(-\bar{l}_1)}{s_1 \exp(\log(l_{1,j} - \bar{l}_1)) + s_2 \exp(\log(\hat{p}_t^{\text{BS}}(\mathbf{y}) - \bar{l}_1))}}, \end{aligned} \quad (7.13)$$

where $\bar{l}_1^* = \text{median}(\log(l_{1,i}))$ and $\bar{l}_2^* = \text{median}(\log(l_{2,i}))$ are to moderate the extreme log marginal likelihood values and calculated at the first time when we apply (7.13) to update from the initial guess, i. e.

$$\hat{p}_1^{\text{BS}}(\mathbf{y}) = \frac{\frac{1}{N_2} \sum_{i=1}^{N_2} \frac{\exp(\log(l_{2,i}) - \bar{l}_2)}{s_1 \exp(\log(l_{2,i}) - \bar{l}_2) + s_2 \exp(\log(\hat{p}_0^{\text{BS}}(\mathbf{y}) - \bar{l}_2))}}{\frac{1}{N_1} \sum_{j=1}^{N_1} \frac{-\exp(-\bar{l}_1)}{s_1 \exp(\log(l_{1,j} - \bar{l}_1)) + s_2 \exp(\log(\hat{p}_0^{\text{BS}}(\mathbf{y}) - \bar{l}_1))}}, \quad (7.14)$$

where $\log(\hat{p}_0^{BS}(\mathbf{y}))$ is the initialisation of the log likelihood.

Indeed, Equations (7.3), (7.5) and (7.7) are all special forms of (7.9) (Gronau et al., 2017a). Specifically, if $h(\theta) = \frac{1}{p(\theta)}$ and $g(\theta) = p(\theta)$, the Monte Carlo estimate (7.3) is recovered,

$$\begin{aligned}\hat{p}^{BS}(\mathbf{y}) &= \frac{\frac{1}{N_2} \sum_{i=1}^{N_2} p(\mathbf{y}|\tilde{\theta}_i) p(\tilde{\theta}_i) \frac{1}{p(\tilde{\theta}_i)}}{\frac{1}{N_1} \sum_{j=1}^{N_1} \frac{1}{p(\theta_j^*)} p(\theta_j^*)} \\ &= \frac{1}{N_2} \sum_{i=1}^{N_2} p(\mathbf{y}|\tilde{\theta}_i) \quad \tilde{\theta}_i \sim p(\theta).\end{aligned}\tag{7.15}$$

If $h(\theta) = 1/g(\theta)$, the importance sampling estimate (7.5) is recovered,

$$\begin{aligned}\hat{p}^{BS}(\mathbf{y}) &= \frac{\frac{1}{N_2} \sum_{i=1}^{N_2} p(\mathbf{y}|\tilde{\theta}_i) p(\tilde{\theta}_i) \frac{1}{g(\tilde{\theta}_i)}}{\frac{1}{N_1} \sum_{j=1}^{N_1} \frac{1}{g(\theta_j^*)} g(\theta_j^*)} \\ &= \frac{1}{N_2} \sum_{i=1}^{N_2} \frac{p(\mathbf{y}|\tilde{\theta}_i) p(\tilde{\theta}_i)}{g(\tilde{\theta}_i)}, \quad \tilde{\theta}_i \sim g(\theta).\end{aligned}\tag{7.16}$$

If $h(\theta) = \frac{1}{p(\theta)p(\mathbf{y}|\theta)}$, the reciprocal importance sampling estimator (7.7) is recovered,

$$\begin{aligned}\hat{p}^{BS}(\mathbf{y}) &= \frac{\frac{1}{N_2} \sum_{i=1}^{N_2} p(\mathbf{y}|\tilde{\theta}_i) p(\tilde{\theta}_i) \frac{1}{p(\tilde{\theta}_i)p(\mathbf{y}|\tilde{\theta}_i)}}{\frac{1}{N_1} \sum_{j=1}^{N_1} \frac{1}{p(\theta_j^*)p(\mathbf{y}|\theta_j^*)} g(\theta_j^*)} \\ &= \frac{1}{\frac{1}{N_1} \sum_{j=1}^{N_1} \frac{g(\theta_j^*)}{p(\theta_j^*)p(\mathbf{y}|\theta_j^*)}}, \quad \theta_j^* \sim p(\theta|\mathbf{y}).\end{aligned}\tag{7.17}$$

7.2.2 Uncertainty quantification

Next, we introduce the approach for measuring the variability of the marginal likelihood estimate using bridge sampling, and this prompts us to quantify uncertainty associated with the estimate. A standard method is to repeat the evaluation multiple times and calculate the sample standard deviation. This is the most straightforward way to assess the uncertainty though it might be computationally expensive. Another way is to approximate the expected relative mean-square error of the estimate (Gronau et al., 2017a). Specifically, the relative mean-square error (RMSE), also called the relative variance, of a marginal likelihood estimate is

$$\text{RMSE} = \frac{\mathbb{E}[(\hat{p}(\mathbf{y}) - p(\mathbf{y}))^2]}{p^2(\mathbf{y})},\tag{7.18}$$

where $p(\mathbf{y})$ is the groundtruth marginal likelihood, and $\hat{p}(\mathbf{y})$ is the estimate of marginal likelihood by either approaches. The error measures the stability and precision of the bridge sampling estimate. Frühwirth-Schnatter (2004) developed an approximation to the RMSE for the bridge sampler in (7.9)

$$\widehat{\text{RMSE}} = \frac{1}{N_2} \frac{\mathbb{V}_g(h(\theta)p(\theta|\mathbf{y}))}{\mathbb{E}_p^2(h(\theta)g(\theta))} + \frac{\rho_f(0)}{N_1} \frac{\mathbb{V}_p(h(\theta)g(\theta))}{\mathbb{E}_p^2(h(\theta)g(\theta))}, \quad (7.19)$$

where $f(\theta^m) = h(\theta^m)g(\theta^m)$ is a process over m and $\rho_f(0)$ denotes the normalised spectral density of the process f at the frequency 0. The form of the (7.19) suggests that this approximation considers the auto-correlation between posterior samples. For the independent θ , $\rho_f(0)$ takes the value of 1. $\mathbb{E}$ and $\mathbb{V}$ refer to the expectation and variance, and the subscripts indicate the distribution that the expectations and variances are computed over. Note that we abbreviate the expectation with respect to the posterior distribution $p(\theta|\mathbf{y})$ rather than the prior as $\mathbb{E}_p$.

As special cases of bridge sampling, we can derive the relative mean square errors for the estimates (7.3), (7.5), (7.7) and (7.11). Equation (7.3) can be considered a special case of the bridge sampling setup when $h(\theta) = \frac{1}{p(\theta)}$ and $g(\theta) = p(\theta)$, and the RMSE is

$$\begin{aligned} \widehat{\text{RMSE}}^{\text{MC}} &= \frac{1}{N_2} \frac{\mathbb{V}_g(\frac{1}{p(\theta)}p(\theta|\mathbf{y}))}{\mathbb{E}_p^2(\frac{1}{p(\theta)}p(\theta))} + \frac{\rho_f(0)}{N_1} \frac{\mathbb{V}_p(\frac{1}{p(\theta)}p(\theta))}{\mathbb{E}_p^2(\frac{1}{p(\theta)}p(\theta))} \\ &= \frac{1}{N_2} \mathbb{V}_g\left(\frac{p(\theta|\mathbf{y})}{p(\theta)}\right). \end{aligned} \quad (7.20)$$

Equation (7.5) is a special case of the bridge sampling setup when $h(\theta) = \frac{1}{g(\theta)}$, and the RMSE is

$$\begin{aligned} \widehat{\text{RMSE}}^{\text{IS}} &= \frac{1}{N_2} \frac{\mathbb{V}_g(\frac{1}{g(\theta)}p(\theta|\mathbf{y}))}{\mathbb{E}_p^2(\frac{1}{g(\theta)}g(\theta))} + \frac{\rho_f(0)}{N_1} \frac{\mathbb{V}_p(\frac{1}{g(\theta)}g(\theta))}{\mathbb{E}_p^2(\frac{1}{g(\theta)}g(\theta))} \\ &= \frac{1}{N_2} \mathbb{V}_g\left(\frac{p(\theta|\mathbf{y})}{g(\theta)}\right). \end{aligned} \quad (7.21)$$

Equation (7.7) is a special case of the bridge sampling setup when $h(\theta) = \frac{1}{p(\theta)p(\mathbf{y}|\theta)}$, and the RMSE is

$$\begin{aligned}\widehat{\text{RMSE}}^{\text{HM}} &= \frac{1}{N_2} \frac{\mathbb{V}_g\left(\frac{1}{p(\mathbf{y}|\theta)p(\theta)}p(\theta|\mathbf{y})\right)}{\mathbb{E}_p^2\left(\frac{1}{p(\mathbf{y}|\theta)p(\theta)}g(\theta)\right)} + \frac{\rho_f(0)}{N_1} \frac{\mathbb{V}_p\left(\frac{1}{p(\mathbf{y}|\theta)p(\theta)}g(\theta)\right)}{\mathbb{E}_p^2\left(\frac{1}{p(\mathbf{y}|\theta)p(\theta)}g(\theta)\right)} \\ &= \frac{\rho_{f_1}(0)}{N_1} \frac{\mathbb{V}_p\left(\frac{1}{p(\theta|\mathbf{y})p(\mathbf{y})}g(\theta)\right)}{\mathbb{E}_g^2\left(\frac{1}{p(\theta|\mathbf{y})p(\mathbf{y})}p(\theta|\mathbf{y})\right)} \\ &= \frac{\rho_{f_1}(0)}{N_1} \mathbb{V}_p\left(\frac{g(\theta)}{p(\theta|\mathbf{y})}\right),\end{aligned}\quad (7.22)$$

where

$$f_1(\theta) = \frac{g(\theta)}{p(\mathbf{y}|\theta)p(\theta)}. \quad (7.23)$$

Next the RMSE under the optimal bridge function (7.10) is

$$\begin{aligned}\widehat{\text{RMSE}}^{\text{BS}} &= \frac{1}{N_2} \frac{\mathbb{V}_g\left(\frac{p(\theta|\mathbf{y})}{s_1p(\mathbf{y}|\theta)p(\theta)+s_2p(\mathbf{y})g(\theta)}\right)}{\mathbb{E}_p^2\left(\frac{g(\theta)}{s_1p(\mathbf{y}|\theta)p(\theta)+s_2p(\mathbf{y})g(\theta)}\right)} + \frac{\rho_f(0)}{N_1} \frac{\mathbb{V}_p\left(\frac{g(\theta)}{s_1p(\mathbf{y}|\theta)p(\theta)+s_2p(\mathbf{y})g(\theta)}\right)}{\mathbb{E}_p^2\left(\frac{g(\theta)}{s_1p(\mathbf{y}|\theta)p(\theta)+s_2p(\mathbf{y})g(\theta)}\right)} \\ &= \frac{1}{N_2} \frac{\mathbb{V}_g\left(\frac{p(\theta|\mathbf{y})}{s_1p(\theta|\mathbf{y})+s_2g(\theta)}\right)}{\mathbb{E}_g^2\left(\frac{p(\theta|\mathbf{y})}{s_1p(\theta|\mathbf{y})+s_2g(\theta)}\right)} + \frac{\rho_{f_2}(0)}{N_1} \frac{\mathbb{V}_p\left(\frac{g(\theta)}{s_1p(\theta|\mathbf{y})+s_2g(\theta)}\right)}{\mathbb{E}_p^2\left(\frac{g(\theta)}{s_1p(\theta|\mathbf{y})+s_2g(\theta)}\right)},\end{aligned}\quad (7.24)$$

where

$$f_2(\theta) = \frac{g(\theta)}{s_1p(\mathbf{y}|\theta)p(\theta)+s_2p(\mathbf{y})g(\theta)}. \quad (7.25)$$

We applied the same method as in Gronau et al. (2017b) and evaluated $\rho_f(0)$ by coda which fits a spectral density at frequency zero through an autoregressive model f_2 . This approximation prompts us to evaluate the RMSE along with the bridge sampler (7.12), because all the quantities in (7.24) are easy to evaluate.

The relative mean square errors (7.20)-(7.24) indicate the relative advantage of bridge sampling (Frühwirth-Schnatter, 2004). If $g(\theta)$ has a thinner tail than $p(\theta|\mathbf{y})$, then $g(\theta) \ll p(\theta|\mathbf{y})$ in the tails,

$$\begin{aligned}\frac{p(\theta|\mathbf{y})}{g(\theta)} &\rightarrow \infty, & \frac{g(\theta)}{p(\theta|\mathbf{y})} &\rightarrow 0, \\ \frac{p(\theta|\mathbf{y})}{s_1p(\theta|\mathbf{y})+s_2g(\theta)} &\rightarrow \frac{1}{s_1}, & \frac{g(\theta)}{s_1p(\theta|\mathbf{y})+s_2g(\theta)} &\rightarrow 0,\end{aligned}\quad (7.26)$$

as approaching the tail of $g(\theta)$. Consequently, (7.21) $\rightarrow \infty$, (7.22) $\rightarrow \infty$ and (7.24) $\rightarrow 0$, and the RMSE of the importance sampling estimate (7.21) is thus much more inflated than (7.22) and (7.24). Thus, importance sampling assumes the importance density q satisfies $q(\theta) = 0$ implies $p(\theta|\mathbf{y}) = 0$, i.e. absolute continuity. In contrast, if $g(\theta)$ has a heavier tail than $p(\theta|\mathbf{y})$, the harmonic mean estimate has a high standard error (7.22) and is less stable than (7.21) and (7.24).

7.2.3 Wrap-III bridge sampling

In addition, Meng and Schilling (2002) proved that the asymptotic relative error (7.18) is governed by the distance between the posterior $p(\theta|\mathbf{y})$ and the proposal $g(\theta)$, and thus the marginal likelihood estimate is the most accurate if the distance between the posterior and proposal is small and the overlap is strong. When the posterior distribution is highly skewed, the multivariate normal distribution is not the most appropriate choice as a proposal distribution. To handle this problem, it was proposed to use the standard multivariate normal proposal and manipulate the posterior samples so that the first three moments of the two densities match, which is called the warp-III bridge sampler (Gronau et al., 2019a; Meng and Schilling, 2002).

Recall that we drew N_1 samples from the posterior distribution, and the samples were denoted as θ^* in (7.9). For simplicity of demonstration, we suppose all the parameters take values on the real line. We abbreviated the unnormalised posterior distribution as $q(\theta) = p(\mathbf{y}|\theta)p(\theta)$. If the posterior samples follow a multivariate normal distribution, we denote the mean vector and covariance matrix to be $\hat{\mu}_\theta$ and $\hat{\Sigma}_\theta$ respectively, which can be estimated by MLE, and the Cholesky decomposition of $\hat{\Sigma}_\theta$ as $\mathbf{L}_{\hat{\Sigma}_\theta}$. The standardised parameters are

$$\theta^{II} = \mathbf{L}_{\hat{\Sigma}_\theta}^{-1}(\theta^* - \hat{\mu}_\theta), \quad (7.27)$$

where θ^{II} refers to the warp-II transformation, i.e. the mean and variance of transformed parameters match those of standard normal distributions, and its density is

$$q(\theta^{II}) = |\mathbf{L}_{\hat{\Sigma}_\theta}|q(\theta^*) = |\mathbf{L}_{\hat{\Sigma}_\theta}|q(\mathbf{L}_{\hat{\Sigma}_\theta}\theta^{II} + \hat{\mu}_\theta). \quad (7.28)$$

To further remove the skewness of the posterior samples, we followed Meng and Schilling (2002) and applied a symmetric transformation,

$$q(\theta) = \frac{q(|\theta|)}{2}, \quad \theta \in \mathbb{R}. \quad (7.29)$$

The density of warp-III transformed parameters θ^{III} is

$$\begin{aligned} q(\theta^{III}) &= |\mathbf{L}_{\hat{\Sigma}_\theta}| \frac{q(\mathbf{L}_{\hat{\Sigma}_\theta} \theta^{III} + \hat{\mu}_\theta)}{2} \\ &= \frac{|\mathbf{L}_{\hat{\Sigma}_\theta}|}{2} [q(\mathbf{L}_{\hat{\Sigma}_\theta} \theta^{III} + \hat{\mu}_\theta) + q(-\mathbf{L}_{\hat{\Sigma}_\theta} \theta^{III} + \hat{\mu}_\theta)]. \end{aligned} \quad (7.30)$$

Recall from (7.9) and (7.10) that the bridge sampler under the optimal bridge function was given by

$$\begin{aligned} p^{WBS}(\mathbf{y}) &= \frac{\int \frac{p(\mathbf{y}|\theta)p(\theta)}{s_1 p(\mathbf{y}|\theta)p(\theta) + s_2 p(\mathbf{y})g(\theta)} g(\theta) d\theta}{\int \frac{g(\theta)}{s_1 p(\mathbf{y}|\theta)p(\theta) + s_2 p(\mathbf{y})g(\theta)} p(\theta|\mathbf{y}) d\theta} \\ &= \frac{\int \frac{\frac{p(\mathbf{y}|\theta)p(\theta)}{g(\theta)}}{s_1 \frac{p(\mathbf{y}|\theta)p(\theta)}{g(\theta)} + s_2 p(\mathbf{y})} g(\theta) d\theta}{\int \frac{1}{s_1 \frac{p(\mathbf{y}|\theta)p(\theta)}{g(\theta)} + s_2 p(\mathbf{y})} p(\theta|\mathbf{y}) d\theta}. \end{aligned} \quad (7.31)$$

As in Sinharay and Stern (2005) and Gronau et al. (2019b), we substituted the warped posterior distribution (7.30) into (7.31). It is straightforward to evaluate the numerator in (7.31) because we can generate samples from the standard normal proposal distribution,

$$\int \frac{\frac{|\mathbf{L}_{\hat{\Sigma}_\theta}|}{2} [q(\mathbf{L}_{\hat{\Sigma}_\theta} \theta + \hat{\mu}_\theta) + q(-\mathbf{L}_{\hat{\Sigma}_\theta} \theta + \hat{\mu}_\theta)]}{g(\theta)} g(\theta) d\theta. \quad (7.32)$$

It is challenging to evaluate the denominator as the expectation is with respect to the warped posterior distribution. For convenience we prefer to sample from the unwrapped posterior distribution. Thus, we transformed the denominator back by assuming $\eta = \mathbf{L}_{\hat{\Sigma}_\theta} \theta + \hat{\mu}_\theta$, and the resulting denominator is

$$\int \frac{1}{s_1 \frac{|\mathbf{L}_{\hat{\Sigma}_\theta}|}{2} [q(\eta) + q(2\hat{\mu}_\theta - \eta)] + s_2 p(\mathbf{y})} q(\eta) d\eta, \quad (7.33)$$

where details are in Appendix D.

Therefore, the iterative algorithm of the warp-III bridge sampling follows a similar form as,

$$\hat{p}_{t+1}^{\text{WBS}}(\mathbf{y}) = \frac{\frac{1}{N_2} \sum_{i=1}^{N_2} \frac{l_{2,i}}{s_1 l_{2,i} + s_2 \hat{p}_t^{\text{WBS}}(\mathbf{y})}}{\frac{1}{N_1} \sum_{j=1}^{N_1} \frac{1}{s_1 l_{1,j} + s_2 \hat{p}_t^{\text{WBS}}(\mathbf{y})}}, \quad (7.34)$$

where

$$\begin{aligned} l_{1,i} &= \frac{\frac{|\mathbf{L}_{\hat{\Sigma}_\theta}|}{2} [q(\tilde{\theta}_i) + q(2\hat{\mu}_\theta - \tilde{\theta}_i)]}{g(\mathbf{L}_{\hat{\Sigma}_\theta}^{-1}(\tilde{\theta}_i - \hat{\mu}_\theta))} \\ l_{2,j} &= \frac{\frac{|\mathbf{L}_{\hat{\Sigma}_\theta}|}{2} [q(\mathbf{L}_{\hat{\Sigma}_\theta} \theta_j^* + \hat{\mu}_\theta) + q(-\mathbf{L}_{\hat{\Sigma}_\theta} \theta_j^* + \hat{\mu}_\theta)]}{g(\theta_j^*)}. \end{aligned} \quad (7.35)$$

where $\tilde{\theta}_i$ is generated from the unwarped posterior distribution, and θ_j^* is drawn from the standard normal proposal distribution.

Gronau et al. (2017b) suggested that the approximation of the relative mean square error (7.19) is not reliable for the warp-III bridge sampling in their experience. In particular, the RMSE follows the same equation as (7.24), and

$$\begin{aligned} f_2(\theta) &= \frac{g(\theta)}{s_1 p(\mathbf{y}|\theta) p(\theta) + s_2 p(\mathbf{y}) g(\theta)} \\ &= \frac{g(\theta)}{s_1 \frac{|\mathbf{L}_{\hat{\Sigma}_\theta}|}{2} [q(\mathbf{L}_{\hat{\Sigma}_\theta} \theta + \hat{\mu}_\theta) + q(-\mathbf{L}_{\hat{\Sigma}_\theta} \theta + \hat{\mu}_\theta)] + s_2 p(\mathbf{y}) g(\theta)}. \end{aligned} \quad (7.36)$$

Gronau et al. (2017b) indicated that $q(\mathbf{L}_{\hat{\Sigma}_\theta} \theta + \hat{\mu}_\theta) + q(-\mathbf{L}_{\hat{\Sigma}_\theta} \theta + \hat{\mu}_\theta)$ may change the autocorrelation structure of posterior samples, and this makes f_2 no longer an autoregressive model. The approximation for the RMSE of the warp-III bridge sampler will be accurate if the normalised spectral density of (7.36) at 0 is reliably estimated through alternative approaches. Here we follow the same practice as Gronau et al. (2017b). To assess the uncertainty of the marginal likelihood estimate obtained through warp-III bridge sampling, Gronau et al. (2017b) recommended to repeat the warp-III bridge sampling several times, and to calculate the interquartile range and the range of the marginal likelihood estimate.

7.3 Implementation of the bridge sampling algorithm

In principle, `bridgesampling` provides an automatic evaluation of the marginal likelihood of a Bayesian model formulated in a Stan text file. However, it is not straightforward to apply it in a Gaussian process based model (also see Section 4.4). Then we implemented the bridge sampling algorithm in R 3.6.0. The codes are available at <https://github.com/BDI-pathogens/phyloscanner/blob/master/phyloflows/inst/misc/bs.R> for the standard bridge sampler and <https://github.com/BDI-pathogens/phyloscanner/blob/master/phyloflows/inst/misc/bsw.R> for the warp-III bridge sampler.

To implement the bridge sampling algorithm, we first set an arbitrary initial guess of the marginal likelihood, and then update the estimates through (7.12). For example, we set an initial guess of the log marginal likelihood to -2000 , but any initial guesses are allowed here.

To explain, we took (6.2) as an example. For completeness, we recall from Section 6.2.1 that its likelihood function is

$$p(\mathbf{y}|\theta) = p(\mathbf{n}|\lambda, \xi) = \left(\prod_{a,b} \text{Poisson}(n_{ab}; \lambda_{ab} \xi_a \xi_b) \right), \quad (7.37)$$

and the prior distribution is

$$p(\theta) = p(\mu, \nu, \xi, \ell, \sigma, \beta) = \text{Normal}(\mu; 0, 10^2) \text{Normal}(\nu; 0, 10^2) p(\xi; \mathbf{s}, \mathbf{X}) \\ \text{Normal}(\beta; 0, 10^2) \text{Half-Normal}(\sigma; 0, 10^2) \text{Inverse-Gamma}(\ell; \alpha, \gamma), \quad (7.38)$$

To work on (7.12), the first step is to determine the proposal distribution. Given the large parameter spaces in (7.37) with 2088 parameters, we chose a larger sample size than the parameter space to characterise the posterior distribution. To be precise, we used 10,000 posterior samples of parameters, transformed them into the real line, assumed the transform parameters to follow a multivariate normal distribution as described in Section 7.2.1, and estimated the parameters of the distribution through MLE. In (6.2), the set of model parameters includes $\theta = \{\mu, \nu, \xi, \ell, \sigma\}$, and we applied the log transformation to ℓ and σ , and the logit transformation to ξ . To find the MLE of the parameters of multivariate normal

distribution, $\hat{\mu}_\theta$ and $\hat{\Sigma}_\theta$, we used Rfast version 2.0.3. The proposal distribution is then

$$\begin{aligned}
 g(\theta) &= g(\mu, \nu, \xi, \ell, \sigma) \\
 &= \text{Multivariate Normal}((\mu, \nu, \xi^*, \ell^*, \sigma^*); \hat{\mu}_\theta, \hat{\Sigma}_\theta) \\
 &\quad \Pi_i \frac{d\text{logit}(\xi_i)}{d\xi_i} \Pi_i \frac{d\log(\ell_i)}{d\ell_i} \Pi_i \frac{d\log(\sigma_i)}{d\sigma_i} \\
 &= \text{Multivariate Normal}((\mu, \nu, \xi^*, \ell^*, \sigma^*); \hat{\mu}_\theta, \hat{\Sigma}_\theta) \Pi_i \left(\frac{1}{\xi_i} + \frac{1}{1-\xi_i} \right) \Pi_i \frac{1}{\sigma_i} \Pi_i \frac{1}{\ell_i},
 \end{aligned} \tag{7.39}$$

where the transformed parameters are denoted by adding a superscript asterisk, $\xi^* = \text{logit}(\xi)$, $\ell^* = \log(\ell)$ and $\sigma^* = \log(\sigma)$. After obtaining the proposal distribution, we followed the same practice as Gronau et al. (2017a) and discarded the Monte Carlo samples used for specifying the multivariate normal proposal distribution in the subsequent marginal likelihood estimation.

Note that in Chapter 6, we made inference on (6.2) for 10^4 iterations, and achieved effective sample sizes $> 5 \times 10^3$. However, we need to run the chains for longer and obtain more samples to estimate the marginal likelihood, because the first 10^4 samples were all used to determine the proposal distribution. Therefore, we ran Stan's no-U-turn sampler (NUTS) algorithm for another 11×10^4 post-warm-up iterations. Then we are ready to apply (7.13) to update the estimate of the marginal likelihood from our initial guess. Each update requires the samples drawn from the posterior distribution and the proposal as in (7.13), and we decided to use 10^4 samples per iteration again for safety, taking a large number of parameters into account. These choices prompt us to update the marginal likelihood 11 times. The samples from the posterior distribution are directly from the Hamiltonian Monte Carlo (HMC), and the samples from the proposal distribution are generated using (7.39). As the bridge function (7.10) is optimal under the assumption of independent posterior samples, we used the median effective sample size of posterior samples from the model as N_1 . It is straightforward to evaluate all the other terms in (7.13). The first update from the initial guess $\log(p_0^{BS}(\mathbf{y}))$ was written explicitly in (7.14) for example.

We then check whether the iterative algorithm converges to a stable estimate, and to assess the uncertainty associated with the estimate (7.24). To measure the convergence of the algorithm, we computed the differences of the log marginal likelihood estimates in the two consecutive steps t and $t + 1$,

$$|\log(\hat{p}_{t+1}^{BS}(\theta)) - \log(\hat{p}_t^{BS}(\theta))| = \left| \log\left(\frac{\hat{p}_{t+1}^{BS}(\theta)}{\hat{p}_t^{BS}(\theta)}\right) \right|, \quad t = 0, \dots, 10. \tag{7.40}$$

As a diagnostic, we expect (7.40) to drop over the bridge sampling iterations, because $\hat{p}_t^{BS}(\theta)$ and $\hat{p}_{t+1}^{BS}(\theta)$ become more and more similar as the algorithm converges. Plotting (7.40) against the bridge sampling iterations is a useful indicator of convergence. Upon the convergence, we obtained the value that the algorithm converges to, and considered it as the marginal likelihood estimate. The RMSE of the estimate was approximated by (7.24), and reflects the uncertainty of the estimate.

To perform warp-III bridge sampling, we followed the same procedure as discussed for the standard bridge sampling with two minor differences. First, the $l_{1,i}$ and $l_{2,i}$ in (7.13) are now set as specified in (7.35). Second, the draws from the proposal follow the standard normal distribution. Therefore, the two changes usually do not require dramatically more computational efforts for the warp-III bridge sampling.

7.4 Comparison of models that describe age-specific transmission flows

In this Section, we explore the seven candidate models which describe age-specific transmission flows, evaluate the marginal likelihood using the bridge sampler and the warp-III bridge sampler in Section 7.3, and make a comparison. The seven competing models that we investigate are variations of (6.2). The first model that we considered is (6.2) that assumes that the age-specific pattern and the baseline transmission intensities vary between the male-to-female and female-to-male transmissions,

$$M_1 : \log \lambda = \mu \mathbf{1} + v \mathbb{1}_{mf} + \mathbb{1}_{mf} f_{mf} + \mathbb{1}_{fm} f_{fm}, \quad (7.41)$$

where $\lambda \in \mathbb{R}^{24K^2}$ represents the transmission intensities by age, gender and location, μ is the baseline log intensity for the transmission from female to male, v is the difference in the baseline log intensities between male-to-female and female-to-male transmissions, f_{mf} and f_{fm} are two functions of ages of sources and recipients, for male-to-female and female-to-male respectively. Then we summarised the model parameters by $\theta = (\mu, v, \xi, \ell, \sigma, \beta)$ where $\mu \in \mathbb{R}$, $v \in \mathbb{R}$, $\text{logit}(\xi) \in \mathbb{R}^{280}$, $\beta \in \mathbb{R}^{m \times 2}$, $\log(\ell) \in \mathbb{R}^{2 \times 2}$, $\log(\sigma) \in \mathbb{R}^2$. The second model relaxed the assumption of constant log baselines intensities across the locations,

$$M_2 : \log \lambda = \mu \mathbf{1} + v \mathbb{1}_{mf} + \sum_x \mu_x \mathbb{1}_x + \mathbb{1}_{mf} f_{mf} + \mathbb{1}_{fm} f_{fm}, x = \text{eh, el, hh, hl, lh}, \quad (7.42)$$

where μ is the log baseline intensity for the transmissions from female in inland to male in inland, μ_x represents the differences in the log baseline intensities between x and within-inland transmissions. This means the transmission intensities by age have the same shape for male-to-female transmissions, but the whole surface might be shifted up and down by combinations of the locations. For example, the transmission intensities within the fishing sites are higher than those between inland and fishing communities. The model contains the parameters $\mu \in \mathbb{R}^6$, $\nu \in \mathbb{R}$, $\text{logit}(\xi) \in \mathbb{R}^{280}$, $\beta \in \mathbb{R}^{m \times 2}$, $\log(\ell) \in \mathbb{R}^{2 \times 2}$, $\log(\sigma) \in \mathbb{R}^2$. Next, we considered the third and fourth models as the analogy to the first and second models in terms of the baseline intensities. The change is to allow the varying age dynamics by both gender and the location of recipients, and the resulting models are given by

$$M_3 : \log \lambda = \mu \mathbf{1} + \nu \mathbb{1}_{mf} + \mathbb{1}_{mf} f_{mf} + \mathbb{1}_{fm} f_{fm} + \mathbb{1}_{.h} f_{.h} + \mathbb{1}_{.l} f_{.l}, \quad (7.43)$$

$$M_4 : \log \lambda = \mu \mathbf{1} + \nu \mathbb{1}_{mf} + \sum_x \mu_x \mathbb{1}_x + \mathbb{1}_{mf} f_{mf} + \mathbb{1}_{fm} f_{fm} + \mathbb{1}_{.h} f_{.h} + \mathbb{1}_{.l} f_{.l}, \quad (7.44)$$

$$x = eh, el, hh, hl, lh,$$

where $f_{.h}$ and $f_{.l}$ represent the two functions of the ages for sources and recipient in a pair of strata where the flow destination is in the high and low prevalence locations respectively, i. e. fishing and inland. The resulting model parameters are similar as (7.41) and (7.42), except for the terms related to $f_{.h}$ and $f_{.l}$. That is, $\mu \in \mathbb{R}$, $\nu \in \mathbb{R}$, $\text{logit}(\xi) \in \mathbb{R}^{280}$, $\beta \in \mathbb{R}^{m \times 4}$, $\log(\ell) \in \mathbb{R}^{4 \times 2}$, $\log(\sigma) \in \mathbb{R}^4$ for M_3 in (7.43), and $\mu \in \mathbb{R}^6$, $\nu \in \mathbb{R}$, $\text{logit}(\xi) \in \mathbb{R}^{280}$, $\beta \in \mathbb{R}^{m \times 4}$, $\log(\ell) \in \mathbb{R}^{4 \times 2}$, $\log(\sigma) \in \mathbb{R}^4$ for M_4 in (7.44). The structure of baseline intensities also applies to the fifth and sixth models, except we make the age-specific transmission intensities more flexible and further dependant on the location of sources,

$$M_5 : \log \lambda = \mu \mathbf{1} + \nu \mathbb{1}_{mf} + \mathbb{1}_{mf} f_{mf} + \mathbb{1}_{fm} f_{fm} + \mathbb{1}_{.h} f_{.h} + \mathbb{1}_{.l} f_{.l} \quad (7.45)$$

$$+ \mathbb{1}_{e.} f_{e.} + \mathbb{1}_{h.} f_{h.} + \mathbb{1}_{l.} f_{l.}$$

$$M_6 : \log \lambda = \mu \mathbf{1} + \nu \mathbb{1}_{mf} + \sum_x \mu_x \mathbb{1}_x + \mathbb{1}_{mf} f_{mf} + \mathbb{1}_{fm} f_{fm} + \mathbb{1}_{.h} f_{.h} + \mathbb{1}_{.l} f_{.l} \quad (7.46)$$

$$+ \mathbb{1}_{e.} f_{e.} + \mathbb{1}_{h.} f_{h.} + \mathbb{1}_{l.} f_{l.}$$

$$x = eh, el, hh, hl, lh,$$

where $f_{h.}$, $f_{l.}$ and $f_{e.}$ represent the functions of the ages of sources and recipient for the pair of strata where the origin of flow is in the high prevalence locations, low prevalence locations and external regions respectively. The model parameters are $\mu \in \mathbb{R}$, $\nu \in \mathbb{R}$, $\text{logit}(\xi) \in \mathbb{R}^{280}$, $\beta \in \mathbb{R}^{m \times 7}$, $\log(\ell) \in \mathbb{R}^{7 \times 2}$, $\log(\sigma) \in \mathbb{R}^7$ for M_5 in (7.45), and $\mu \in \mathbb{R}^6$, $\nu \in \mathbb{R}$, $\text{logit}(\xi) \in \mathbb{R}^{280}$,

$\beta \in \mathbb{R}^{m \times 7}$, $\log(\ell) \in \mathbb{R}^{7 \times 2}$, $\log(\sigma) \in \mathbb{R}^7$ for M_6 in (7.46), as there are seven GP components in both models. The final model we considered is a baseline model, which assumes no differences in the age-specific dynamics by gender and location,

$$M_7 : \log \lambda = \mu \mathbf{1} + f, \quad (7.47)$$

where μ is the overall baseline of transmission intensities, and f is a function representing the deviations from the baseline. The model parameters contain $\mu \in \mathbb{R}$, $\text{logit}(\xi) \in \mathbb{R}^{280}$, $\beta \in \mathbb{R}^{m \times 1}$, $\log(\ell) \in \mathbb{R}^{1 \times 2}$, $\log(\sigma) \in \mathbb{R}$.

To compare these models, we use bridge sampling in Section 7.3 to estimate the marginal likelihood associated with each model. We first applied the standard bridge sampling, which took 1.21 (0.89 - 1.77) hours. When the model comprised several GP components, especially as in models (7.45) and (7.46), the marginal likelihood estimates fluctuated dramatically between bridge sampling iterations. Figure 7.1 (left) shows the log marginal likelihood estimates against bridge sampling iterations, and Figure 7.1 (right) shows the corresponding measures of convergence (7.40). Therefore, we cannot obtain the marginal likelihood estimate from the standard bridge sampling. This suggests more posterior samples are required to guarantee the convergence. However, we do not have more computational power to generate more posterior samples, as it has already taken 74.09 days to generate 12×10^4 posterior samples, and the samples occupy 114.43 gigabyte space for (7.41), the relatively simple model of interest.

As discussed in Section 7.3, it is often more efficient to apply the warp-III bridge sampling algorithm, when the multivariate normal distribution cannot well describe the posterior distribution. We applied the warp-III bridge sampling (7.30) to estimate the marginal likelihood of models (7.41)-(7.47), which took 2.07 (1.03 - 4.08) hours. Convergence was quickly achieved within 5 steps of the warp-III bridge sampling algorithm for all candidate models, as illustrated in Figure 7.2. This enabled us to continue with our ultimate aim, model comparison.

Figure 7.2 illustrates the log marginal likelihood of the seven candidate models in colours. We investigated the final bridge sampling estimates of the log marginal likelihoods and found that the model M_2 in (7.42), in which the age dynamics varies by gender and the transmission intensity baselines depend on both genders and locations, describes the data best across (7.41)-(7.46). In addition, Figure 7.2 shows that model M_1 (7.41) has marginally inferior performance to the model M_2 (7.42). The benchmark model (7.47) performs as well as (7.42) with regard to the log marginal likelihood. However, the marginal likelihood penalises

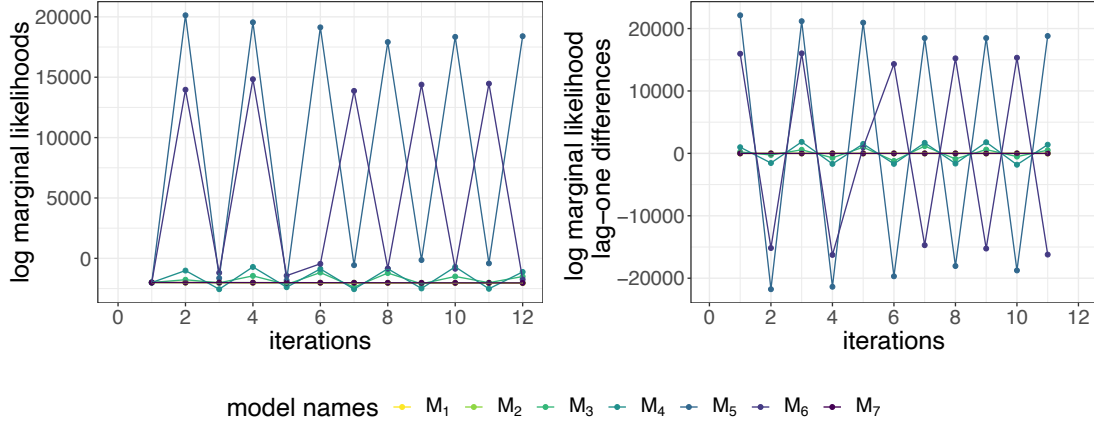


Figure 7.1 **Log marginal likelihood estimates for the candidate models over iterations, using the standard bridge sampling.** We estimated the log marginal likelihood using the standard bridge sampling scheme in (7.13). The estimates were updated along with the iterative scheme, and are shown against the iterations (in the x-axis). The legend specifies the model features, including the number of Gaussian process components (in the line types) and the number of baseline terms (in the colours).

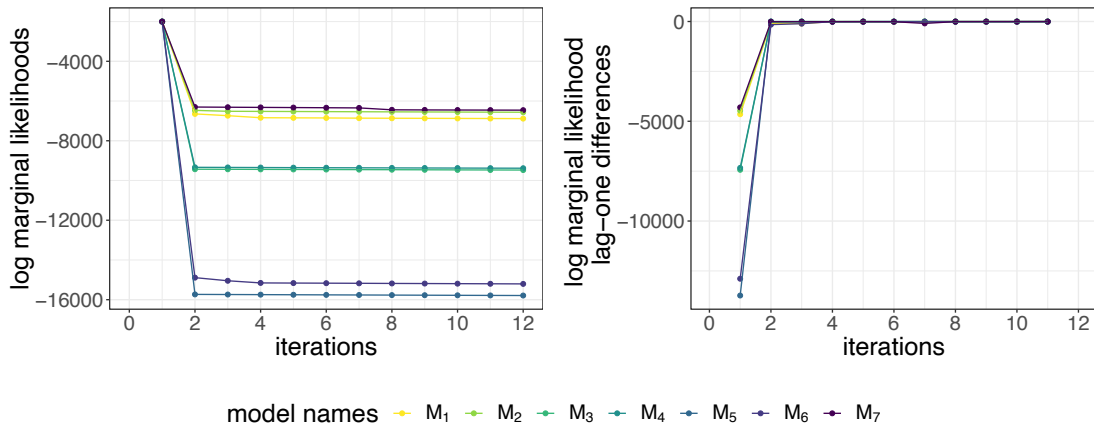


Figure 7.2 **Log marginal likelihood estimates for the candidate models over iterations, using the warp-III bridge sampling.** We estimated the log marginal likelihood using the warp-III bridge sampling in (7.34). The estimates were updated along with the iterative scheme, and are shown against the iterations (in the x-axis). The legend specifies the model features, including the number of Gaussian process components (in the line types) and the number of baseline terms (in the colours).

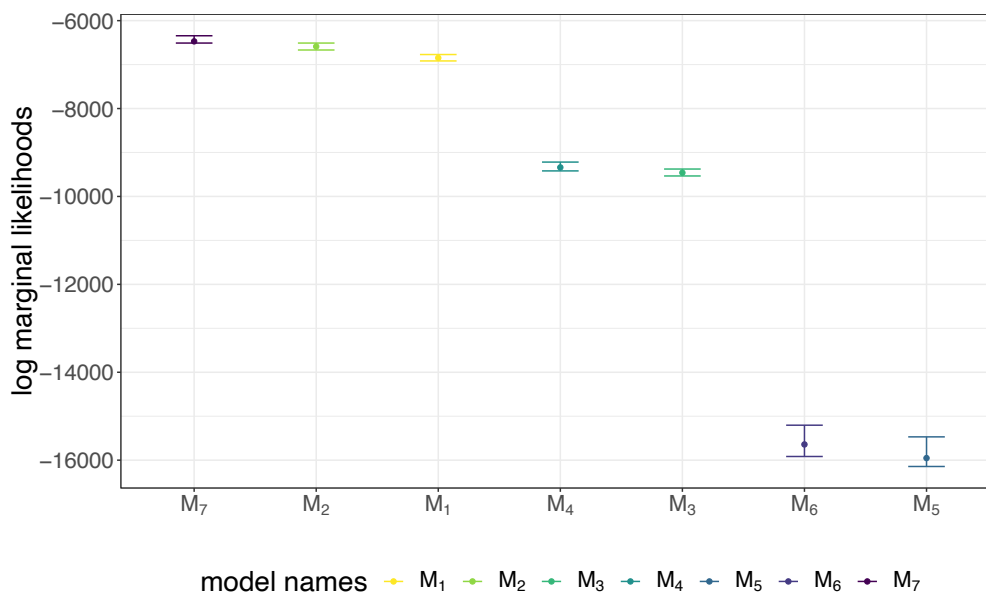


Figure 7.3 **Comparison of log marginal likelihood for the candidate models.** We evaluated the log marginal likelihoods using the warp-III bridge sampler for seven candidate models in (7.41)-(7.47), and show the median and range of the estimates in the points and error bars respectively. The colours represent models and are comparable to those in Figure 7.2.

the large model parameter space by its nature. Considering the enormous differences in the parameter space of about 1,000 and the minor improvement, we would be cautious to conclude the performance of the benchmark model.

As explained in Section 7.2.2, we follow the golden rule (Gronau et al., 2017b) and repeat bridge sampling procedure multiple times. However, we are not able to obtain more posterior samples due to the limited computational power. We chose to perform the estimation 20 times, based on the current posterior samples. Figure 7.3 shows the median and range of the log marginal likelihood estimates in (7.11). This approach suggests that model (7.41), (7.42) and (7.47) perform better than the other models and result in the highest log marginal likelihood. There are fewer differences between M_2 (7.42) and M_7 (7.47) than between M_1 (7.41) and M_2 (7.42).

7.5 Summary

The `bridgesampling` package is often a useful tool to evaluate model marginal likelihoods for arbitrary Stan models, though we did not manage to get it to run on our Gaussian process based models. Especially, our models comprise a large number of parameters, rang-

ing from 2,088 in (7.41) to 6,328 in (7.46), and even the baseline model in (7.47) has 1,184 parameters. Thus we implemented the algorithm ourselves. According to the large parameter space, we used 10^4 samples per update. In total, we generated 12×10^4 Hamiltonian Monte Carlo samples from the joint posterior distribution of all parameters in each model. This required massive computational resources. In this Chapter, we first attempted to use the standard bridge sampler with a multivariate normal proposal distribution for estimating the marginal log likelihoods. We were not able to achieve convergence given the samples. The warp-III bridge sampler made the proposal function more efficient by altering the posterior samples, and we were able to reach convergence in 5 bridge sampling iterations. This enabled us to compare various models of age-specific transmission dynamics. We found that the best model assumes the varying age-specific patterns between male-to-female and female-to-male transmissions, and the baseline intensities further depend on the locations of a source-recipient pair.

Chapter 8

COVID-19 flow model

8.1 Introduction

The main focus of this PhD thesis is to provide advanced methods for estimating human immunodeficiency virus (HIV) transmission flows between subpopulations from HIV deep sequence data. The analysis starts from reconstructing partially observed transmission networks through the bioinformatics tools, and the outputs were subsequently analysed using statistical models of transmission flows. Similar flow analyses are useful also in other contexts, such as the Coronavirus disease (COVID-19) pandemic. In particular, I contributed to analysing the role of age groups in driving the re-surfing COVID-19 epidemics in the US. The analyses described in this Chapter do not involve any viral sequences, and are based on detailed age-specific mortality data. My primary contribution to this work was in model development, and post-processing the model outcomes to reconstruct transmission flows as generated quantities. Section 8.2.1 describes the framework for the age-specific renewal model. Section 8.2.2 introduces the data that are used to construct the prior. Section 8.2.3 constructs the prior distributions for the model parameters. Section 8.2.4 defines the age-specific reproduction numbers and transmission flows that are deduced from the model parameters. Section 8.3 presents the results on these generated quantities. Section 8.4 compares the COVID-19 flow model with main theme of the thesis, and Section 8.5 ends with the summary.

8.2 Development of a Bayesian age-specific renewal model

8.2.1 Model framework

We analysed longitudinal and age-specific COVID-19 mortality data together with contact and mobility data, in order to explore the interaction between changes in human contact patterns and COVID-19 epidemics. To do this, we built the link between the daily deaths and cases, and described the cases by the changing contact patterns. In particular, we developed a discrete time age-specific renewal model for the case evolution, where the age is defined by stratifying the population into 5-year age groups

$$a \in A = \{[0-4], [5-9], \dots, [80-84], [85+]\}. \quad (8.1)$$

The starting point of this analysis is the reported COVID-19 age-specific daily deaths by states $D_{m,t,a}$, and we modelled them with a negative binomial distribution parameterised by the expected daily deaths and an overdispersion parameter ϕ . We denote the expected number of new deaths at location m , on day t and in age group a as $d_{m,t,a}$. The expected deaths and cases were linked through

$$d_{m,t,a} = \sum_{s=1}^{t-1} c_{m,t,a} h_{m,t-s,a}, \quad (8.2)$$

where $c_{m,t,a}$ represents the expected number of new cases at location m , on day t and in age group a , and $h_{m,t-s,a}$ is the probability of an individual aged a at location m dying after $t-s$ days of infections.

Then the age-specific new COVID-19 cases were described by an age-specific renewal equation with the time-changing contact intensities,

$$c_{m,t,a'} = s_{m,t,a'} \rho_{a'} \sum_a C_{m,t,a,a'} \sum_{s=1}^{t-1} c_{m,s,a} g(t-s), \quad (8.3)$$

where $s_{m,t,a'}$ is the proportion of susceptible individuals aged a' in location m at time t , $\rho_{a'}$ is the probability of transmissions upon contact to an individual in a' , $C_{m,t,a,a'}$ is the time-varying contact that one person in a has with another person in a' at location m in time t , and $g(t-s)$ describes the relative infectiousness $t-s$ days after infection. In particular, assuming that the previously infected individuals were immune to reinfection during the

analysis period,

$$s_{m,t,a'} = 1 - \frac{\sum_{s=1}^{t-1} c_{m,s,a'}}{N_{m,a'}}, \quad (8.4)$$

where $N_{m,a'}$ is the population size in the age group a' and location m . The generation time distribution g kept fixed and followed the same choice as in Flaxman et al. (2020).

8.2.2 Data source to inform the prior

Contact and mobility in adults

First, Mélodie Monod and Dr. Alexandra Blenkinsop in our research group predicted human contact matrices for the baseline period prior to the emergence of COVID-19 in each study location. This was used as the baseline pre-pandemic contacts that one person in a has with another person in a' in location m , denoted by $C_{m,a,a'}$. Next, Dr. Oliver Ratmann obtained and analysed age-specific mobility trend data of over 10 million individuals, and these data were collected from mobile phones through Foursquare's location technology, Pilgrim. Specifically, the mobility data were available for individuals aged over 20 years. We denoted the mobility trend for an individual aged a in the location m at time t as $X_{m,t,a}$ for $a \geq 20$, and the three covariates were derived from the mobility data as follows,

- a baseline mobility trend prior to the emergence of COVID-19

$$X_{m,t,a}^{\text{base}} = \begin{cases} X_{m,t,a}, & t < t_m^{\text{dip}}, \\ 1, & t \geq t_m^{\text{dip}}, \end{cases} \quad (8.5)$$

- an eased mobility trend which represents a permanent reduction in contacts during the early epidemic,

$$X_{m,t,a}^{\text{ease}} = \begin{cases} 1, & t < t_m^{\text{dip}}, \\ X_{m,t,a}, & t_m^{\text{dip}} \leq t < t_m^{\text{rebound}}, \\ \frac{X_{m,a}^{\text{weekday}}}{X_{m,a}}, & t \geq t_m^{\text{rebound}}, t \text{ is a weekday}, \\ \frac{X_{m,a}^{\text{weekend}}}{X_{m,a}}, & t \geq t_m^{\text{rebound}}, t \text{ is a weekend}, \end{cases} \quad (8.6)$$

- an upswing multiplier which describes long term increases in mobility after the initial reduction,

$$X_{m,t,a}^{\text{upswing}} = \begin{cases} 1, t < t_m^{\text{dip}}, \\ 1, t_m^{\text{dip}} \leq t < t_m^{\text{rebound}}, \\ X_{m,t,a} / \overline{X_{m,a}^{\text{weekday}}}, t \geq t_m^{\text{rebound}}, t \text{ is a weekday}, \\ X_{m,t,a} / \overline{X_{m,a}^{\text{weekend}}}, t \geq t_m^{\text{rebound}}, t \text{ is a weekend}, \end{cases} \quad (8.7)$$

where t_m^{dip} is the dip date when the per capita visits in the location m start to decrease, and t_m^{rebound} is the rebound date when the per capita visits reach the minimum, $\overline{X_{m,a}^{\text{weekday}}}$ is the average mobility trend $X_{m,t,a}$ over 5 weekdays before rebound, and $\overline{X_{m,a}^{\text{weekend}}}$ is the average mobility trend $X_{m,t,a}$ over 4 weekends before rebound. The decomposition satisfies $X_{m,t,a} = X_{m,t,a}^{\text{base}} X_{m,t,a}^{\text{ease}} X_{m,t,a}^{\text{upswing}}$, and prompts us to decouple the impact of mobility trend on contacts and thus characterise transmission risks in different stages of epidemics.

Infection-to-death distribution

The discretised infection-to-death distribution $h_{m,t-s,a}$ describes the probability of deaths, and we characterised the quantity through the infection fatality ratio (IFR) in location m and age band a , and infection-to-death distribution. Our specification of infection-to-death distribution is the same as Flaxman et al. (2020), and the age-specific IFR follows a recent meta-analysis (Levin et al., 2020).

Susceptibility

Zhang et al. (2020a) conducted a contact tracing and testing study in mainland China, and found the increasing susceptibility with age. Specifically, individuals aged 0 – 14 have lower risk of infections than those aged 15 – 64, with the odd ratio 0.34 [0.24 - 0.49], and individuals aged 65+ have higher susceptibility than those aged 15 – 64, with the odd ratio 1.47 [1.12 - 1.92].

8.2.3 Prior specification

Time-varying contact matrices

The baseline contacts $C_{m,a,a'}$ and the mobility data enabled us to model the contact matrix at time t through

$$C_{m,t,a,a'} = \eta_{m,t,a} C_{m,a,a'} \eta_{m,t,a'}, \quad (8.8)$$

where $\eta_{m,t,a}$ refers to the average reduction in contacts from an individual in group a in location m at time t and $\eta_{m,t,a'}$ denotes the reduction in contacted individual in a' . Then we modelled $\eta_{m,t,a}$ through a regression with three covariates derived from the mobility data. The three covariates, $X_{m,t,a}^{\text{base}}$, $X_{m,t,a}^{\text{ease}}$ and $X_{m,t,a}^{\text{upswing}}$, respectively represent the pre-pandemic, dip and rebound phases in mobility trends, and the multiplier $\eta_{m,t,a}$ was then described by

$$\eta_{m,t,a} = \exp(\log(X_{m,t,a}^{\text{base}}) + \beta_m^{\text{ease}} \log(X_{m,t,a}^{\text{ease}}) + \beta_{m,t,a}^{\text{upswing}} \log(X_{m,t,a}^{\text{upswing}})), \quad (8.9)$$

where β_m^{ease} and $\beta_{m,t,a}^{\text{upswing}}$ are regression coefficients and describe the mean changes in the multiplier $\eta_{m,t,a}$ for one unit changes in the log decomposed mobility trends $\log(X_{m,t,a}^{\text{ease}})$ and $\log(X_{m,t,a}^{\text{upswing}})$. We specified the spatial ease random effects based on the fact that a reduction of venue check-ins may imply a greater reduction of contacts in populous areas, and then

$$\begin{aligned} \beta_m^{\text{ease}} &\sim \text{Normal}(\beta^{\text{ease}}, \sigma_{\text{ease}}^2) \\ \beta^{\text{ease}} &\sim \text{Normal}(0, 1) \\ \sigma_{\text{ease}}^2 &\sim \text{Exponential}(1.5), \end{aligned} \quad (8.10)$$

The upswing random effects are designed to capture the potential disease-relevant, unobserved behavioural factors and investigate the association of the resurgent epidemics with young people. We decomposed the upswing random effect by

$$\beta_{m,t,a}^{\text{upswing}} = (\beta^{\text{upswing-base}} + \beta_{m,a}^{\text{upswing-age}}) \beta_{m,t}^{\text{upswing-time}}, \quad (8.11)$$

where $\beta^{\text{upswing-base}}$ is a baseline effect of mobility increases since the rebound time

$$\beta^{\text{upswing-base}} \sim \text{Half-Normal}(0, 1), \quad (8.12)$$

$\beta_{m,a}^{\text{upswing-age}}$ represents the age-specific random effects that test the impact of individuals aged 20-49 on transmission risks,

$$\begin{aligned}\beta_{m,[20-49]}^{\text{upswing-age}} &\sim \text{Exponential}(\beta_{[20-49]}^{\text{upswing-age}}) \\ \beta_{[20-49]}^{\text{upswing-age}} &\sim \text{Exponential}(0, 0.1) \\ \beta_{m,a}^{\text{upswing-age}} &= 0, a \notin [20-49],\end{aligned}\tag{8.13}$$

and $\beta_{m,t}^{\text{upswing-time}}$ is the time effect that allows the different effect on the contacts over consecutive two week intervals,

$$\begin{aligned}\beta_{m,t}^{\text{upswing-time}} &= \varepsilon_{m,c(m,t)/2} \\ \varepsilon_{m,1} &\sim \text{Half-Normal}(0, 0.025^2) \\ \varepsilon_{m,s} &\sim \text{Half-Normal}(\varepsilon_{m,s-1}, \sigma_\varepsilon^2), s > 1 \\ \log(\sigma_\varepsilon) &\sim \text{Normal}(-1.2, 0.2^2),\end{aligned}\tag{8.14}$$

where $c(m, t)$ is a function that maps the time indices in the location m to calendar weeks, and the mean of σ_ε takes the value of 0.3 that favours the smoothness. For lack of further data, we assumed the $\eta_{m,t,a}$ to be a constant for $a < 20$, and imposed a vague prior on $\eta^{\text{children}} \sim \text{Uniform}(0.1, 1.0)$. To moderate the interaction between children and teenagers, we introduced a multiplier γ , and used the prior $\log \gamma \sim \text{Normal}(0, 0.35)$ that centres at the null hypothesis of no effect.

Infection-to-death distribution

The discredited infection-to-death distribution $h_{m,t-s,a}$ is described by

$$\begin{aligned}h_{m,1,a} &= \gamma_{m,a} \int_0^{1.5} h(u) du \\ h_{m,s,a} &= \gamma_{m,a} \int_{s-0.5}^{s+0.5} h(u) du, s = 2, 3, \dots,\end{aligned}\tag{8.15}$$

where $\gamma_{m,a}$ is the IFR in location m and age band a and $h(u)$ follows

$$h(u) \sim \text{Gamma}(5.1, 0.86) + \text{Gamma}(17.8, 0.45),\tag{8.16}$$

as in Flaxman et al. (2020). We decomposed $\gamma_{m,a}$ into the age-specific IFR $\gamma_a^{\text{IFR-meta}}$ and the location-specific random effects in terms of three age bands $[20-49]$, $[50-69]$ and $[70+]$,

denoted as $\delta_{m,a}^{\text{IFR}}$. We characterised $\gamma_a^{\text{IFR-meta}}$ by fitting log-normal distributions to the 95% credible intervals of IFRs in Levin et al. (2020). The location-specific departures from the baseline for the three age bands were described by a random effect model,

$$\begin{aligned}\delta_{m,[20-49]}^{\text{IFR}} &\sim \text{Exponential}(\delta_{[20-49]}^{\text{IFR}}) \\ \delta_{m,[50-69]}^{\text{IFR}} &\sim \text{Exponential}(\delta_{[50-69]}^{\text{IFR}}) \\ \delta_{m,[70+]}^{\text{IFR}} &\sim \text{Exponential}(\delta_{[70+]}^{\text{IFR}}) \\ \delta_{[20-49]}^{\text{IFR}}, \delta_{[50-69]}^{\text{IFR}}, \delta_{[70+]}^{\text{IFR}} &\sim \text{Exponential}(0.1).\end{aligned}\tag{8.17}$$

Susceptibility

Recall that we denoted the probability that the contact with an infectious individual leads to the infection of the person in a as ρ_a . We modelled it as a product of a baseline parameter $\rho_{0,m}$ and the relative susceptibility ρ_a^S . In particular, we considered three age bands $[0-14]$, $[14-65]$ and $[65+]$, and set $\rho_{[14-65]}^S$ to be 0. The relative susceptibility of an individual aged a is relative to $[14-65]$ was informed by a contact tracing and testing study in mainland China (Zhang et al., 2020a). Again we fitted the log normal distributions to the 95% credibility intervals in Zhang et al. (2020a). The baseline $\rho_{0,m}$ was obtained by scaling the basic reproduction number R_0 (Davies et al., 2020) by the average contacts of one person in location m at baseline

$$\rho_{0,m} = R_0 / \sum_a p_{m,a} \sum_{a'} C_{m,a,a'}^{\text{weekday}},\tag{8.18}$$

where $p_{m,a}$ is the proportion of the population of location m in age band a , and $C_{m,a,a'}^{\text{weekday}}$ is the baseline contacts that an individual in a have with the individual in a' in the location m during the weekdays.

Dispersion parameter

Here we applied the same vague prior as in Flaxman et al. (2020),

$$\phi \sim \text{Half-Normal}(0, 5).\tag{8.19}$$

8.2.4 Generated quantities under the flow model

I was primarily in charge of deriving and implementing the posterior transmission flows that are implied by the age-specific COVID-19 transmission models of the previous Section.

For ease of presentation, most of the generated quantities were presented in terms of seven age groups,

$$d \in D = \{[0-9], [10-19], [20-34], [35-49], [50-64], [65-79], [80+]\}. \quad (8.20)$$

Here we present several generated quantities that will be used in the following Section 8.3. By aggregation, the number of new infections in the group d in location m at time t is

$$c_{m,t,d} = \sum_{a \in d} c_{m,t,a}. \quad (8.21)$$

The first important quantity we introduce is the time-varying reproduction number. To define it, we first evaluate the estimated number of infectious individuals by weighting the cases by their infectiousness on day t since the infection on day s ,

$$c_{m,t,d}^* = \sum_{s=1}^{t-1} c_{m,s,d} g(t-s). \quad (8.22)$$

The time-varying reproduction number in age group a describes the number of secondary cases generated by an infected individual in a , and is defined by

$$R_{m,t,a} = \sum_{a'} s_{m,t,a'} \rho_{a'} C_{m,t,a,a'} = \frac{c_{m,t,a}}{c_{m,t,a}^*}. \quad (8.23)$$

Then the time-varying reproduction number in age group d is

$$R_{m,t,d} = \frac{c_{m,t,d}}{c_{m,t,d}^*} = \frac{\sum_{a \in d} c_{m,t,a}}{c_{m,t,d}^*} = \frac{\sum_{a \in d} R_{m,t,a} c_{m,t,a}^*}{c_{m,t,d}^*} = \sum_{a \in d} \frac{c_{m,t,a}^*}{c_{m,t,d}^*} R_{m,t,a}, \quad (8.24)$$

so $R_{m,t,d}$ is the weighted sum of $R_{m,t,a}$ for $a \in d$.

Next, we introduced a quantity that is closely related to the theme of the thesis, the contribution of the age group d and the new infections in age group d . First we define the number of transmissions from a to a' ,

$$z_{m,t,a,a'} = s_{m,t,a'} \rho_{a'} C_{m,t,a,a'} \left(\sum_{s=1}^{t-1} c_{m,s,a} g(t-s) \right), \quad (8.25)$$

and then the aggregated number of transmissions from d to d' is

$$z_{m,t,d,d'} = \sum_{a \in d, a' \in d'} z_{m,t,a,a'}. \quad (8.26)$$

Using this quantity, we determine the contribution of the age group d to onward transmissions,

$$\pi_{m,t,d,\cdot} = \frac{\sum_{d'} z_{m,t,d,d'}}{\sum_d \sum_{d'} z_{m,t,d,d'}}, \quad (8.27)$$

and the proportion of new infections in the age group d' ,

$$\pi_{m,t,\cdot,d'} = \frac{\sum_d z_{m,t,d,d'}}{\sum_d \sum_{d'} z_{m,t,d,d'}}. \quad (8.28)$$

8.3 Results

The age-specific transmission model was fitted to the mortality data deposited in <https://github.com/ImperialCollegeLondon/US-covid19-agespecific-mortality-data> in cmdStan 2.23.0 with Stan's adaptive Hamiltonian Monte Carlo (HMC) sampler, and we obtained 8 chains of length 2,000 where the first 1,500 iterations were specified as warm-up. There are 4,000 iterations after burn-in across 8 chains, and 10 parameters with the lowest effective sample sizes were assessed. Their effective sample sizes are from 121 to 500, and $\hat{R}$ are from 1.01 to 1.0687. Using cmdStan instead of rstan, we were able to parallelise computations within one Hamiltonian Monte Carlo chain, and speed up calculations substantially.

Using the generated quantities presented in Section 8.2.4, we were able to describe important aspects of COVID-19's age-specific transmission dynamics across the United States. Particularly, we focused on the effective reproduction number in (8.24), the contribution to onward transmissions by age (8.27), and the age distribution of new daily infections (8.28). First, Figure 8.1 illustrates time trends in the estimated effective reproduction number over time in seven age bands, across 40 states. We found that across most locations, reproduction numbers were consistently highest in individuals aged 20-34 and 35-49. However, the figure also reveals important heterogeneity of age-specific dynamics over time. While the model consistently estimates effective reproduction numbers close to or above one across all locations from adults aged 35-49, disease dynamics are more variable from young adults aged 20-34, with some states (Arizona, Florida, Texas) showing sustained transmission from young adults in May and June, and other states (e.g. Colorado, Illinois, Wisconsin) show-

ing sustained transmission from young adults since August. This suggests that additional interventions to adults aged 20-49 could bring resurgent COVID-19 epidemics under control.

Next, we visualise the estimated new infections by age in Figure 8.2 and the contribution by age during October in Figure 8.3. We found no evidence for substantial shifts in the age composition of new infections. Based on the combined mobility and death data, we found that the reconstructed fluctuations in age-specific reproduction numbers had only a relatively modest impact on the contribution of age groups to onward spread over time, and no evidence that young adults aged 20-34 were the primary source of resurgent COVID-19 in the United States over summer 2020. In contrast, Boehmer et al. (2020) suggests the share of individuals aged 20 to 29 among reported cases increased most, which is inconsistent with our findings. These results underscore that, when testing rates are heterogeneous and not population-representative, it is challenging to determine the age-specific pattern of transmission based only on the reported case data.

Last, Figure 8.3 explores the source of COVID-19 infections in October 2020 and is an analogy to Figure 6.6 in HIV flow analysis. The figure indicates that adults aged 20 to 34 and 35 to 49 continue to be the only age groups that contribute disproportionately to COVID-19 spread relative to their size in the population, which is consistent with the findings in Figure 8.1 on effective reproduction numbers.

8.4 Discussion

We draw a comparison between the COVID study that are based on standard epidemiological surveillance and our HIV study using the viral sequence data. Phyloscanner reconstructs the transmission dynamics with limited assumptions and no external data, while the COVID study depends heavily on the reported deaths, mobility and contact data. First, Rossen et al. (2020) reported the excess data is 38% higher than the reported COVID attributed deaths, and suggested under-reporting of COVID-19 deaths. Kobilov et al. (2021) also indicates the stringency of the policies and the healthcare capabilities bias the death reporting over the locations. Though less biased reported deaths in contrast with the reported cases, the research outcome would be heavily sensitive to the bias of the officially reported deaths. Second, the COVID analysis is model-based and depends on several assumptions. For instance, (8.2) depends on the assumptions of infection-to-death distribution and infection fatality ratios, and (8.3) depends on the predicted baseline contact matrices in the US, the influence of mobility on contacts, probability of infection upon contacts and generation time distribution. Though most of the choices are made from the literature, we are in danger of propagating

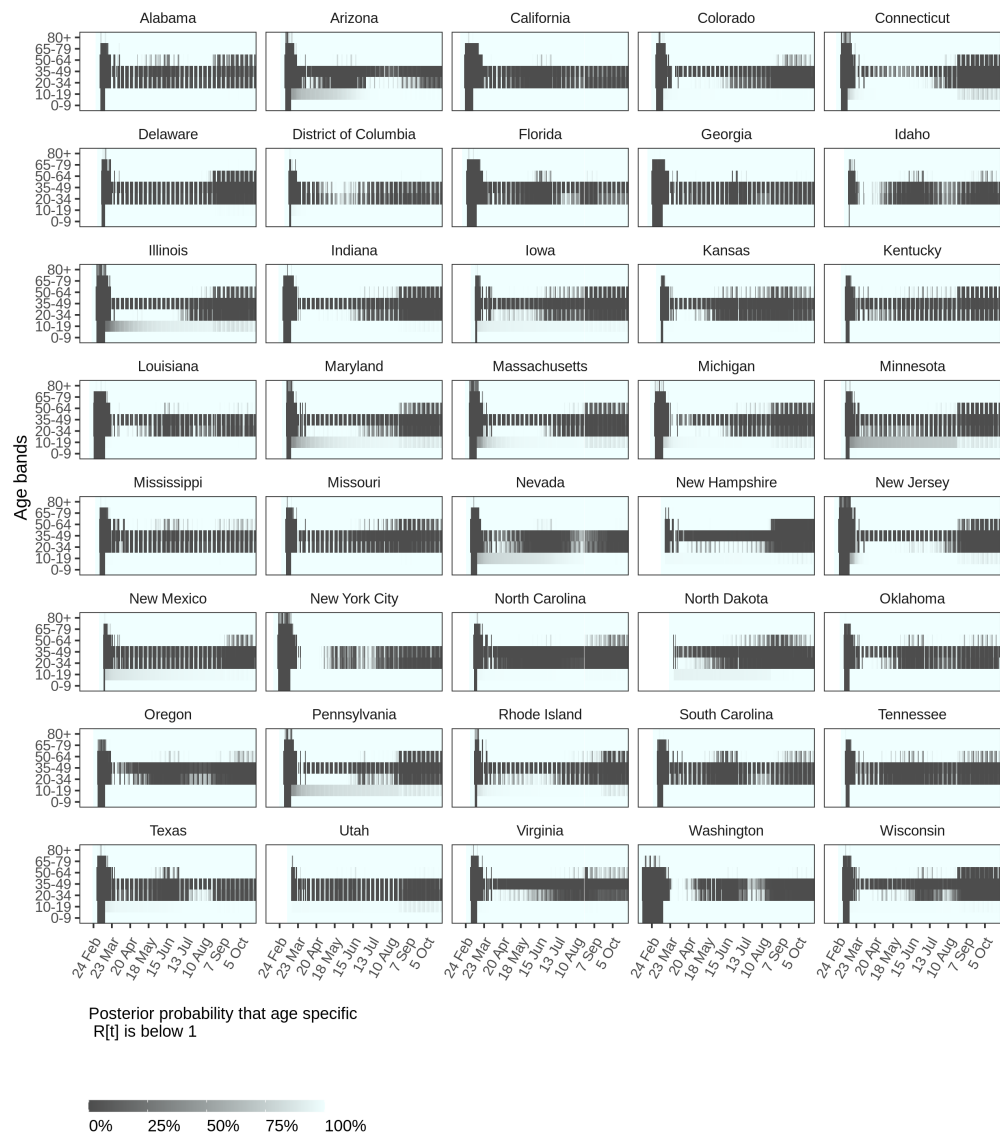


Figure 8.1 **Time evolution of age-specific reproduction numbers across states.** Each panel shows the estimated posterior probability (in colour) that the daily effective reproduction number in each age group (in the y-axis) is < 1 over time (in the x-axis).

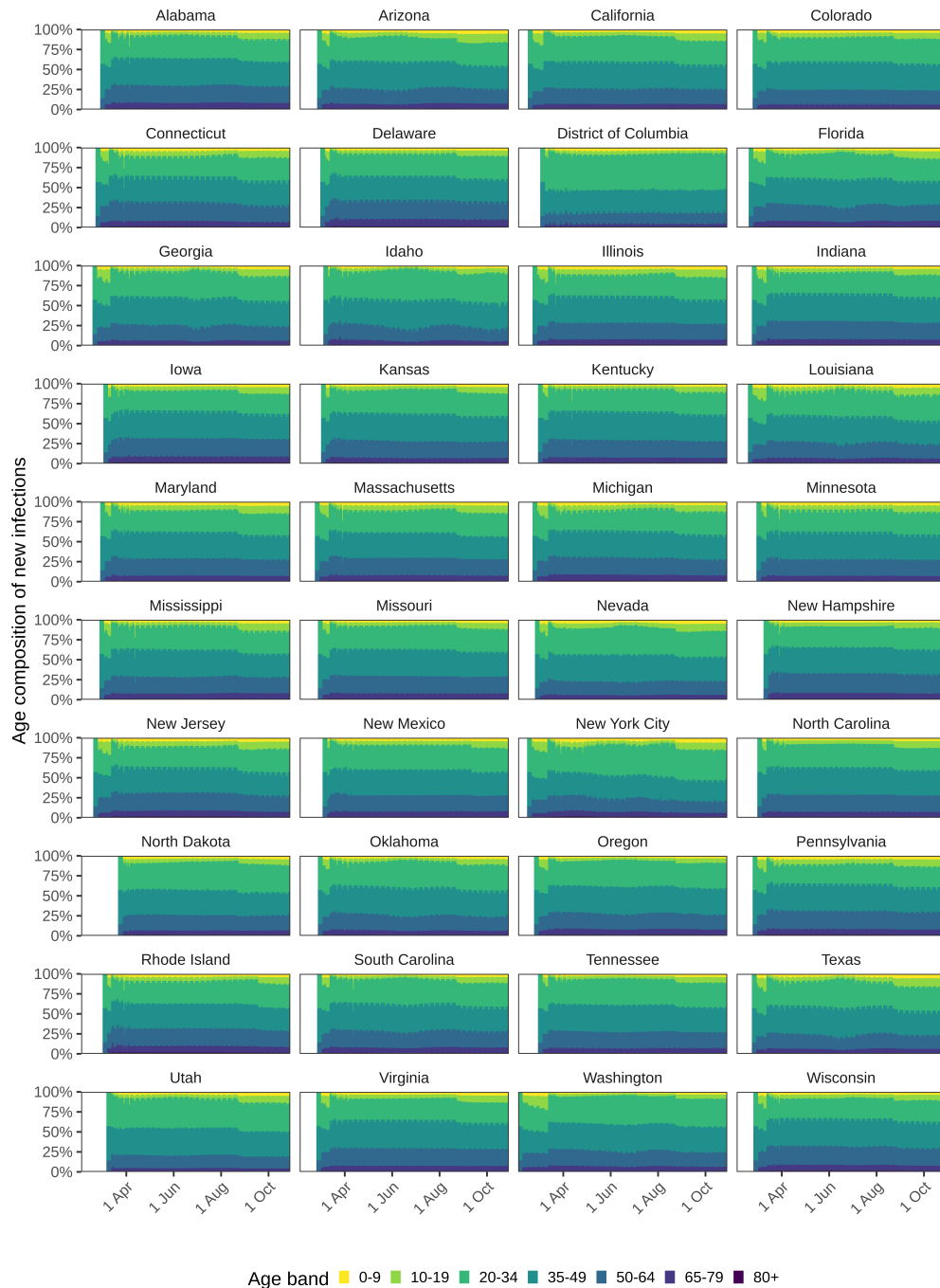


Figure 8.2 New cases by age. Each panel shows the proportion of new daily COVID-19 infections (in the y-axis) for seven age bands (in colour) along the time (in the x-axis).

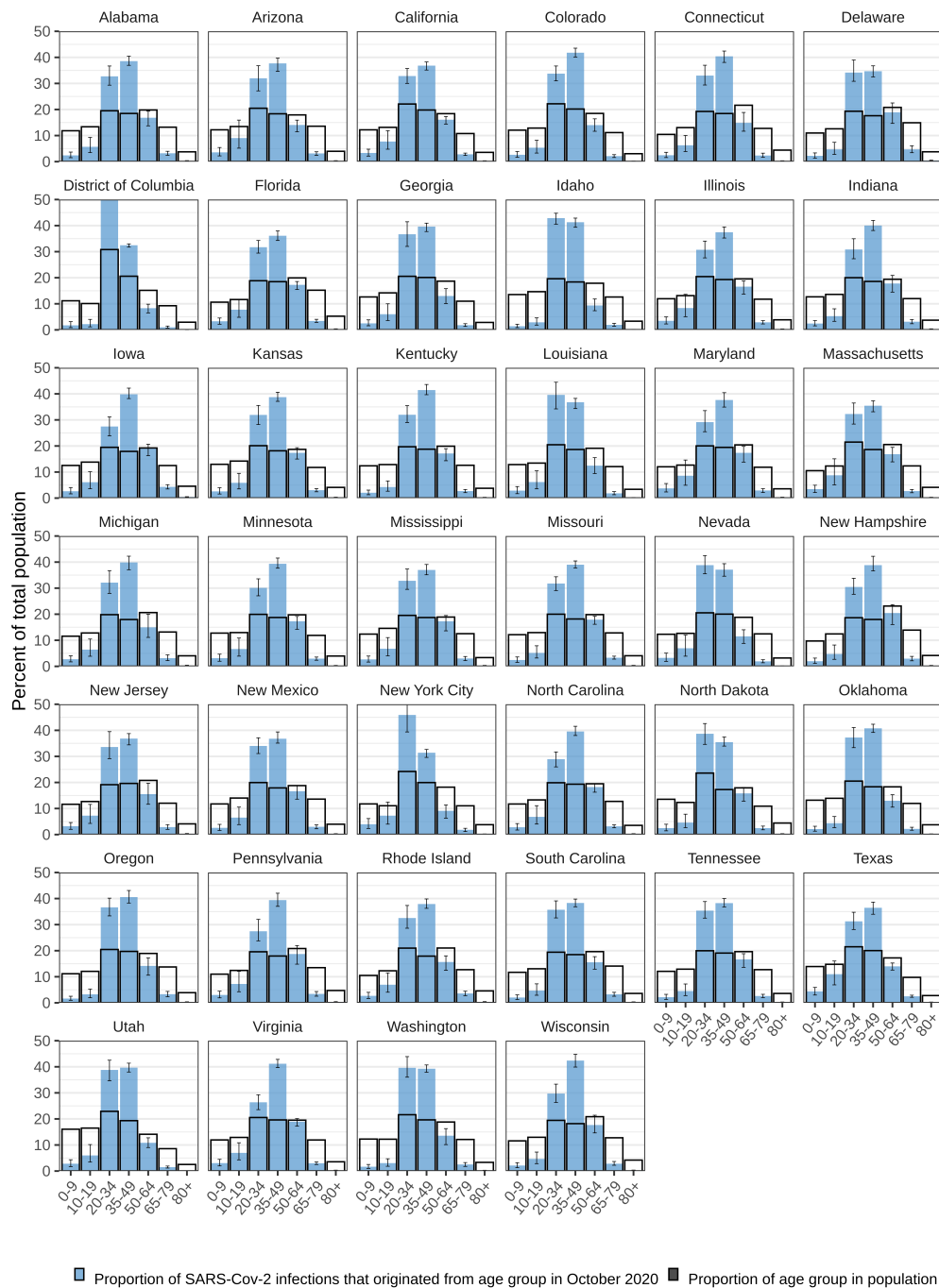


Figure 8.3 Estimated accumulated contribution of age groups to COVID-19 cases for October, versus the proportion of the population in the same age group. Posterior median estimates of the contributions from each age group (blue fill bars with 95% credible intervals) are compared against the population age composition of each state (black contour bars).

multiple sources of bias. Also, neglecting further complexity in population-level disease spread may generate unreliable results as evidence accumulates. Even though the analyses via viral sequences are less subject to reporting bias and assumptions, the comprehensive study, such as the Rakai Community Cohort Study (RCCS), requires loads of effort to conduct household census and sample collection, and is thus expensive and rare in practice, especially on the national scale. Another advantage of the model-based COVID study is the ability to simulate intervention scenarios and project into the future, while the aim of the HIV study is to reconstruct population-level transmissions in a recent time frame. For example, sources and destinations of flows in (8.27) and (8.28) are equivalent to (5.2a) and (5.2b), but (8.27) and (8.28) can be evaluated over time and (5.2a) and (5.2b) are defined between August 2011 and January 2015.

8.5 Summary

In this Chapter, we integrated the mobility data and mortality data to uncover the transmission dynamics between the age groups during the resurgent COVID-19 epidemics in the United States in 2020. To do this, we built an age-specific renewal equation and embedded it into a Bayesian hierarchical model. This work suggested a broad age group aged 20-49, especially 35-49, drive the resurgent COVID-19 in the United States. The share of age groups in new cases remained almost constant over time. We finally made a comparison with the main theme of the thesis.

Chapter 9

Conclusion

This Chapter summarises our contributions and mentions the future work. The primary objective of the thesis is to develop statistical methods for reconstructing transmission dynamics at the population level using human immunodeficiency virus (HIV) deep sequence data. The use of these new types of data opens up a new avenue of phylodynamics research, which requires new methods for its analysis. Specifically, we update the bio-informative pipeline of deep sequence analysis and scale it up to a larger dataset. The key feature of the pre-processing tailored for deep sequences is the ability to reconstruct source-recipient pairs by analysing the topological relationship in the phylogenies with fewer assumptions and external data than the approaches tailored for Sanger sequences. It prompts us to apply Bayesian methods for interpreting the pre-processing outputs which are computationally easier to fit than phylogenies. We are then able to transfer the Gaussian processes on directional data such as wind directions to phylodynamics and explore the structures of transmission directions without strong assumptions. It also allows for addressing the challenging problems in the current phylodynamics methods, like spatial sampling heterogeneity, migration and high resolution flows. To do this, 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, and easier model comparison. We template the model with and without Hilbert space approximations for scalable inference of high-resolution flows in generic Stan model files, make it possible for selecting the model that best describe the data and hope the canonical structure of the semi-parametric Poisson flow model will be helpful in other movement, origin-destination or flow applications as well (Faye et al., 2015;

Hazelton, 2001; Lindström et al., 2013; Raymer et al., 2013; Sun et al., 2021; Tebaldi and West, 1998; van de Kastele et al., 2017). In the final Chapter, I describe my contributions to the inference of high-resolution flows using non-viral sequence data and compare with the deep sequence based flows.

Future research should study the reliability of deep sequence analysis more carefully. Though Ratmann et al. (2019) and we assess the accuracy of the inference of transmission directions on limited clinical data, a simulation study on deep sequence data would be beneficial to examine and guarantee the performance of source attribution through *phyloscanner*. Second, the scripts to automate the bioinformatics pipeline in Chapter 3 make it easier to generalise the transmission network analysis to other cohorts in Phylogenetics And Networks for Generalised Epidemics in Africa consortium (PANGAEA), and thereby facilitate the comparison between cohorts. Third, we analyse transmission networks through the sequences collected during January 2011 and May 2018. Repeating the analysis using the semi-parametric model, especially the model in Chapter 6 provides useful epidemiological insights on shifts in age dynamics between study periods. Another stimulating epidemiological question is whether age dynamics differ between inland and fishing communities. Our preliminary findings suggest it is more likely that women aged 15-24 in inland communities transmit to men aged 35+ while those in fishing sites pass more infections to men aged 25-34. Future work is needed to consolidate the findings. Fourth, the non-parametric model in Chapter 4 is a start to transfer the ideas in meteorology to phylodynamics and maximise the utility of the deep sequence data, where as less as possible assumptions are introduced. We may work on the model further to be more realistic such as considering non-stationary and non-separable kernels, and incorporate more individual covariates. Fifth, the model frameworks in Chapter 5 and 6 can be extended for soft thresholds, rather than hard thresholds, to propagate the evidence on transmission directions into flow inference. It can also be generalised to consider the case when the sampling process is more complicated and not missing at random. For example, individuals on the treatment and short-term mobile hard-to-reach individuals may confound our results. It is possible to further extend the model framework by selecting variables and identifying the drivers of HIV transmissions in Rakai, which used to be challenging through standard epidemiological data. Sixth, it is worthwhile to attempt other popular Sanger sequence based approaches such as structured coalescent models, compare with the emerging deep sequence based approaches, and well understand both their benefits and drawbacks. Moreover, an efficient method for comparing Gaussian process models can be developed, considering the huge computational resources required in

our approach. Finally, the regularising trick in Chapter 6 might be beneficial for the accurate and robust inference of COVID-19 flow by age.

In conclusion, this thesis mainly introduces a flexible class of Bayesian semi-parametric Poisson models for estimating cross-sectional infectious disease transmission flows, that is applicable to emerging pathogen deep-sequence data. The central feature of the model is that high-dimensional transmission flows can be efficiently estimated from sparsely and heterogeneously sampled pair data, which is possible because external information on population sampling can be directly incorporated into the model, and because efficient Gaussian process approximations facilitate smoothing in flow space. In general, we found deep viral sequence provides unprecedented opportunities, rich information, and a less biased way to discover transmission patterns than epidemiological surveillance data, and provide an alternative way to tackle the challenging problems in standard phylodynamics.

Bibliography

- L. Abeler-Dörner, M. K. Grabowski, A. Rambaut, D. Pillay, C. Fraser, et al. PANGEA-HIV 2: Phylogenetics and networks for generalised epidemics in Africa. *Current Opinion in HIV and AIDS*, 14(3):173–180, 2019.
- F. Ailloud, X. Didelot, S. Woltemate, G. Pfaffinger, J. Overmann, R. C. Bader, C. Schulz, P. Malfertheiner, and S. Suerbaum. Within-host evolution of helicobacter pylori shaped by niche-specific adaptation, intragastric migrations and selective sweeps. *Nature communications*, 10(1):1–13, 2019.
- K. E. Ainslie, C. E. Walters, H. Fu, S. Bhatia, H. Wang, X. Xi, M. Baguelin, S. Bhatt, A. Boonyasiri, O. Boyd, et al. Evidence of initial success for China exiting COVID-19 social distancing policy after achieving containment. *Wellcome Open Research*, 5, 2020.
- Ç. Ak, Ö. Ergönül, İ. Şencan, M. A. Torunoğlu, and M. Gönen. Spatiotemporal prediction of infectious diseases using structured Gaussian processes with application to Crimean–Congo hemorrhagic fever. *PLoS neglected tropical diseases*, 12(8):e0006737, 2018.
- A. Akullian, A. Bershteyn, D. Klein, A. Vandormael, T. Bärnighausen, and F. Tanser. Sexual partnership age pairings and risk of HIV acquisition in rural South Africa. *AIDS (London, England)*, 31(12):1755, 2017.
- S. Alizon and C. Fraser. Within-host and between-host evolutionary rates across the HIV-1 genome. *Retrovirology*, 10(1):1–10, 2013.
- R. M. Anderson and R. M. May. *Infectious diseases of humans: dynamics and control*. Oxford University Press, 1992.
- S. O. Aral, E. Torrone, and K. Bernstein. Geographical targeting to improve progression through the sexually transmitted infection/HIV treatment continua in different populations. *Current Opinion in HIV and AIDS*, 10(6):477, 2015.
- R. Baker and D. Jackson. Using journal impact factors to correct for the publication bias of medical studies. *Biometrics*, 62(3):785–792, 2006.
- J. E. Balkus, G. Nair, E. T. Montgomery, A. Mishra, T. Palanee-Phillips, G. Ramjee, R. Panchia, P. Selepe, B. A. Richardson, Z. M. Chirenje, et al. Age-disparate partnerships and risk of HIV-1 acquisition among South African women participating in the VOICE trial. *Journal of Acquired Immune Deficiency Syndromes*, 70(2):212, 2015.
- S. Banerjee, B. P. Carlin, and A. E. Gelfand. *Hierarchical modeling and analysis for spatial data*. Chapman and Hall/CRC, 2003.

- F. Barré-Sinoussi, S. S. Abdool Karim, J. Albert, L.-G. Bekker, C. Beyrer, P. Cahn, A. Calmy, B. Grinsztejn, A. Grulich, A. Kamarulzaman, et al. Expert consensus statement on the science of HIV in the context of criminal law. *Journal of the International AIDS Society*, 21(7):e25161, 2018.
- N. Bbosa, D. Ssemwanga, R. N. Nsubuga, J. F. Salazar-Gonzalez, M. G. Salazar, M. Nanyonjo, M. Kuteesa, J. Seeley, N. Kiwanuka, B. S. Bagaya, et al. Phylogeography of HIV-1 suggests that Ugandan fishing communities are a sink for, not a source of, virus from general populations. *Scientific reports*, 9(1):1–8, 2019.
- N. Bbosa, D. Ssemwanga, A. Ssekagiri, X. Xi, Y. Mayanja, U. Bahemuka, J. Seeley, D. Pillay, L. Abeler-Dörner, T. Golubchik, et al. Phylogenetic and demographic characterization of directed HIV-1 transmission using deep sequences from high-risk and general population cohorts/groups in Uganda. *Viruses*, 12(3):331, 2020.
- J. O. Berger, J. M. Bernardo, and D. Sun. Overall objective priors. *Bayesian Analysis*, 10(1): 189–221, 2015.
- M. Betancourt. Diagnosing biased inference with divergences. https://betanalpha.github.io/assets/case_studies/divergences_and_bias.html, 2017a. Accessed: 2022-03-11.
- M. Betancourt. Robust Gaussian processes in Stan. https://betanalpha.github.io/assets/case_studies/gp_part3/part3.html, 2017b. Accessed: 2021-11-23.
- M. Betancourt. Identifying Bayesian mixture models. https://mc-stan.org/users/documentation/case-studies/identifying_mixture_models.html, 2017c. Accessed: 2022-03-11.
- M. Betancourt and M. Girolami. Hamiltonian Monte Carlo for hierarchical models. *Current trends in Bayesian methodology with applications*, 79(30):2–4, 2015.
- S. Bhatia, N. Imai, G. Cuomo-Dannenburg, M. Baguelin, A. Boonyasiri, A. Cori, Z. Cucunubá, I. Dorigatti, R. FitzJohn, H. Fu, et al. Estimating the number of undetected COVID-19 cases among travellers from mainland China. *Wellcome Open Research*, 5, 2020.
- S. Bhatt, E. Cameron, S. R. Flaxman, D. J. Weiss, D. L. Smith, and P. W. Gething. Improved prediction accuracy for disease risk mapping using Gaussian process stacked generalization. *Journal of The Royal Society Interface*, 14(134):20170520, 2017.
- V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10): P10008, 2008.
- T. K. Boehmer, J. DeVies, E. Caruso, K. L. van Santen, S. Tang, C. L. Black, K. P. Hartnett, A. Kite-Powell, S. Dietz, M. Lozier, et al. Changing age distribution of the COVID-19 pandemic—United States, May–August 2020. *Morbidity and Mortality Weekly Report*, 69(39):1404, 2020.

- D. Bonsall, T. Golubchik, M. de Cesare, M. Limbada, B. Kosloff, G. MacIntyre-Cockett, M. Hall, C. Wymant, M. A. Ansari, L. Abeler-Dörner, et al. A comprehensive genomics solution for HIV surveillance and clinical monitoring in a global health setting. *bioRxiv*, page 397083, 2018.
- R. Bunnell, A. Opio, J. Musinguzi, W. Kirungi, P. Ekwaru, V. Mishra, W. Hladik, J. Kafuko, E. Madraa, and J. Mermin. HIV transmission risk behavior among HIV-infected adults in Uganda: results of a nationally representative survey. *AIDS*, 22(5):617–624, 2008.
- N. B. Carnegie, R. Wang, V. Novitsky, and V. De Gruttola. Linkage of viral sequences among HIV-infected village residents in Botswana: estimation of linkage rates in the presence of missing data. *PLoS Computational Biology*, 10(1):e1003430, 2014.
- B. Carpenter, A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell. Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1), 2017.
- CDC. Basic Statistics., 2021.
- S. K. Chan, L. R. Thornton, K. J. Chronister, J. Meyer, M. Wolverton, C. K. Johnson, R. R. Arafat, M. P. Joyce, W. M. Switzer, W. Heneine, et al. Likely female-to-female sexual transmission of HIV—Texas, 2012. *CDC Morbidity and Mortality Weekly Report*, 63(10): 209, 2014.
- L. W. Chang, M. K. Grabowski, R. Ssekubugu, F. Nalugoda, G. Kigozi, B. Nantume, J. Lessler, S. M. Moore, T. C. Quinn, S. J. Reynolds, et al. Heterogeneity of the HIV epidemic: an observational epidemiologic study of agrarian, trading, and fishing communities in Rakai, Uganda. *The Lancet HIV*, 3(8):e388, 2016.
- S. Chib. Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, 90(432):1313–1321, 1995.
- S. Chib and I. Jeliazkov. Marginal likelihood from the Metropolis–Hastings output. *Journal of the American statistical association*, 96(453):270–281, 2001.
- A. Cornish-Bowden. Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984. *Nucleic acids research*, 13(9):3021, 1985.
- N. G. Davies, A. J. Kucharski, R. M. Eggo, A. Gimma, W. J. Edmunds, T. Jombart, K. O’Reilly, A. Endo, J. Hellewell, E. S. Nightingale, et al. Effects of non-pharmaceutical interventions on COVID-19 cases, deaths, and demand for hospital services in the UK: a modelling study. *The Lancet Public Health*, 5(7):e375–e385, 2020.
- N. De Maio, C.-H. Wu, K. M. O’Reilly, and D. Wilson. New routes to phylogeography: a Bayesian structured coalescent approximation. *PLoS genetics*, 11(8):e1005421, 2015.
- T. De Oliveira, A. B. Kharsany, T. Gräf, C. Cawood, D. Khanyile, A. Grobler, A. Puren, S. Madurai, C. Baxter, Q. A. Karim, et al. Transmission networks and risk of HIV infection in KwaZulu-Natal, South Africa: a community-wide phylogenetic study. *The Lancet HIV*, 4(1):e41–e50, 2017.

- R. C. Dellar, S. Dlamini, and Q. A. Karim. Adolescent girls and young women: key populations for HIV epidemic control. *Journal of the International AIDS Society*, 18: 19408, 2015.
- S. Dellicour, G. Baele, G. Dudas, N. R. Faria, O. G. Pybus, M. A. Suchard, A. Rambaut, and P. Lemey. Phylodynamic assessment of intervention strategies for the West African Ebola virus outbreak. *Nature Communications*, 9(1):1–9, 2018.
- X. Didelot, C. Fraser, J. Gardy, and C. Colijn. Genomic infectious disease epidemiology in partially sampled and ongoing outbreaks. *Molecular Biology and Evolution*, 34(4): 997–1007, 2017.
- P. J. Diggle, J. A. Tawn, and R. A. Moyeed. Model-based geostatistics. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 47(3):299–350, 1998.
- A. J. Drummond and A. Rambaut. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC evolutionary biology*, 7(1):1–8, 2007.
- S. Duval and R. Tweedie. Trim and fill: a simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, 56(2):455–463, 2000.
- L. Dwyer-Lindgren, M. A. Cork, A. Sligar, K. M. Steuben, K. F. Wilson, N. R. Provost, B. K. Mayala, J. D. VanderHeide, M. L. Collison, J. B. Hall, et al. Mapping HIV prevalence in sub-Saharan Africa between 2000 and 2017. *Nature*, 570(7760):189, 2019.
- J. Edmonds. Paths, trees, and flowers. *Canadian Journal of mathematics*, 17:449–467, 1965.
- S. H. Eshleman, S. E. Hudelson, A. D. Redd, L. Wang, R. Debes, Y. Q. Chen, C. A. Martens, S. M. Ricklefs, E. J. Selig, S. F. Porcella, et al. Analysis of genetic linkage of HIV from couples enrolled in the HIV Prevention Trials Network 052 trial. *The Journal of infectious diseases*, 204(12):1918–1926, 2011.
- N. R. Faria, A. Rambaut, M. A. Suchard, G. Baele, T. Bedford, M. J. Ward, A. J. Tatem, J. D. Sousa, N. Arinaminpathy, J. Pépin, et al. The early spread and epidemic ignition of HIV-1 in human populations. *Science*, 346(6205):56–61, 2014.
- O. Faye, P.-Y. Boëlle, E. Heleze, O. Faye, C. Loucoubar, N. Magassouba, B. Soropogui, S. Keita, T. Gakou, L. Koivogui, et al. Chains of transmission and control of Ebola virus disease in Conakry, Guinea, in 2014: an observational study. *The Lancet Infectious Diseases*, 15(3):320–326, 2015.
- S. Flaxman, S. Mishra, A. Gandy, H. J. T. Unwin, T. A. Mellan, H. Coupland, C. Whittaker, H. Zhu, T. Berah, J. W. Eaton, et al. Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. *Nature*, 584(7820):257–261, 2020.
- B. T. Foley, T. K. Leitner, C. Apetrei, B. Hahn, I. Mizrachi, J. Mullins, A. Rambaut, S. Wolinsky, and B. T. M. Korber. HIV sequence compendium 2015. 10 2015. doi: 10.2172/1222684. URL <https://www.osti.gov/biblio/1222684>.
- S. D. Frost, O. G. Pybus, J. R. Gog, C. Viboud, S. Bonhoeffer, and T. Bedford. Eight challenges in phylodynamic inference. *Epidemics*, 10:88–92, 2015.

- S. Frühwirth-Schnatter. Estimating marginal likelihoods for mixture and Markov switching models using bridge sampling techniques. *The Econometrics Journal*, 7(1):143–167, 2004.
- H. Fu, H. Wang, X. Xi, A. Boonyasiri, Y. Wang, W. Hinsley, K. J. Fraser, R. McCabe, D. O. Mesa, J. Skarp, et al. Database of epidemic trends and control measures during the first wave of COVID-19 in mainland China. *International Journal of Infectious Diseases*, 102: 463–471, 2021.
- A. Gall, B. Ferns, C. Morris, S. Watson, M. Cotten, M. Robinson, N. Berry, D. Pillay, and P. Kellam. Universal amplification, next-generation sequencing, and assembly of HIV-1 genomes. *Journal of Clinical Microbiology*, 50(12):3838–3844, 2012.
- L. Geidelberg, O. Boyd, D. Jorgensen, I. Siveroni, F. F. Nascimento, R. Johnson, M. Ragonnet-Cronin, H. Fu, H. Wang, X. Xi, et al. Genomic epidemiology of a densely sampled COVID-19 outbreak in China. *Virus evolution*, 7(1):veaa102, 2021.
- A. E. Gelfand and D. K. Dey. Bayesian model choice: asymptotics and exact calculations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 56(3):501–514, 1994.
- A. E. Gelfand and E. M. Schliep. Spatial statistics and Gaussian processes: A beautiful marriage. *Spatial Statistics*, 18:86–104, 2016.
- A. E. Gelfand, H.-J. Kim, C. Sirmans, and S. Banerjee. Spatial modeling with spatially varying coefficient processes. *Journal of the American Statistical Association*, 98(462): 387–396, 2003.
- A. Gelman and X.-L. Meng. Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical science*, pages 163–185, 1998.
- J. Geweke. Bayesian inference in econometric models using Monte Carlo integration. *Econometrica: Journal of the Econometric Society*, pages 1317–1339, 1989.
- G. H. Givens, D. Smith, and R. Tweedie. Publication bias in meta-analysis: a Bayesian data-augmentation approach to account for issues exemplified in the passive smoking debate. *Statistical Science*, pages 221–240, 1997.
- T. Golubchik, O. Ratmann, C. Wymant, M. Hall, D. Bonsall, M. K. Grabowski, O. Laeyendecker, and C. Fraser. Quantifying within-host viral diversification using deep sequencing data: recent vs chronic HIV infection, 2017.
- M. K. Grabowski, J. Lessler, A. D. Redd, J. Kagaayi, O. Laeyendecker, A. Ndyanaabo, M. I. Nelson, D. A. Cummings, J. B. Bwanika, A. C. Mueller, et al. 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(3), 2014.
- M. K. Grabowski, D. M. Serwadda, R. H. Gray, G. Nakigozi, G. Kigozi, J. Kagaayi, R. Ssekubugu, F. Nalugoda, J. Lessler, T. Lutalo, et al. HIV prevention efforts and incidence of HIV in Uganda. *New England Journal of Medicine*, 377(22):2154–2166, 2017.

- M. K. Grabowski, S. J. Reynolds, J. Kagaayi, R. H. Gray, W. Clarke, L. Chang, G. Nakigozi, O. Laeyendecker, A. D. Redd, V. Goud-Billoux, et al. The validity of self-reported antiretroviral use in persons living with HIV: a population-based study. *AIDS*, 32(3):363, 2018.
- M. K. Grabowski, J. Lessler, J. Bazaale, D. Nabukalu, J. Nankinga, B. Nantume, J. Ssekasanvu, S. J. Reynolds, R. Ssekubugu, F. Nalugoda, et al. Migration, hotspots, and dispersal of HIV infection in Rakai, Uganda. *Nature communications*, 11(1):1–12, 2020.
- N. C. Grassly, M. Pons-Salort, E. P. Parker, P. J. White, N. M. Ferguson, K. Ainslie, M. Baguelin, S. Bhatt, A. Boonyasiri, N. Brazeau, et al. Comparison of molecular testing strategies for COVID-19 control: a mathematical modelling study. *The Lancet Infectious Diseases*, 20(12):1381–1389, 2020.
- Q. F. Gronau, A. Sarafoglou, D. Matzke, A. Ly, U. Boehm, M. Marsman, D. S. Leslie, J. J. Forster, E.-J. Wagenmakers, and H. Steingroever. A tutorial on bridge sampling. *Journal of mathematical psychology*, 81:80–97, 2017a.
- Q. F. Gronau, H. Singmann, and E.-J. Wagenmakers. bridgesampling: An R package for estimating normalizing constants. *arXiv preprint arXiv:1710.08162*, 2017b.
- Q. F. Gronau, A. Heathcote, and D. Matzke. Computing Bayes factors for evidence-accumulation models using Warp-III bridge sampling. *Behavior research methods*, pages 1–20, 2019a.
- Q. F. Gronau, E.-J. Wagenmakers, D. W. Heck, and D. Matzke. A simple method for comparing complex models: Bayesian model comparison for hierarchical multinomial processing tree models using Warp-III bridge sampling. *psychometrika*, 84(1):261–284, 2019b.
- M. Guan and J. Vandekerckhove. A Bayesian approach to mitigation of publication bias. *Psychonomic Bulletin & Review*, 23(1):74–86, 2016.
- M. Hall, M. Woolhouse, and A. Rambaut. Epidemic reconstruction in a phylogenetics framework: transmission trees as partitions of the node set. *PLoS computational biology*, 11(12):e1004613, 2015.
- M. Hall, T. Golubchik, D. Bonsall, L. Abeler-Dorner, M. Limbada, B. Kosloff, A. Schaap, M. de Cesare, G. Mackintyre-Cockett, W. Probert, et al. Demographic characteristics of sources of HIV-1 transmission in Zambia. *medRxiv*, 2021.
- M. D. Hall, M. E. Woolhouse, and A. Rambaut. The effects of sampling strategy on the quality of reconstruction of viral population dynamics using Bayesian skyline family coalescent methods: A simulation study. *Virus evolution*, 2(1), 2016.
- J. Hammersley. *Monte Carlo methods*. Springer Science & Business Media, 2013.
- R. W. Hamming. Error detecting and error correcting codes. *The Bell system technical journal*, 29(2):147–160, 1950.

- K. Hanke, N. R. Faria, D. Kühnert, K. P. Yousef, A. Hauser, K. Meixenberger, A. Hofmann, V. Bremer, B. Bartmeyer, O. Pybus, et al. Reconstruction of the genetic history and the current spread of HIV-1 subtype A in Germany. *Journal of virology*, 93(12):e02238–18, 2019.
- G. Harling, M.-L. Newell, F. Tanser, I. Kawachi, S. Subramanian, and T. Barnighausen. Do age-disparate relationships drive HIV incidence in young women? Evidence from a population cohort in rural KwaZulu-natal, South Africa. *Journal of acquired immune deficiency syndromes (1999)*, 66(4):443, 2014.
- R. J. Hayes, D. Donnell, S. Floyd, N. Mandla, J. Bwalya, K. Sabapathy, B. Yang, M. Phiri, A. Schaap, S. H. Eshleman, E. Piwowar-Manning, B. Kosloff, A. James, T. Skalland, E. Wilson, L. Emel, D. Macleod, R. Dunbar, M. Simwinga, N. Makola, V. Bond, G. Hod-dinott, A. Moore, S. Griffith, N. Deshmane Sista, S. H. Vermund, W. El-Sadr, D. N. Burns, J. R. Hargreaves, K. Hauck, C. Fraser, K. Shanaube, P. Bock, N. Beyers, H. Ayles, and S. Fidler. Effect of Universal Testing and Treatment on HIV Incidence — HPTN 071 (PopART). *New England Journal of Medicine*, 381(3):207–218, 2019. doi: 10.1056/NEJMoa1814556. PMID: 31314965.
- M. L. Hazelton. Inference for origin–destination matrices: estimation, prediction and reconstruction. *Transportation Research Part B: Methodological*, 35(7):667–676, 2001.
- A. B. Hogan, B. L. Jewell, E. Sherrard-Smith, J. F. Vesga, O. J. Watson, C. Whittaker, A. Hamlet, J. A. Smith, P. Winskill, R. Verity, et al. Potential impact of the COVID-19 pandemic on HIV, tuberculosis, and malaria in low-income and middle-income countries: a modelling study. *The Lancet Global Health*, 8(9):e1132–e1141, 2020.
- C. F. Houlihan, D. Frampton, R. B. Ferns, J. Raffle, P. Grant, M. Reidy, L. Hail, K. Thomson, F. Mattes, Z. Kozlakidis, et al. Use of whole-genome sequencing in the investigation of a nosocomial influenza virus outbreak. *The Journal of Infectious Diseases*, 218(9): 1485–1489, 2018.
- S. Hué, D. Pillay, J. P. Clewley, and O. G. Pybus. Genetic analysis reveals the complex structure of HIV-1 transmission within defined risk groups. *Proceedings of the National Academy of Sciences*, 102(12):4425–4429, 2005.
- J. P. Ioannidis. Why most published research findings are false. *PLoS medicine*, 2(8):e124, 2005.
- X. Jin, J. Cheng, and J. Lou. Infer HIV transmission dynamics from gene sequences among young men who have sex with men in China. *Infectious Disease Modelling*, 6:832–838, 2021.
- Z.-C. Jin, X.-H. Zhou, and J. He. Statistical methods for dealing with publication bias in meta-analysis. *Statistics in Medicine*, 34(2):343–360, 2015.
- T. Jombart, A. Cori, X. Didelot, S. Cauchemez, C. Fraser, and N. Ferguson. Bayesian reconstruction of disease outbreaks by combining epidemiologic and genomic data. *PLoS Comput Biol*, 10(1):e1003457, 2014.

- G. Jona-Lasinio, A. Gelfand, and M. Jona-Lasinio. Spatial analysis of wave direction data using wrapped Gaussian processes. *The Annals of Applied Statistics*, 6(4):1478–1498, 2012.
- G. Jona Lasinio, M. Santoro, and G. Mastrantonio. Circspacetime: an R package for spatial and spatio-temporal modelling of circular data. *Journal of Statistical Computation and Simulation*, 90(7):1315–1345, 2020.
- A. Jordan, F. Krüger, and S. Lerch. Evaluating probabilistic forecasts with scoringRules. *arXiv preprint arXiv:1709.04743*, 2017.
- J. Kagaayi, L. W. Chang, V. Ssempijja, M. K. Grabowski, R. Ssekubugu, G. Nakigozi, G. Kigozi, D. M. Serwadda, R. H. Gray, F. Nalugoda, et al. Impact of combination HIV interventions on HIV incidence in hyperendemic fishing communities in Uganda: a prospective cohort study. *The Lancet HIV*, 6(10):e680–e687, 2019.
- M. D. Karcher, J. A. Palacios, T. Bedford, M. A. Suchard, and V. N. Minin. Quantifying and mitigating the effect of preferential sampling on phylodynamic inference. *PLoS Computational Biology*, 12(3):e1004789, 2016.
- R. E. Kass and A. E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795, 1995.
- K. Katoh and D. M. Standley. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution*, 30(4):772–780, 2013.
- E. S. Knock, L. K. Whittles, J. A. Lees, P. N. Perez-Guzman, R. Verity, R. G. FitzJohn, K. A. Gaythorpe, N. Imai, W. Hinsley, L. C. Okell, et al. Key epidemiological drivers and impact of interventions in the 2020 SARS-CoV-2 epidemic in england. *Science Translational Medicine*, 13(602):eabg4262, 2021.
- B. Kobilov, E. Rouen, and G. Serafeim. Predictable country-level bias in the reporting of COVID-19 deaths. *Journal of Government and Economics*, 2:100012, 2021.
- D. Kühnert, T. Stadler, T. G. Vaughan, and A. J. Drummond. Phylodynamics with migration: a computational framework to quantify population structure from genomic data. *Molecular biology and evolution*, 33(8):2102–2116, 2016.
- D. Kühnert, R. Kouyos, G. Shirreff, J. Pečerska, A. U. Scherrer, J. Böni, S. Yerly, T. Klimkait, V. Aubert, H. F. Günthard, et al. Quantifying the fitness cost of HIV-1 drug resistance mutations through phylodynamics. *PLoS pathogens*, 14(2):e1006895, 2018.
- D. T. Larose and D. K. Dey. Modeling publication bias using weighted distributions in a Bayesian framework. *Computational statistics & Data Analysis*, 26(3):279–302, 1998.
- A. Larsson. AliView: a fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics*, 30(22):3276–3278, 2014.
- N. Lartillot and H. Philippe. Computing Bayes factors using thermodynamic integration. *Systematic biology*, 55(2):195–207, 2006.

- S. Le Vu, O. Ratmann, V. Delpech, A. E. Brown, O. N. Gill, A. Tostevin, D. Dunn, C. Fraser, E. M. Volz, and U. H. D. R. Database. HIV-1 transmission patterns in Men Who Have Sex with Men: Insights from genetic source attribution analysis. *AIDS Research and Human Retroviruses*, 35(9):805–813, 2019.
- J. Y. Lee, P. J. Green, and L. M. Ryan. On the "Poisson trick" and its extensions for fitting multinomial regression models. *arXiv preprint arXiv:1707.08538*, 2017.
- T. Leitner and E. Romero-Severson. Phylogenetic patterns recover known HIV epidemiological relationships and reveal common transmission of multiple variants. *Nature Microbiology*, 3(9):983–988, 2018.
- P. Lemey, A. Rambaut, A. J. Drummond, and M. A. Suchard. Bayesian phylogeography finds its roots. *PLoS Computational Biology*, 5(9):e1000520, 2009.
- A. T. Levin, W. P. Hanage, N. Owusu-Boaitey, K. B. Cochran, S. P. Walsh, and G. Meyerowitz-Katz. Assessing the age specificity of infection fatality rates for COVID-19: systematic review, meta-analysis, and public policy implications. *European journal of epidemiology*, pages 1–16, 2020.
- D. Lewandowski, D. Kurowicka, and H. Joe. Generating random correlation matrices based on vines and extended onion method. *Journal of multivariate analysis*, 100(9):1989–2001, 2009.
- T. Lindström, D. A. Gear, M. Buhnerkempe, C. T. Webb, R. S. Miller, K. Portacci, and U. Wennergren. A Bayesian approach for modeling cattle movements in the United States: scaling up a partially observed network. *PLoS One*, 8(1):e53432, 2013.
- U. Mai and S. Mirarab. TreeShrink: fast and accurate detection of outlier long branches in collections of phylogenetic trees. *BMC genomics*, 19(5):23–40, 2018.
- X.-L. Meng and S. Schilling. Warp bridge sampling. *Journal of Computational and Graphical Statistics*, 11(3):552–586, 2002.
- X.-L. Meng and W. H. Wong. Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Statistica Sinica*, pages 831–860, 1996.
- J. Mermin, J. Musinguzi, A. Opio, W. Kirungi, J. P. Ekwaru, W. Hladik, F. Kaharuza, R. Downing, and R. Bunnell. Risk factors for recent HIV infection in Uganda. *Jama*, 300(5):540–549, 2008.
- H. J. Miller, S. Dodge, J. Miller, and G. Bohrer. Towards an integrated science of movement: converging research on animal movement ecology and human mobility science. *International Journal of Geographical Information Science*, 33(5):855–876, 2019.
- M. Monod, A. Blenkinsop, X. Xi, D. Hebert, S. Bershan, S. Tietze, M. Baguelin, V. C. Bradley, Y. Chen, H. Coupland, et al. Age groups that sustain resurging COVID-19 epidemics in the United States. *Science*, 371(6536):eabe8372, 2021.
- N. F. Müller, D. A. Rasmussen, and T. Stadler. The structured coalescent and its approximations. *Molecular Biology and Evolution*, 34(11):2970–2981, 2017.

- N. F. Müller, D. Rasmussen, and T. Stadler. MASCOT: parameter and state inference under the marginal structured coalescent approximation. *Bioinformatics*, 34(22):3843–3848, 2018.
- S. A. Nadeau, T. G. Vaughan, J. Scire, J. S. Huisman, and T. Stadler. The origin and early spread of SARS-CoV-2 in Europe. *Proceedings of the National Academy of Sciences*, 118(9), 2021.
- R. M. Neal. Slice sampling. *The Annals of Statistics*, 31(3):705–767, 2003.
- M. E. Newman. Modularity and community structure in networks. *Proceedings of the national academy of sciences*, 103(23):8577–8582, 2006.
- L.-T. Nguyen, H. A. Schmidt, A. Von Haeseler, and B. Q. Minh. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular biology and evolution*, 32(1):268–274, 2015.
- P. Nouvellet, S. Bhatia, A. Cori, K. E. Ainslie, M. Baguelin, S. Bhatt, A. Boonyasiri, N. F. Brazeau, L. Cattarino, L. V. Cooper, et al. Reduction in mobility and COVID-19 transmission. *Nature communications*, 12(1):1–9, 2021.
- Y. Ogata. A Monte Carlo method for high dimensional integration. *Numerische Mathematik*, 55(2):137–157, 1989.
- T. Paananen, J. Piironen, P.-C. Bürkner, and A. Vehtari. Implicitly adaptive importance sampling. *Statistics and Computing*, 31(2):1–19, 2021.
- O. Papaspiliopoulos, G. O. Roberts, and M. Sköld. A general framework for the parametrization of hierarchical models. *Statistical Science*, pages 59–73, 2007.
- E. Paradis, J. Claude, and K. Strimmer. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics*, 20(2):289–290, 2004.
- A. F. Poon, R. Gustafson, P. Daly, L. Zerr, S. E. Demlow, J. Wong, C. K. Woods, R. S. Hogg, M. Krajden, D. Moore, et al. Near real-time monitoring of HIV transmission hotspots from routine HIV genotyping: an implementation case study. *The Lancet HIV*, 3(5):e231–e238, 2016.
- A. Popinga, T. Vaughan, T. Stadler, and A. J. Drummond. Inferring epidemiological dynamics with Bayesian coalescent inference: the merits of deterministic and stochastic models. *Genetics*, 199(2):595–607, 2015.
- M. N. Price, P. S. Dehal, and A. P. Arkin. FastTree 2—approximately maximum-likelihood trees for large alignments. *PloS one*, 5(3):e9490, 2010.
- W. Probert, M. Hall, X. Xi, R. Sauter, T. Golubchik, D. Bonsall, L. Abeler-Dörner, M. Pickles, A. Cori, J. Bwalya, S. Floyd, N. Mandla, K. Shanaube, B. Yang, H. Ayles, P. Bock, D. Donnell, K. Grabowski, D. Pillay, A. Rambaut, O. Ratmann, S. Fidler, R. Hayes, C. Fraser, P. consortium, and the HPTN 071 (PopART) study team. 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, 2019.

- A. Purvis and P.-M. Agapow. Phylogeny imbalance: taxonomic level matters. *Systematic Biology*, 51(6):844–854, 2002.
- C. E. Rasmussen and C. Williams. *Gaussian processes for Machine Learning*. MIT Press, 2006.
- D. A. Rasmussen, E. Wilkinson, A. Vandormael, F. Tanser, D. Pillay, T. Stadler, and T. De Oliveira. Tracking external introductions of HIV using phylodynamics reveals a major source of infections in rural KwaZulu-Natal, South Africa. *Virus Evolution*, 4(2):vey037, 2018.
- O. Ratmann, A. Van Sighem, D. Bezemer, A. Gavryushkina, S. Jurriaans, A. Wensing, F. De Wolf, P. Reiss, C. Fraser, et al. Sources of HIV infection among men having sex with men and implications for prevention. *Science Translational Medicine*, 8(320):320ra2, 2016.
- O. Ratmann, E. B. Hodcroft, M. Pickles, A. Cori, M. Hall, S. Lycett, C. Colijn, B. Dearlove, X. Didelot, S. Frost, et al. Phylogenetic tools for generalized HIV-1 epidemics: findings from the PANGEA-HIV methods comparison. *Molecular Biology and Evolution*, 34(1):185–203, 2017.
- O. Ratmann, M. K. Grabowski, M. Hall, T. Golubchik, C. Wymant, L. Abeler-Dörner, D. Bonsall, A. Hoppe, A. L. Brown, T. de Oliveira, et al. Inferring HIV-1 transmission networks and sources of epidemic spread in Africa with deep-sequence phylogenetic analysis. *Nature Communications*, 10(1):1–13, 2019.
- O. Ratmann, J. Kagaayi, M. Hall, T. Golubchick, G. Kigozi, X. Xi, C. Wymant, G. Nakigozi, L. Abeler-Dörner, D. Bonsall, et al. Quantifying HIV transmission flow between high-prevalence hotspots and surrounding communities: a population-based study in Rakai, Uganda. *The Lancet HIV*, 7(1):PE173–E183, 2020.
- J. Raymer, A. Wiśniowski, J. J. Forster, P. W. Smith, and J. Bijak. Integrated modeling of European migration. *Journal of the American Statistical Association*, 108(503):801–819, 2013.
- A. Rhoads and K. F. Au. PacBio sequencing and its applications. *Genomics, Proteomics & Bioinformatics*, 13(5):278–289, 2015.
- G. Riutort-Mayol, P.-C. Bürkner, M. R. Andersen, A. Solin, and A. Vehtari. Practical hilbert space approximate bayesian gaussian processes for probabilistic programming. *arXiv preprint arXiv:2004.11408*, 2020.
- E. O. Romero-Severson, I. Bulla, and T. Leitner. Phylogenetically resolving epidemiologic linkage. *Proceedings of the National Academy of Sciences*, 113(10):2690–2695, 2016.
- R. Rose, M. Hall, A. D. Redd, S. Lamers, A. E. Barbier, S. F. Porcella, S. E. Hudelson, E. Piwowar-Manning, M. McCauley, T. Gamble, et al. Phylogenetic methods inconsistently predict the direction of HIV transmission among heterosexual pairs in the HPTN 052 Cohort. *The Journal of Infectious Diseases*, 220(9):1406–1413, 2019.

- L. M. Rossen, A. M. Branum, F. B. Ahmad, P. Sutton, and R. N. Anderson. Excess deaths associated with COVID-19, by age and race and ethnicity—United States, January 26–October 3, 2020. *Morbidity and Mortality Weekly Report*, 69(42):1522, 2020.
- H. Rue and L. Held. *Gaussian Markov random fields: theory and applications*. CRC press, 2005.
- J. S. Santelli, I. Chen, F. Makumbi, Y. Wei, F. Nalugoda, T. Lutalo, E. Spindler, S. A. Grilo, A. Deisher, K. Grabowski, et al. Household wealth and HIV incidence over time, rural Uganda, 1994–2018. *AIDS (London, England)*, 35(11):1835, 2021.
- J. Saul, G. Bachman, S. Allen, N. Toiv, C. Cooney, and T. Beamon. Determined resilient empowered AIDS-free mentored and safe (DREAMS): What is the core package and why now. *PLOS One*, 13(12):e0208167, 2018.
- J. Scire, J. Barido-Sottani, D. Kühnert, T. G. Vaughan, and T. Stadler. Improved multi-type birth-death phylodynamic inference in BEAST 2. *bioRxiv*, page 2020.01.06.895532, 2020.
- N. P. Silliman. Hierarchical selection models with applications in meta-analysis. *Journal of the American Statistical Association*, 92(439):926–936, 1997.
- S. Sinharay and H. S. Stern. An empirical comparison of methods for computing Bayes factors in generalized linear mixed models. *Journal of Computational and Graphical Statistics*, 14(2):415–435, 2005.
- J. Skilling. Nested sampling. In *AIP Conference Proceedings*, volume 735, pages 395–405. American Institute of Physics, 2004.
- P. Skums, A. Zelikovsky, R. Singh, W. Gussler, Z. Dimitrova, S. Knyazev, I. Mandric, S. Ramachandran, D. Campo, D. Jha, et al. QUENTIN: reconstruction of disease transmissions from viral quasispecies genomic data. *Bioinformatics*, 34(1):163–170, 2018.
- A. Solin and S. Särkkä. Hilbert space methods for reduced-rank Gaussian process regression. *Statistics and Computing*, 30:419–446, 2020.
- T. Stadler. Sampling-through-time in birth-death trees. *Journal of Theoretical Biology*, 267(3):396–404, 2010.
- T. Stadler and S. Bonhoeffer. Uncovering epidemiological dynamics in heterogeneous host populations using phylogenetic methods. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 368(1614):20120198, 2013.
- T. Stadler, R. Kouyos, V. von Wyl, S. Yerly, J. Böni, P. Bürgisser, T. Klimkait, B. Joos, P. Rieder, D. Xie, et al. Estimating the basic reproductive number from viral sequence data. *Molecular biology and evolution*, 29(1):347–357, 2012.
- T. Stadler, D. Kühnert, S. Bonhoeffer, and A. J. Drummond. Birth–death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). *Proceedings of the National Academy of Sciences*, 110(1):228–233, 2013.
- A. Stamatakis. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9):1312–1313, 2014.

- Stan Development Team. Stan modeling language users guide and reference manual, version 2.18.0, 2018. URL <http://mc-stan.org/>.
- K. Sun, W. Wang, L. Gao, Y. Wang, K. Luo, L. Ren, Z. Zhan, X. Chen, S. Zhao, Y. Huang, et al. Transmission heterogeneities, kinetics, and controllability of SARS-CoV-2. *Science*, 371(6526), 2021.
- C. Tebaldi and M. West. Bayesian inference on network traffic using link count data. *Journal of the American Statistical Association*, 93(442):557–573, 1998.
- H. A. Thompson, N. Imai, A. Dighe, K. E. Ainslie, M. Baguelin, S. Bhatia, S. Bhatt, A. Boonyasiri, O. Boyd, N. F. Brazeau, et al. SARS-CoV-2 infection prevalence on repatriation flights from Wuhan City, China. *Journal of travel medicine*, 27(8):taaa135, 2020.
- E. Todesco, M. Wirten, R. Calin, A. Simon, S. Sayon, F. Barin, C. Katlama, V. Calvez, A.-G. Marcelin, and S. Hué. Caution is needed in interpreting HIV transmission chains by ultradeep sequencing. *AIDS*, 33(4):691–699, 2019.
- P. Touloupou, N. Alzahrani, P. Neal, S. E. Spencer, and T. J. McKinley. Efficient model comparison techniques for models requiring large scale data augmentation. *Bayesian Analysis*, 13(2):437–459, 2018.
- Uganda AIDS Commission. National strategic plan for HIV & AIDS, 2014, 2014. URL http://www.nationalplanningcycles.org/sites/default/files/country_docs/Uganda/national_strategic_plan_for_hiv_aids_2011_2015.pdf.
- A.-C. Uhlemann, J. Dordel, J. R. Knox, K. E. Raven, J. Parkhill, M. T. Holden, S. J. Peacock, and F. D. Lowy. Molecular tracing of the emergence, diversification, and transmission of *S. aureus* sequence type 8 in a New York community. *Proceedings of the National Academy of Sciences*, 111(18):6738–6743, 2014.
- UNAIDS. 90-90-90: An ambitious treatment target to help end the AIDS epidemic., 2017.
- UNAIDS. Women and girls and HIV., 2018a.
- UNAIDS. Miles to go: closing gaps, breaking barriers, righting justice., 2018b. URL https://www.unaids.org/sites/default/files/media_asset/miles-to-go_en.pdf.
- UNAIDS. UNAIDS Data 2019, 2019. URL <https://www.unaids.org/en/resources/documents/2019/2019-UNAIDS-data>.
- UNAIDS. Seizing the moment - Tackling entrenched inequalities to end epidemics., 2020.
- UNAIDS. Fact Sheet 2021., 2021.
- H. J. T. Unwin, S. Mishra, V. C. Bradley, A. Gandy, T. A. Mellan, H. Coupland, J. Ish-Horowicz, M. A. Vollmer, C. Whittaker, S. L. Filippi, et al. State-level tracking of COVID-19 in the United States. *Nature communications*, 11(1):1–9, 2020.
- J. van de Kastele, J. van Eijkeren, J. Wallinga, et al. Efficient estimation of age-specific social contact rates between men and women. *The Annals of Applied Statistics*, 11(1): 320–339, 2017.

- T. F. Van de Mortel et al. Faking it: social desirability response bias in self-report research. *Australian Journal of Advanced Nursing*, 25(4):40, 2008.
- J. Vanhatalo and A. Vehtari. Sparse log Gaussian processes via MCMC for spatial epidemiology. In *Gaussian processes in practice*, pages 73–89. PMLR, 2007.
- T. G. Vaughan and A. J. Drummond. A stochastic simulator of birth–death master equations with application to phylodynamics. *Molecular Biology and Evolution*, 30(6):1480–1493, 2013.
- T. G. Vaughan, D. Kühnert, A. Poppinga, D. Welch, and A. J. Drummond. Efficient Bayesian inference under the structured coalescent. *Bioinformatics*, 30(16):2272–2279, 2014.
- A. Vehtari, D. Simpson, A. Gelman, Y. Yao, and J. Gabry. Pareto smoothed importance sampling. *arXiv preprint arXiv:1507.02646*, 2015.
- A. Vehtari, A. Gelman, and J. Gabry. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and computing*, 27(5):1413–1432, 2017.
- A. Vehtari, A. Gelman, D. Simpson, B. Carpenter, and P.-C. Bürkner. Rank-normalization, folding, and localization: An improved $\hat{r}$ for assessing convergence of MCMC. *arXiv preprint arXiv:1903.08008*, 2019.
- R. Verity, L. C. Okell, I. Dorigatti, P. Winskill, C. Whittaker, N. Imai, G. Cuomo-Dannenburg, H. Thompson, P. G. Walker, H. Fu, et al. Estimates of the severity of coronavirus disease 2019: a model-based analysis. *The Lancet infectious diseases*, 20(6):669–677, 2020.
- E. Volz and S. Frost. Scalable relaxed clock phylogenetic dating. *Virus evolution*, 3(2), 2017.
- E. M. Volz and S. D. Frost. Inferring the source of transmission with phylogenetic data. *PLoS Comput Biol*, 9(12):e1003397, 2013.
- E. M. Volz and I. Siveroni. Bayesian phylodynamic inference with complex models. *PLoS Computational Biology*, 14(11):e1006546, 2018.
- E. M. Volz, S. L. K. Pond, M. J. Ward, A. J. L. Brown, and S. D. Frost. Phylodynamics of infectious disease epidemics. *Genetics*, 183(4):1421–1430, 2009.
- E. M. Volz, E. Ionides, E. O. Romero-Severson, M.-G. Brandt, E. Mokotoff, and J. S. Koopman. HIV-1 transmission during early infection in Men who have Sex with Men: a phylodynamic analysis. *PLoS Medicine*, 10(12):e1001568, 2013.
- P. G. Walker, C. Whittaker, O. J. Watson, M. Baguelin, P. Winskill, A. Hamlet, B. A. Djafaara, Z. Cucunubá, D. Olivera Mesa, W. Green, et al. The impact of COVID-19 and strategies for mitigation and suppression in low-and middle-income countries. *Science*, 369(6502):413–422, 2020.
- F. Wang and A. E. Gelfand. Modeling space and space-time directional data using projected Gaussian processes. *Journal of the American Statistical Association*, 109(508):1565–1580, 2014.

- M. J. Wawer, R. H. Gray, N. K. Sewankambo, D. Serwadda, X. Li, O. Laeyendecker, N. Kiwanuka, G. Kigozi, M. Kiddugavu, T. Lutalo, et al. Rates of HIV-1 transmission per coital act, by stage of HIV-1 infection, in Rakai, Uganda. *The Journal of infectious diseases*, 191(9):1403–1409, 2005.
- C. Wymant, M. Hall, O. Ratmann, D. Bonsall, T. Golubchik, M. de Cesare, A. Gall, M. Cornelissen, C. Fraser, STOP-HCV Consortium, The Maela Pneumococcal Collaboration, and The BEEHIVE Collaboration. PHYLOSCANNER: inferring transmission from within- and between-host pathogen genetic diversity. *Molecular Biology and Evolution*, 35(3): 719–733, 2017.
- C. Wymant, F. Blanquart, T. Golubchik, A. Gall, M. Bakker, D. Bezemer, N. J. Croucher, M. Hall, M. Hillebregt, S. H. Ong, et al. Easy and accurate reconstruction of whole HIV genomes from short-read sequence data with shiver. *Virus evolution*, 4(1):vey007, 2018.
- X. Xi, S. E. Spencer, M. Hall, M. K. Grabowski, J. Kagaayi, and O. Ratmann. 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*, 2022.
- W. Xie, P. O. Lewis, Y. Fan, L. Kuo, and M.-H. Chen. Improving marginal likelihood estimation for Bayesian phylogenetic model selection. *Systematic biology*, 60(2):150–160, 2011.
- Z. Yang. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *Journal of Molecular evolution*, 39(3):306–314, 1994.
- Z. Yang. A space-time process model for the evolution of DNA sequences. *Genetics*, 139(2): 993–1005, 1995.
- J. Zhang, M. Litvinova, Y. Liang, Y. Wang, W. Wang, S. Zhao, Q. Wu, S. Merler, C. Viboud, A. Vespignani, et al. Changes in contact patterns shape the dynamics of the COVID-19 outbreak in China. *Science*, 368(6498):1481–1486, 2020a.
- Y. Zhang, C. Wymant, O. Laeyendecker, M. K. Grabowski, M. Hall, S. Hudelson, E. Piwowar-Manning, M. McCauley, T. Gamble, M. C. Hosseinipour, et al. Evaluation of phylogenetic methods for inferring the direction of HIV transmission: HPTN 052. *Clinical Infectious Diseases*, Epub ahead of print:ciz1247, 2020b.
- Y. Zhou, A. M. Johansen, and J. A. Aston. Toward automatic model comparison: an adaptive sequential Monte Carlo approach. *Journal of Computational and Graphical Statistics*, 25 (3):701–726, 2016.
- K. Zuma, E. Gouws, B. Williams, and M. Lurie. Risk factors for HIV infection among women in Carletonville, South Africa: migration, demography and sexually transmitted diseases. *International Journal of STD & AIDS*, 14(12):814–817, 2003.

Appendix A

Modelling and estimation of the sampling cascade

As motivated in Chapter 5.2, we plan to model the sampling process separately and then integrate with modelling transmission flows. In this Appendix we will describe the statistical models to characterise each step of sampling cascade in Figure 5.8.

To recall our setting, we defined transmission events as Human Immunodeficiency Virus (HIV) infection events to individuals who lived in Rakai Community Cohort Study (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 (5.1-5.2) 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 to 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 49, yielding $A = 140$ population strata. The actual transmission events from group a to group b are

$$z_{ab} = \sum_{i \in a, j \in b} z_{ij}, \quad (\text{A.1})$$

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

$$n_{ab} = \sum_{i \in a, j \in b} \mathbb{1} \{s_i^S = 1\} \mathbb{1} \{s_j^R = 1\} z_{ij}. \quad (\text{A.2})$$

The sampling cascade involved that sources and recipients participate in the cohort, report no antiretroviral therapy (ART) use, and that virus was successfully deep-sequenced. Section A.1 describes the estimation of participation probabilities for source cases, and Section A.2 describes the estimation of sequence sampling probabilities conditional on participation among source cases. Section A.3 describes the estimation of participation probabilities for recipient cases, and Section A.4 describes the estimation of sequence sampling probabilities conditional on participation among recipient cases. Section A.5 illustrates the resulting sampling probabilities of transmission events. In each step, we assessed model performance through 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).

A.1 Modelling and estimating participation probabilities of source cases

We are interested in participation probabilities among infected individuals who transmitted the 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 $\xi^P = (\xi_a^P)_{a \in \mathcal{A}}$. 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

$$\begin{aligned}
 N_a^p &\sim \text{Beta-Binomial}(N_a^e, \xi_a^p, \gamma), \forall a \\
 \text{logit}(\xi^p) &= \beta_0 \mathbf{1} + \beta \mathbf{X} \\
 \beta_0 &\sim \text{Normal}(0, 100) \\
 \beta &\sim \text{Normal}(0, 10\mathbf{I}) \\
 \gamma &\sim \text{Exp}(1),
 \end{aligned} \tag{A.3}$$

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, $\beta \in \mathbb{R}^p$ is the vector of regression coefficients, and $\mathbf{I}$ is the $p \times p$ identity matrix. Four models were considered and assessed using 10-fold cross-validation.

- Model P1. Model (A.3) 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 (A.3) 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 A.1 shows the proportion of hold-out data in 95% PPCrI, the MAE, and the ELPD for each model in 10-fold cross-validations for each model. Based on these statistics, model P4 was chosen to model participation probabilities across population strata. Figure A.1 summarises the marginal posterior densities of the regression coefficients under model P4. Figure A.2 shows the marginal posterior participation probabilities by location, gender, and age bands.

Table A.1 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)

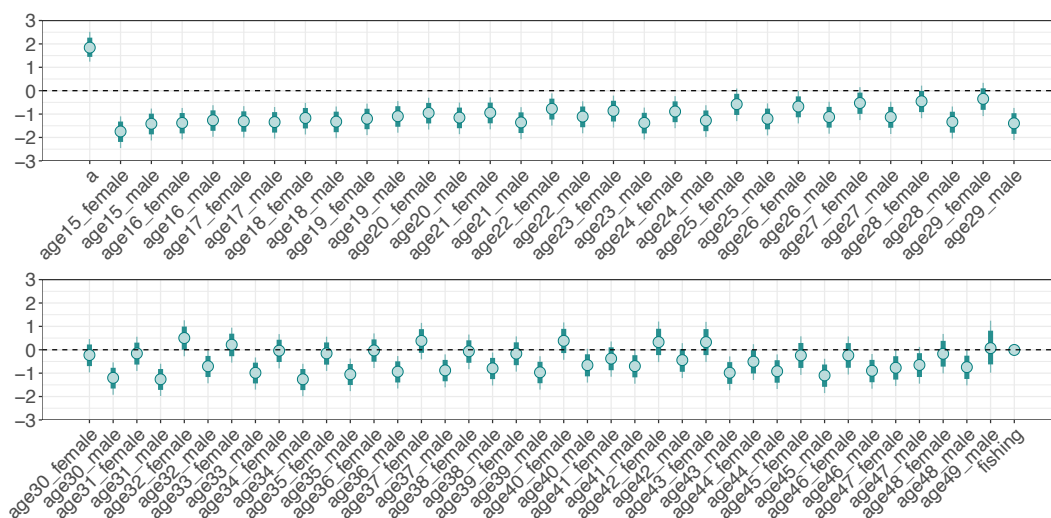


Figure A.1 Marginal posterior densities of regression coefficients of model P4 to estimate population participation probabilities. Circle: median; bar: interquartile range; line: 95% credibility interval.

A.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 a virus successfully deep-sequenced. Considerable demographic and clinical data were collected among RCCS participants, which allowed us to relax the assumptions of Section A.1 when modelling and estimating the sequencing probabilities of source cases. In general, many infected individuals do not transmit the virus onwards, and individuals

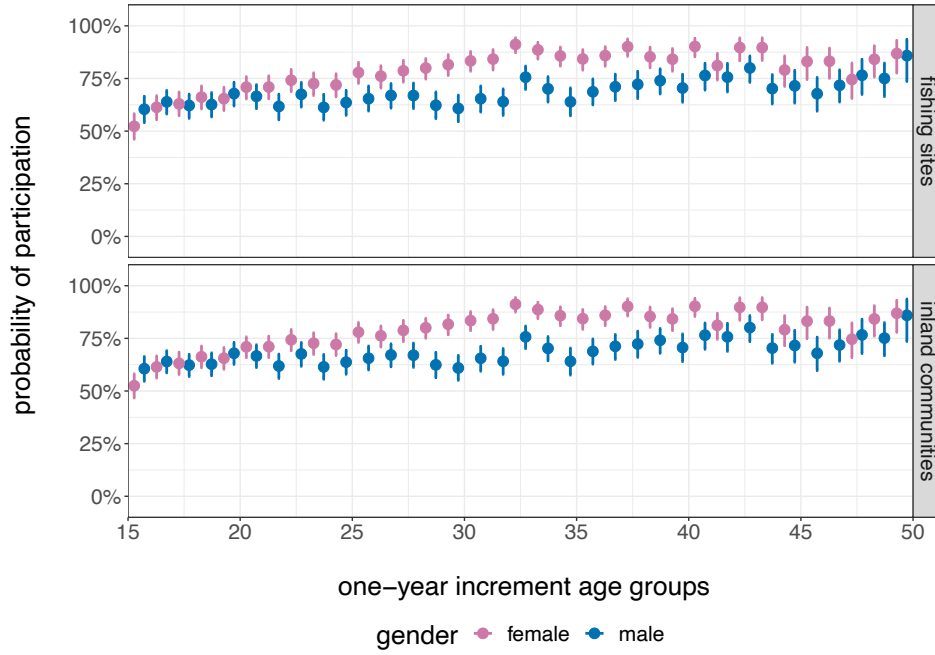


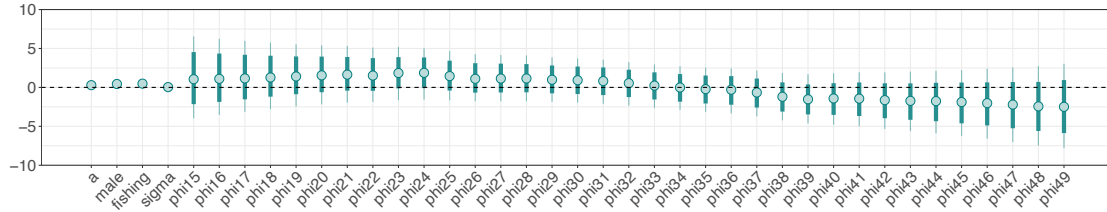
Figure A.2 **Marginal posterior densities of population participation probabilities under model P4.** Circle: median; line: 95% credibility interval.

who transmitted the 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 $\xi^s = (\xi_a^s)_{a \in \mathcal{A}}$, 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 the notation of (A.3) and further denote by N_a^{naive} the number of infected participants in a who did not report consistent

Table A.2 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)

Figure A.3 **Marginal posterior densities of regression coefficients of model S5 to estimate conditional deep-sequencing probabilities.** Circle: median; bar: interquartile range; line: 95% credibility interval.

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

$$\begin{aligned}
 N_a^s &\sim \text{Beta-Binomial}(N_a^{\text{naive}}, \xi_a^s, \gamma), \forall a \\
 \text{logit}(\xi^s) &= \beta_0 \mathbf{1} + \beta \mathbf{X} \\
 \beta_0 &\sim \text{Normal}(0, 100) \\
 \beta &\sim \text{Normal}(0, 10\mathbf{I}) \\
 \gamma &\sim \text{Exp}(1),
 \end{aligned} \tag{A.4}$$

and similar models with first-order intrinsic conditional auto-regressive (ICAR) prior distributions across adjacent age bands (Rue and Held, 2005). In specific, neighbouring age groups are linked in the graph associated with ICAR. Eight models were considered and assessed using 10-fold cross-validation.

- Model S1. Model (A.4) 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 (A.4) 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 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.

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 A.2 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 A.1 shows the marginal posterior densities of the regression coefficients under model S5. Figure A.4 summarises the marginal posterior deep-sequencing probabilities by location, gender, and age bands.

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

$$\xi_a^S = \xi_a^P * \xi_a^s, \quad (\text{A.5})$$

for all population sub-groups a .

A.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 as in Section A.1 and used the RCCS census data to approximate the participation probabilities

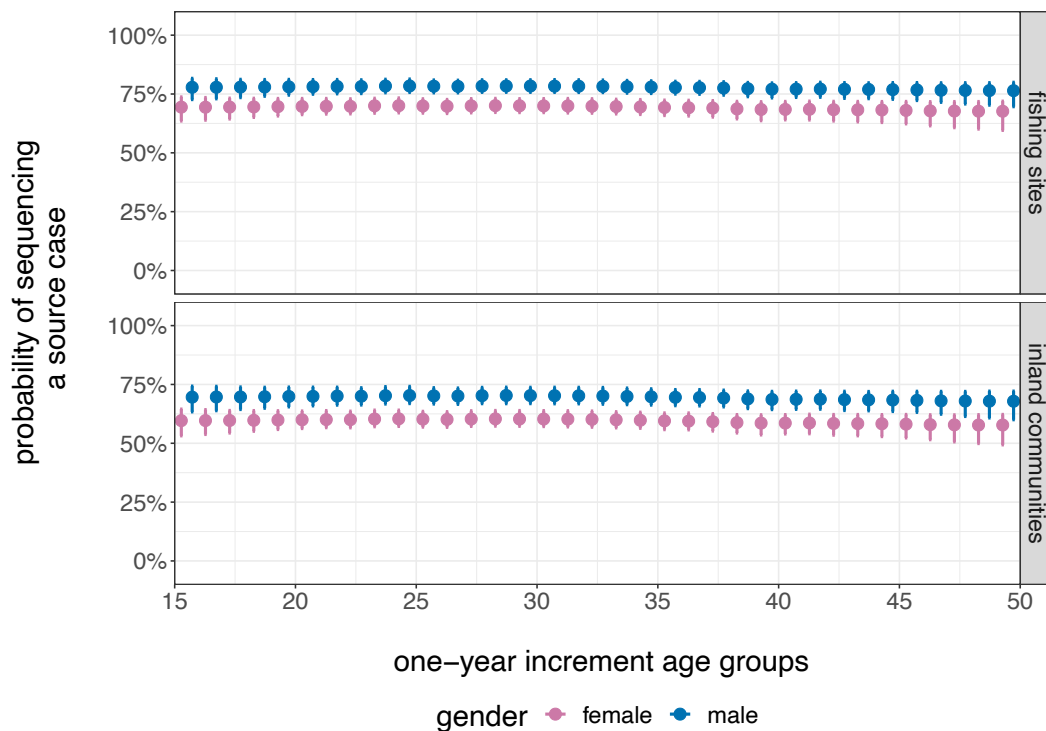


Figure A.4 **Marginal posterior densities of source sequencing probabilities under model S5.** Circle: median; line: 95% credibility interval.

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 A.1, we adjusted for heterogeneity in participation probabilities among newly infected individuals using the posterior estimates under model P4, reported in Figure A.2.

A.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 A.2, we combined both steps and sought

Table A.3 Prediction accuracy of models N1-N3 on hold-out infection time data in 10-fold cross-validation.

Model	Hold-out data in 95% PPCrI	MAE	ELPD
	median (range)	median (range)	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)

to estimate the probability that newly infected participants had a virus successfully deep-sequenced.

Based on available data collected among RCCS participants, we could again relax the assumptions of Section A.3. Our approach here differed in two aspects from the approach we employed to estimate sequencing probabilities among source cases in Section A.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 (Golubchik et al., 2017; Grabowski 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 A.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 $\mathcal{T}$. Information on the 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.

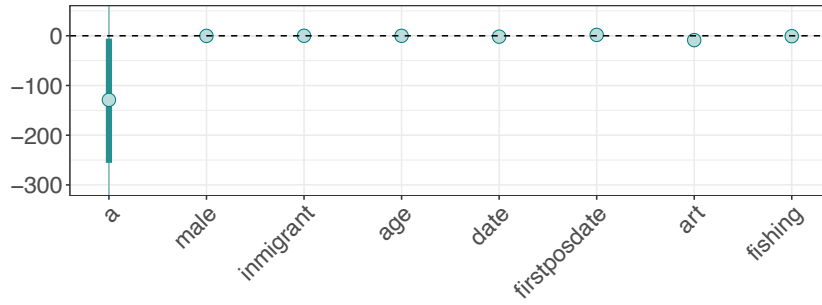


Figure A.5 **Marginal posterior densities of regression coefficients of model N3 to estimate conditional deep-sequencing probabilities.** Circle: median; bar: interquartile range; line: 95% credibility interval.

The Bernoulli regression models had the form

$$\begin{aligned}
 Y_i &\sim \text{Bernoulli}(\eta_i), \forall i \\
 \text{logit}(\eta) &= \beta_0 \mathbf{1} + \beta \mathbf{X} \\
 \beta_0 &\sim \text{Normal}(0, 100) \\
 \beta &\sim \text{Normal}(0, 10\mathbf{I}),
 \end{aligned} \tag{A.6}$$

where η_i is the probability that the i th 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 the time of first positive test (indicator variable), and ART use at first visit date when infected (indicator). Three models were considered.

- Model N1. Model (A.6) with the aforementioned covariates except for ART status and time of first positive test, leading to $p = 5$ predictors.
- Model N2. Model (A.6) with the aforementioned covariates except ART status, leading to $p = 6$ predictors.
- Model N3. Model (A.6) 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 A.3 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 A.5 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 A.4 summarises the number of infected participants classified as having acquired infection before or during the study period between 2009-10-01 and 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 A.2 we used a Beta-Binomial regression framework to estimate the posterior distribution of deep-sequencing probabilities $\xi^{s'} = (\xi_a^{s'})_{a \in \mathcal{A}}$, 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 A.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

$$\begin{aligned}
 N_a^{s'} &\sim \text{Beta-Binomial}(N_a^{\text{study}}, \xi_a^{s'}, \gamma), \forall a \\
 \text{logit}(\xi^{s'}) &= \beta_0 \mathbf{1} + \beta \mathbf{X} \\
 \beta_0 &\sim \text{Normal}(0, 100) \\
 \beta &\sim \text{Normal}(0, 10\mathbf{I}) \\
 \gamma &\sim \text{Exp}(1),
 \end{aligned} \tag{A.7}$$

and potentially including regularising ICAR prior densities as in Section A.2. We compared eight models S'1-S'8 that were analogous to models S1-S8 in Section A.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 A.5 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 A.6 shows the marginal posterior densities of the regression

Table A.4 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 A.5 Prediction accuracy of models S'1-S'8 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)
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)

coefficients under model S'5, and Figure A.7 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 A.8. 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 the 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

$$\xi_a^R = \xi_a^P * \xi_a^{s'}, \quad (\text{A.8})$$

for all population sub-groups a .

A.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 (A.5) and (A.8), 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

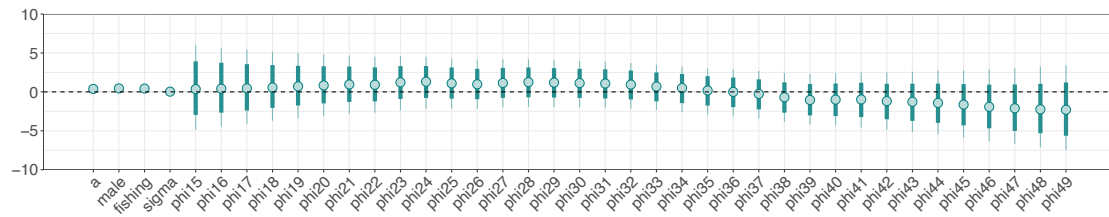


Figure A.6 **Marginal posterior densities of regression coefficients of model S/5 to estimate conditional deep-sequencing probabilities among infected participants who were classified to have acquired HIV during the study period.** Circle: median; bar: interquartile range; line 95% credibility interval.

cases are independently sampled. Figure A.9 illustrates the resulting sampling probabilities of transmission events in inland communities, and Figure A.10 illustrates the sampling probabilities of transmission events in fishing communities.

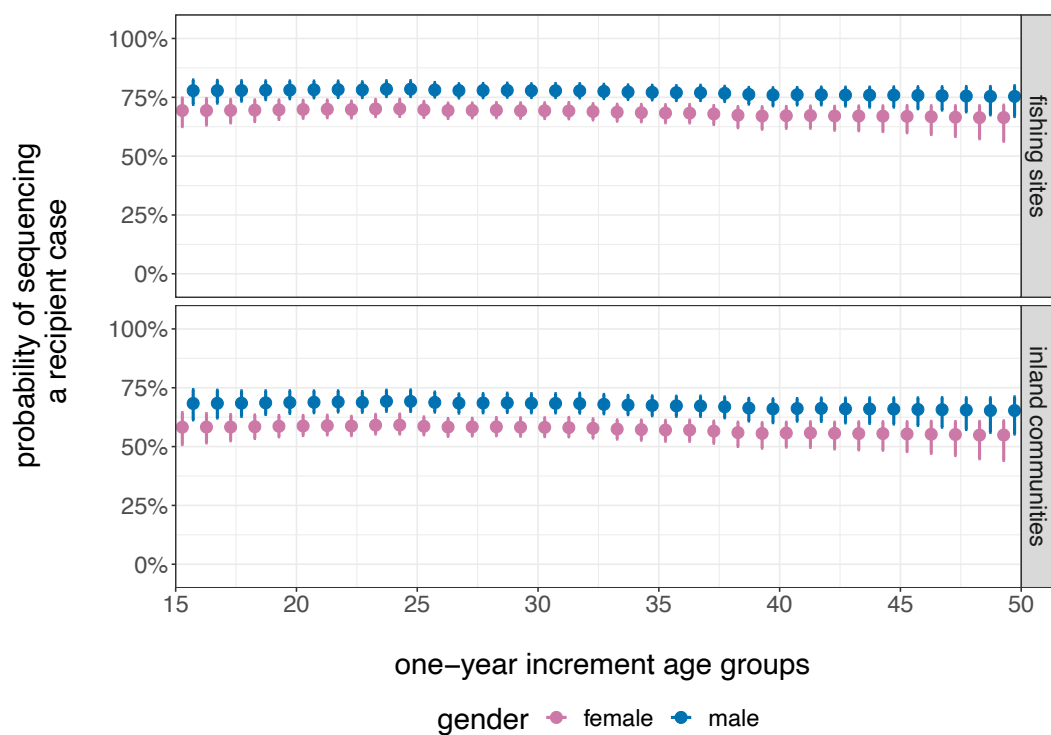


Figure A.7 **Marginal posterior densities of recipient sequencing probabilities under model S'5.** Circle: median; line: 95% credibility interval.

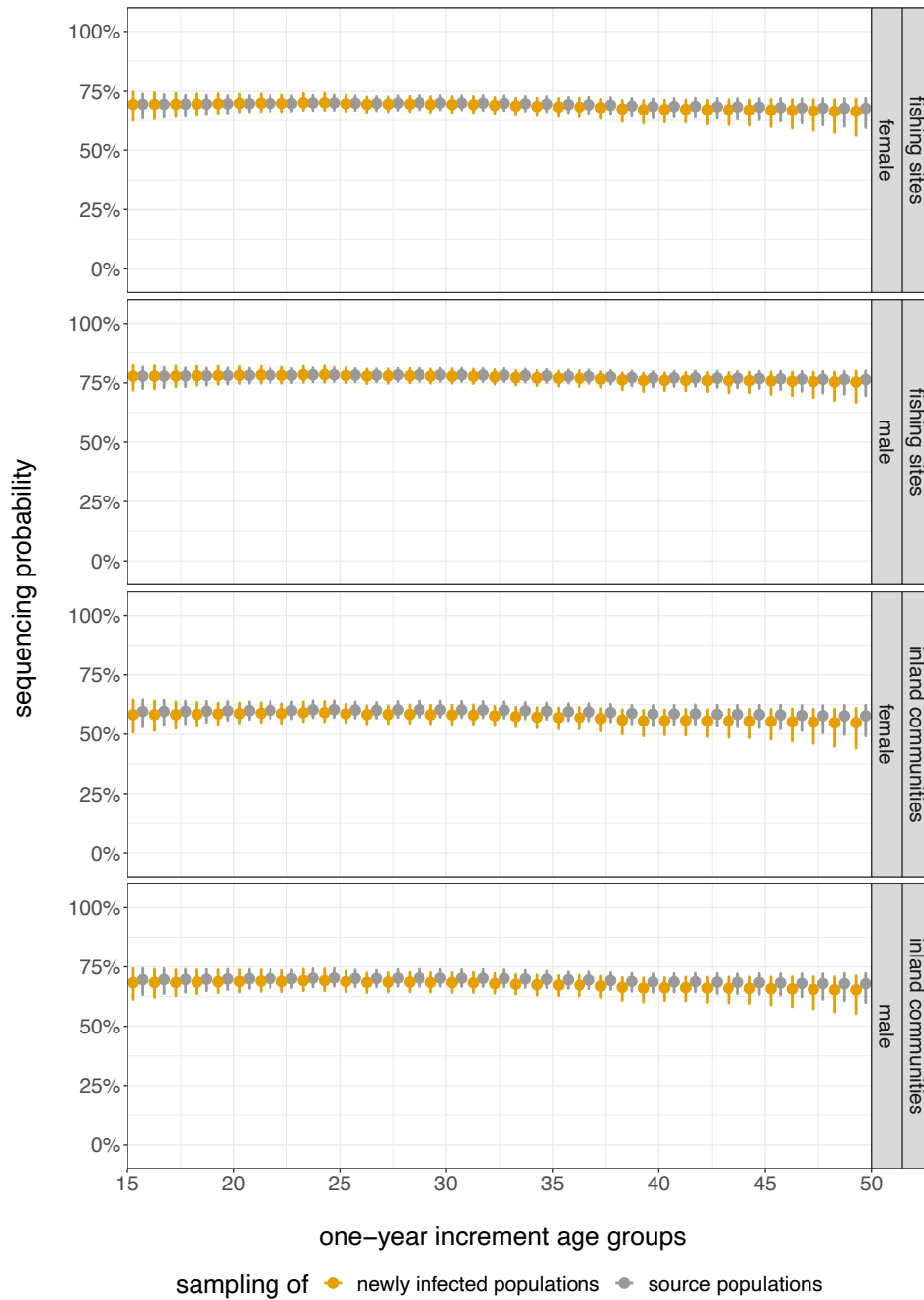


Figure A.8 **Comparison of marginal posterior densities of regression coefficients between model S5 and model S'5.** Circle: median; bar: interquartile range; line 95% credibility interval.

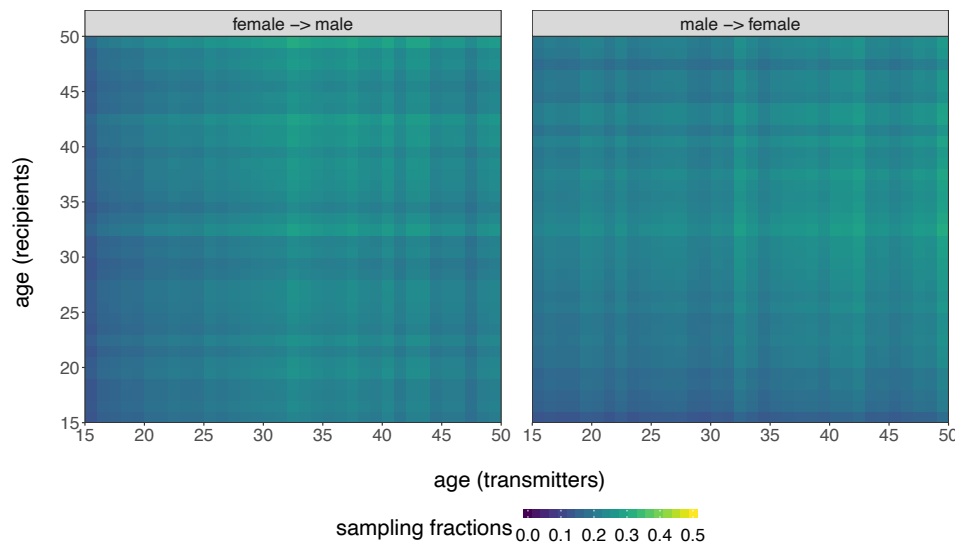


Figure A.9 Estimated sampling probabilities of transmission events in RCCS inland communities, by 1-year age bands. Sampling probabilities of transmission events were calculated by multiplying the sampling probabilities of source and recipient populations. Shown are the posterior median estimates of 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.

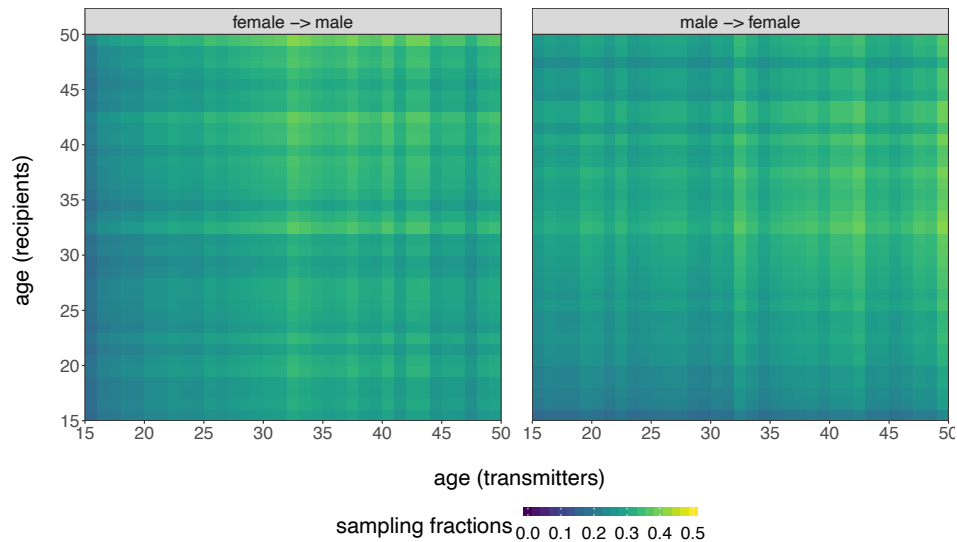


Figure A.10 Estimated sampling probabilities of transmission events in RCCS fishing communities, by 1-year age bands. Shown are the posterior median estimates of the sampling probabilities by the age of source and recipient populations at the midpoint of the study period. Left panel: female to male transmission. Right panel: male to female transmission.

Appendix B

Appendix for Chapter 4

The Appendix derives the covariance matrix for the model that describes gender-specific transmission directions. Recall the source locations $\mathbf{x} \in \mathbb{R}^{n^+ \times 2} = (\mathbf{x}_{\text{mf}}, \mathbf{x}_{\text{fm}})^T$ where $\mathbf{x}_{\text{mf}} \in \mathbb{R}^{n_1^+ \times 2}$ and $\mathbf{x}_{\text{fm}} \in \mathbb{R}^{n_2^+ \times 2}$, the covariance matrix $\mathbf{k}(\mathbf{x}, \mathbf{x})$ is then

$$\mathbf{k}(\mathbf{x}, \mathbf{x}) = \begin{bmatrix} \mathbf{k}_{1,1}(\mathbf{x}, \mathbf{x}) & \mathbf{k}_{1,2}(\mathbf{x}, \mathbf{x}) \\ \mathbf{k}_{2,1}(\mathbf{x}, \mathbf{x}) & \mathbf{k}_{2,2}(\mathbf{x}, \mathbf{x}) \end{bmatrix} \in \mathbb{R}^{2n^+ \times 2n^+}, \quad (\text{B.1})$$

where

$$\begin{aligned} \mathbf{k}_{1,1}(\mathbf{x}, \mathbf{x}) &= \begin{bmatrix} \mathbf{k}_{1,1,\text{mf,mf}}(\mathbf{x}_{\text{mf}}, \mathbf{x}_{\text{mf}}) & \mathbf{k}_{1,1,\text{mf,fm}}(\mathbf{x}_{\text{mf}}, \mathbf{x}_{\text{fm}}) \\ \mathbf{k}_{1,1,\text{fm,mf}}(\mathbf{x}_{\text{fm}}, \mathbf{x}_{\text{mf}}) & \mathbf{k}_{1,1,\text{fm,fm}}(\mathbf{x}_{\text{fm}}, \mathbf{x}_{\text{fm}}) \end{bmatrix} \\ &= \begin{bmatrix} \alpha_{C_1}^2 C_{1,1} \alpha_{H_1}^2 H_{1,1} k(\mathbf{x}_{\text{mf}}, \mathbf{x}_{\text{mf}}) & \alpha_{C_1}^2 C_{1,1} \alpha_{H_1} \alpha_{H_2} H_{1,2} k(\mathbf{x}_{\text{mf}}, \mathbf{x}_{\text{fm}}) \\ \alpha_{C_1}^2 C_{1,1} \alpha_{H_2} \alpha_{H_1} H_{2,1} k(\mathbf{x}_{\text{fm}}, \mathbf{x}_{\text{mf}}) & \alpha_{C_1}^2 C_{1,1} \alpha_{H_2}^2 H_{2,2} k(\mathbf{x}_{\text{fm}}, \mathbf{x}_{\text{fm}}) \end{bmatrix} \quad (\text{B.2}) \\ &= \alpha_{C_1}^2 C_{1,1} \begin{bmatrix} \mathbf{1}_{n_1^+, n_1^+} \alpha_{H_1}^2 H_{1,1} & \mathbf{1}_{n_1^+, n_2^+} \alpha_{H_1} \alpha_{H_2} H_{1,2} \\ \mathbf{1}_{n_2^+, n_1^+} \alpha_{H_2} \alpha_{H_1} H_{2,1} & \mathbf{1}_{n_2^+, n_2^+} \alpha_{H_2}^2 H_{2,2} \end{bmatrix} \circ k(\mathbf{x}, \mathbf{x}), \end{aligned}$$

and $\mathbf{1}_{n_1^+, n_2^+}$ denotes the $n_1^+ \times n_2^+$ matrix of ones. Therefore (B.1) rewrites to

$$\begin{aligned} \mathbf{k}(\mathbf{x}, \mathbf{x}) &= (\text{diag}(\alpha_C) \mathbf{C} \text{diag}(\alpha_C)) \\ &\quad \otimes \left(\begin{bmatrix} \mathbf{1}_{n_1^+, n_1^+} \alpha_{H_1}^2 H_{1,1} & \mathbf{1}_{n_1^+, n_2^+} \alpha_{H_1} \alpha_{H_2} H_{1,2} \\ \mathbf{1}_{n_2^+, n_1^+} \alpha_{H_2} \alpha_{H_1} H_{2,1} & \mathbf{1}_{n_2^+, n_2^+} \alpha_{H_2}^2 H_{2,2} \end{bmatrix} \circ k(\mathbf{x}, \mathbf{x}) \right). \quad (\text{B.3}) \end{aligned}$$

Appendix C

Appendix for Chapter 6

This Appendix derives the spectral density of square exponential kernel with bivariate input (6.4), and verifies eigenfunctions (6.5) of Laplacian operator that are important in Hilbert Space Gaussian Process Approximation (HSGP). The square exponential covariance function in 1D is

$$k(a_1, a_2) = \sigma^2 \exp\left(-\frac{(a_2 - a_1)^2}{2\ell^2}\right). \quad (\text{C.1})$$

Let $\kappa = a_2 - a_1$, and the stationary covariance function depends on κ only. The covariance function can be represented as the Fourier transform of a positive finite measure. If the measure has the density $S(\omega)$, the density is called spectral density or power spectrum corresponding to the covariance function,

$$S(\omega) = \int k(\kappa) \exp(-i\omega\kappa) d\kappa. \quad (\text{C.2})$$

By (C.1) and (C.2),

$$\begin{aligned} S(\omega) &= \int \sigma^2 \exp\left(-\frac{\kappa^2}{2\ell^2}\right) \exp(-i\omega\kappa) d\kappa \\ &= \sigma^2 \int \exp\left(-\frac{\kappa^2 + 2i\omega\kappa\ell^2}{2\ell^2}\right) d\kappa \\ &= \sigma^2 \exp\left(-\frac{\omega^2\ell^4}{2\ell^2}\right) \int \exp\left(-\frac{(\kappa + i\omega\ell^2)^2}{2\ell^2}\right) d\kappa \\ &= \sqrt{2\pi}\ell\sigma^2 \exp\left(-\frac{\omega^2\ell^2}{2}\right). \end{aligned} \quad (\text{C.3})$$

The kernel in the two dimensional space was defined in (6.2), and we simplified the notations,

$$k((a_1, b_1), (a_2, b_2)) = \sigma^2 \exp(-[\frac{(a_2 - a_1)^2}{2\ell_1^2} + \frac{(b_2 - b_1)^2}{2\ell_2^2}]). \quad (\text{C.4})$$

Again let $\kappa_1 = a_2 - a_1$ and $\kappa_2 = b_2 - b_1$, and then the spectral density is

$$\begin{aligned} S(\omega_1, \omega_2) &= \int \int k(\kappa_1, \kappa_2) \exp(-i[\omega_1 \kappa_1 + \omega_2 \kappa_2]) d\kappa_1 d\kappa_2 \\ &= \int \int \sigma^2 \exp(-\frac{\tau_1^2}{\ell_1^2}) \exp(-\frac{\tau_2^2}{\ell_2^2}) \exp(-i\omega_1 \kappa_1) \exp(-i\omega_2 \kappa_2) d\kappa_1 d\kappa_2. \end{aligned} \quad (\text{C.5})$$

Therefore

$$\begin{aligned} S(\omega_1, \omega_2) &= \sqrt{2\pi\sigma^2}\ell_1 \exp(-\frac{\omega_1^2 \ell_1^2}{2}) \sqrt{2\pi\sigma^2}\ell_2 \exp(-\frac{\omega_2^2 \ell_2^2}{2}) \\ &= 2\pi\sigma^2 \prod_{d=1}^2 \ell_d \exp\left(-\frac{1}{2} \sum_{d=1}^2 \ell_d^2 \omega_d^2\right). \end{aligned} \quad (\text{C.6})$$

Then we verified the eigenvalues and eigenfunctions of the Laplacian operator (6.5) in univariate case,

$$\begin{aligned} \Lambda_j &= \left(\frac{j\pi}{2B}\right)^2, \\ \phi_j(x) &= \sqrt{\frac{1}{B}} \sin(\sqrt{\Lambda_j}(x+B)). \end{aligned} \quad (\text{C.7})$$

The Laplacian of $\phi_j(x)$ in (C.7) is

$$\begin{aligned} -\nabla^2 \phi_j(x) &= -\nabla^2 \frac{1}{\sqrt{B}} \sin\left(\frac{j\pi}{2B}(x+B)\right) \\ &= -\nabla \frac{j\pi}{2B} \frac{1}{\sqrt{B}} \cos\left(\frac{j\pi}{2B}(x+B)\right) \\ &= \frac{1}{\sqrt{B}} \left(\frac{j\pi}{2B}\right)^2 \sin\left(\frac{j\pi}{2B}(x+B)\right) \\ &= \Lambda_j \phi_j, \end{aligned} \quad (\text{C.8})$$

where $x \in \Omega = [-B, B]$. On the boundaries,

$$\begin{aligned} -\nabla^2 \phi_j(-B) &= 0 \\ -\nabla^2 \phi_j(B) &= \frac{1}{B} \sin\left(\frac{j\pi}{2B} 2B\right) = 0. \end{aligned} \quad (\text{C.9})$$

So (C.7) satisfies

$$\begin{aligned} -\nabla^2 \phi &= \Lambda \phi, \in \Omega \\ \Lambda &= 0, \in \partial \Omega, \end{aligned} \quad (\text{C.10})$$

and they are eigenvalues and eigenfunctions of Laplacian operator within $[-B, B]$ under Dirichlet boundary condition. In the two dimensional case, we aim to verify Equation (6.5) gives eigenvalues and eigenfunctions of Laplacian operator within the domain $\Omega = [-B_1, B_1] \times [-B_2, B_2]$. The Laplacian of the bivariate function $\phi_j(\mathbf{x})$ is

$$-\nabla^2 \phi_j = -\nabla^2 \phi_{1j} - \nabla^2 \phi_{2j}. \quad (\text{C.11})$$

The eigenvalues and eigenvectors in the first dimension are

$$\begin{aligned} \Lambda_{1j} &= \left(\frac{j\pi}{2B_1}\right)^2 \\ \phi_{1j}(x_1) &= \sqrt{1/B_1} \sin(\sqrt{\Lambda_{1j}}(x_1 + B_1)), \end{aligned} \quad (\text{C.12})$$

and this is of the same form as (C.7), so we followed (C.8), and

$$-\nabla^2 \phi_{1j} = \Lambda_{1j} \phi_{1j}. \quad (\text{C.13})$$

This applies to the second dimension as well. Therefore

$$-\nabla^2 \phi_j = -\nabla^2 \phi_{1j} - \nabla^2 \phi_{2j} = \Lambda_{1j} \phi_{1j} + \Lambda_{2j} \phi_{2j}, \quad (\text{C.14})$$

and the boundary conditions hold as in (C.9).

In the end, we present the identity that enables the transformation from (6.7) to (6.8). Suppose $\mathbf{A} \in \mathbb{R}^{n \times m}$ and $\mathbf{D} \in \mathbb{R}^{m \times m}$ where $\mathbf{D}$ is a diagonal matrix with the i th element D_{ii} . Then

$$(\mathbf{A} \mathbf{D} \mathbf{A}^T)_{ij} = \sum_{k=1}^m A_{ik} D_{kk} A_{kj}^T = \sum_{k=1}^m A_{ik} D_{kk} A_{jk}. \quad (\text{C.15})$$

Appendix D

Appendix for Chapter 7

The Appendix describes the derivation of Equation (7.33).

$$\begin{aligned}
& \int \frac{1}{s_1 \frac{p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{g(\boldsymbol{\theta})} + s_2 p(\mathbf{y})} p(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta} \\
&= \int \frac{1}{s_1 \frac{\frac{|\mathbf{L}_{\hat{\Sigma}_{\boldsymbol{\theta}}}|}{2} [q(\mathbf{L}_{\hat{\Sigma}_{\boldsymbol{\theta}}}\boldsymbol{\theta} + \hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}) + q(-\mathbf{L}_{\hat{\Sigma}_{\boldsymbol{\theta}}}\boldsymbol{\theta} + \hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}})]}{g(\boldsymbol{\theta})} + s_2 p(\mathbf{y})} \frac{|\mathbf{L}_{\hat{\Sigma}_{\boldsymbol{\theta}}}|}{2} [q(\mathbf{L}_{\hat{\Sigma}_{\boldsymbol{\theta}}}\boldsymbol{\theta} + \hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}) + q(-\mathbf{L}_{\hat{\Sigma}_{\boldsymbol{\theta}}}\boldsymbol{\theta} + \hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}})] d\boldsymbol{\theta} \\
&= \int \frac{1}{s_1 \frac{\frac{|\mathbf{L}_{\hat{\Sigma}_{\boldsymbol{\theta}}}|}{2} [q(\boldsymbol{\eta}) + q(2\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}} - \boldsymbol{\eta})]}{g(\mathbf{L}_{\hat{\Sigma}_{\boldsymbol{\theta}}}^{-1}(\boldsymbol{\eta} - \hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}))} + s_2 p(\mathbf{y})} \frac{|\mathbf{L}_{\hat{\Sigma}_{\boldsymbol{\theta}}}|}{2} [q(\boldsymbol{\eta}) + q(2\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}} - \boldsymbol{\eta})] \frac{1}{|\mathbf{L}_{\hat{\Sigma}_{\boldsymbol{\theta}}}|} d\boldsymbol{\eta} \\
&= \frac{1}{2} \int \frac{1}{s_1 \frac{\frac{|\mathbf{L}_{\hat{\Sigma}_{\boldsymbol{\theta}}}|}{2} [q(\boldsymbol{\eta}) + q(2\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}} - \boldsymbol{\eta})]}{g(\mathbf{L}_{\hat{\Sigma}_{\boldsymbol{\theta}}}^{-1}(\boldsymbol{\eta} - \hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}))} + s_2 p(\mathbf{y})} q(\boldsymbol{\eta}) d\boldsymbol{\eta} \\
&\quad + \frac{1}{2} \int \frac{1}{s_1 \frac{\frac{|\mathbf{L}_{\hat{\Sigma}_{\boldsymbol{\theta}}}|}{2} [q(\boldsymbol{\eta}) + q(2\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}} - \boldsymbol{\eta})]}{g(\mathbf{L}_{\hat{\Sigma}_{\boldsymbol{\theta}}}^{-1}(\boldsymbol{\eta} - \hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}))} + s_2 p(\mathbf{y})} q(2\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}} - \boldsymbol{\eta}) d\boldsymbol{\eta} \\
&= \int \frac{1}{s_1 \frac{\frac{|\mathbf{L}_{\hat{\Sigma}_{\boldsymbol{\theta}}}|}{2} [q(\boldsymbol{\eta}) + q(2\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}} - \boldsymbol{\eta})]}{g(\mathbf{L}_{\hat{\Sigma}_{\boldsymbol{\theta}}}^{-1}(\boldsymbol{\eta} - \hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}))} + s_2 p(\mathbf{y})} q(\boldsymbol{\eta}) d\boldsymbol{\eta},
\end{aligned}$$

(D.1)

where we used the symmetry of $q(\eta)$ at $\hat{\mu}_\theta$, i. e. $q(\hat{\mu}_\theta - \eta) = q(\hat{\mu}_\theta + \eta)$,

$$\begin{aligned}
& \int \frac{1}{s_1 \frac{|\mathbf{L}_{\hat{\Sigma}_\theta}|^{\frac{1}{2}} [q(\eta) + q(2\hat{\mu}_\theta - \eta)]}{g(\mathbf{L}_{\hat{\Sigma}_\theta}^{-1}(\eta - \hat{\mu}_\theta))} + s_2 p(\mathbf{y})} q(2\hat{\mu}_\theta - \eta) d\eta \\
&= \int \frac{1}{s_1 \frac{|\mathbf{L}_{\hat{\Sigma}_\theta}|^{\frac{1}{2}} [q(\hat{\mu}_\theta + (\eta - \hat{\mu}_\theta)) + q(\hat{\mu}_\theta - (\eta - \hat{\mu}_\theta))]}{g(\mathbf{L}_{\hat{\Sigma}_\theta}^{-1}(\eta - \hat{\mu}_\theta))} + s_2 p(\mathbf{y})} q(\hat{\mu}_\theta - (\eta - \hat{\mu}_\theta)) d\eta \\
&= \int \frac{1}{s_1 \frac{|\mathbf{L}_{\hat{\Sigma}_\theta}|^{\frac{1}{2}} [q(\hat{\mu}_\theta - (\eta - \hat{\mu}_\theta)) + q(\hat{\mu}_\theta + (\eta - \hat{\mu}_\theta))]}{g(\mathbf{L}_{\hat{\Sigma}_\theta}^{-1}(\eta - \hat{\mu}_\theta))} + s_2 p(\mathbf{y})} q(\hat{\mu}_\theta + (\eta - \hat{\mu}_\theta)) d\eta.
\end{aligned} \tag{D.2}$$

Appendix E

Codes

The Appendix describes the important codes in the thesis, including the pre-processing in Appendix E.1 and modelling in Appendix E.2 and E.3.

E.1 Bioinformatics pipeline for processing HIV deep sequence data

The code for running the procedure in Section 3.2 is in https://github.com/olli0601/Phyloscanner.R/blob/master/misc_data_analysis_RCCS1519/Rk1519_02_make_potential_transmission_networks.R, the code for Section 3.3 is in https://github.com/olli0601/Phyloscanner.R/blob/master/misc_data_analysis_RCCS1519/Rk1519_03_make_trees.R, the code for Section 3.4 is in https://github.com/olli0601/Phyloscanner.R/blob/master/misc_data_analysis_RCCS1519/Rk1519_04_pairwise_relationship.R. The workflow is organised in https://github.com/olli0601/Phyloscanner.R/tree/master/misc_data_analysis_RCCS1519/software.

In addition, we discuss several options of phyloscanner in Section 3.4, and here we present the full command of phyloscanner when the down-sampling is not applied,

```
/phyloscanner/phyloscanner_analyse_trees.R "/ptyr1_trees" ptyr1 s,15
--allowMultiTrans --distanceThreshold 0.02 0.05
--fileNameRegex "^(?:.*\D)?([0-9]+)_to_([0-9]+).*$"
--minReadsPerHost 30 --minTipsPerHost 1
--multifurcationThreshold 1e-05 --multinomial
--noProgressBars --normRefFileName "~/normalisation_ByPosition.csv"
--normStandardiseGagPol --outgroupName "REF_CON_H"
```

```
--outputRDA --overwrite --parsimonyBlacklistK 15
--postHocCountBlacklisting --ratioBlacklistThreshold 0.01
--rawBlacklistThreshold 3 --readCountsMatterOnZeroLengthBranches
--relaxedAncestry --seed 42
--tipRegex "^(.*)-fq[0-9]+_read_([0-9]+)_count_([0-9]+) $"
--treeFileExtension ".treefile" --verbose 1.
```

To down-sample to 100 reads for each host, we use

```
--maxReadsPerHost 100.
```

The option in Section 3.4.3 is switched on by

```
--zeroLengthAdjustment.
```

E.2 The basic Bayesian Poisson flow model

In case that the GitHub links do not work in the future, we include the main codes in Stan to reproduce the study. First we present the Stan code to implement the model in Chapter 5.

```
data {
  // number of strata pairs
  int<lower=1> L;
  // transmission count for a strata pair
  int n[L];
  // number of sampling categories
  int<lower=1> N_xi;
  // parameters for the density of sampling fractions
  matrix[N_xi,2] shape;
  // indices of strata for sources
  int xi_id_src[L];
  // indices of strata for recipients
  int xi_id_rec[L];
}

transformed data {
  // parameters of the gamma prior
  real alpha = 0.8/L;
}
```

```

parameters {
  // rates of the Poisson likelihood
  vector<lower=0>[L] lambda;
  // sampling fractions
  vector<lower=0,upper=1>[N_xi] xi;
}

transformed parameters {
  // pairwise sampling fractions
  vector[L] xi_pair;
  // expected actual transmission count for a pair
  vector[L] betav;
  // scale parameter of the gamma prior
  real beta;
  for (i in 1:L){
    xi_pair[i] = xi[xi_id_src[i]] * xi[xi_id_rec[i]];
    if (n[i]==0){
      betav[i] = 1/xi_pair[i]-1;}
    else{
      betav[i] = n[i]/xi_pair[i];}
  }
  beta = 0.8/sum(betav);
}

model {
  for (i in 1:N_xi){
    xi[i]~ beta(shape[i,1],shape[i,2]);
  }
  for (i in 1:L){
    lambda[i] ~ gamma(alpha,beta);
    n[i] ~ poisson(lambda[i] * xi_pair[i]);
  }
}

```

E.3 The semi-parametric Bayesian Poisson flow model

We then present the Stan code to implement the model in Chapter 6. Here we stratify the population by one-year age bands and genders and assume the varying transmissions and baselines by age across the genders. Ideally, we would like to perform the Gaussian process model in Appendix E.3.1, but it is not computationally feasible for the high resolution age bands. Instead, we use the approximated Gaussian process in Appendix E.3.2.

E.3.1 Gaussian process

```
// kernel
functions {
  matrix L_cov_exp_quad_ARD(vector[] x,
                             real alpha,
                             vector rho) {

    int N = size(x);
    matrix[N, N] K;
    real sq_alpha = square(alpha);
    for (i in 1:(N-1)) {
      K[i, i] = sq_alpha + 1e-10;
      for (j in (i + 1):N) {
        K[i, j] = sq_alpha
                  * exp(-0.5 * dot_self((x[i] - x[j]) ./ rho));
        K[j, i] = K[i, j];
      }
    }
    K[N, N] = sq_alpha + 1e-10;
    return cholesky_decompose(K);
  }
}

data {
  // male to female and female to male
  int<lower=1> L_group;
  // number of age group pairs
  int<lower=1> L_per_group;
  // dimension of inputs
```

```

int<lower=1> D;
// age of sources and recipients
vector[D] x[L_per_group];
// transmission count for each strata pair
int n[L_per_group, L_group];
// number of sampling categories
int<lower=1> N_xi;
// parameters for the density of sampling fractions
matrix[N_xi,2] shape;
// indices of strata for sources
int xi_id_src[L_per_group,L_group];
// indices of strata for recipients
int xi_id_rec[L_per_group,L_group];
// indices of the male to female pair
int<lower=0> id_mf;
// indices of the female to male pair
int<lower=0> id_fm;
// parameters for the inverse gamma prior
real d1_invg_par1_mf;
real d1_invg_par2_mf;
real d2_invg_par1_mf;
real d2_invg_par2_mf;
real d1_invg_par1_fm;
real d1_invg_par2_fm;
real d2_invg_par1_fm;
real d2_invg_par2_fm;
}
parameters {
// length scale
vector<lower=0>[D] rho[L_group];
// marginal standard deviation
vector<lower=0>[L_group] alpha;
// differences in baselines between male-to-female and female-to-male
real mu_mf;
// baseline of female-to-male

```

```

    real mu;
    // iid normal
    vector[L_per_group] beta[L_group];
    // sampling fractions
    vector<lower=0,upper=1>[N_xi] xi;
}

transformed parameters {
    // pairwise sampling fractions
    vector[L_per_group] lxi_pair[L_group];
    // log rates of the Poisson likelihood
    vector[L_per_group] llbd[L_group];
    # cholesky decomposition of covariance matrices
    matrix[L_per_group, L_per_group] L_cov[L_group];
    // baselines for male-to-female and female-to-male
    vector[L_group] muv;
    // GP evaluations
    vector[L_per_group] f[L_group];
    muv = rep_vector(0,L_group);
    muv[id_mf] = muv[id_mf] + mu_mf;
    for (i in 1:L_group){
        L_cov[i] = L_cov_exp_quad_ARD(x,alpha[i],rho[i]);
        f[i] = L_cov[i] * beta[i];
        lxi_pair[i] = log(xi[xi_id_src[,i]]) + log(xi[xi_id_rec[,i]]);
        llbd[i] = mu + muv[i] + f[i] + lxi_pair[i];
    }
}

model {
    for (i in 1:N_xi){
        xi[i] ~ beta(shape[i,1],shape[i,2]);
    }
    for (i in 1:L_group){
        beta[i] ~ normal(0,1);
    }
    rho[1][1] ~ inv_gamma(d1_invg_par1_mf, d1_invg_par2_mf);
    rho[1][2] ~ inv_gamma(d2_invg_par1_mf, d2_invg_par2_mf);
}

```

```

rho[2][1] ~ inv_gamma(d1_invg_par1_fm, d1_invg_par2_fm);
rho[2][2] ~ inv_gamma(d2_invg_par1_fm, d2_invg_par2_fm);
alpha ~ normal(0, 10);
mu_mf ~ normal(0, 10);
mu ~ normal(0, 10);
for (i in 1:L_group){
  n[,i] ~ poisson_log(llbd[i]);
}
}

```

E.3.2 Hilbert space Gaussian process approximation

```

//spectral density
functions {
  vector spd_cov_exp_quad(vector[] x, real alpha, vector rho) {
    int L_per_group = dims(x)[1];
    int D = dims(x)[2];
    int Drho = rows(rho);
    vector[L_per_group] out;
    if (Drho == 1) {
      real constant = square(alpha) * (sqrt(2 * pi()) * rho[1])^D;
      real neg_half_rho2 = -0.5 * square(rho[1]);
      for (m in 1:L_per_group) {
        out[m] = constant * exp(neg_half_rho2 * dot_self(x[m]));
      }
    } else {
      real constant = square(alpha) * sqrt(2 * pi())^D * prod(rho);
      vector[Drho] neg_half_rho2 = -0.5 * square(rho);
      for (m in 1:L_per_group) {
        out[m] = constant * exp(dot_product(neg_half_rho2, square(x[m])));
      }
    }
    return out;
  }
}
vector gpa_coefficient(int M_nD, real alpha, vector rho,

```

```

vector beta, vector[] slambda) {
    vector[M_nD] diag_spd = sqrt(spd_cov_exp_quad(slamba,
        alpha, rho));
    return diag_spd .* beta;
}
}

data {
    // male to female and female to male
    int<lower=1> L_group;
    // number of age group pairs
    int<lower=1> L_per_group;
    // number of strata pairs
    int<lower=1> L;
    // dimension of inputs
    int<lower=1> D;
    // age of sources and recipients
    matrix[L_per_group,D] x;
    // transmission count for each strata pair
    int n[L_per_group, L_group];
    // number of basis functions
    int<lower=1> M_nD;
    // eigenvalues
    matrix[L_per_group, M_nD] Xgp;
    // eigenfunctions
    vector[D] slambda[M_nD];
    // number of sampling categories
    int<lower=1> N_xi;
    // parameters for the density of sampling fractions
    matrix[N_xi,2] shape;
    // indices of strata for sources
    int xi_id_src[L_per_group,L_group];
    // indices of strata for recipients
    int xi_id_rec[L_per_group,L_group];
    // indices of the male to female pair
    int<lower=0> id_mf;

```

```

// indices of the female to male pair
int<lower=0> id_fm;
// parameters for the inverse gamma prior
real d1_invg_par1_mf;
real d1_invg_par2_mf;
real d2_invg_par1_mf;
real d2_invg_par2_mf;
real d1_invg_par1_fm;
real d1_invg_par2_fm;
real d2_invg_par1_fm;
real d2_invg_par2_fm;
}
parameters {
  // length scale
  vector<lower=0>[D] rho[L_group];
  // marginal standard deviation
  vector<lower=0>[L_group] alpha;
  // differences in baselines between male-to-female and female-to-male
  real mu_mf;
  // baseline of female-to-male
  real mu;
  // iid normal
  vector[M_nD] beta[L_group];
  // sampling fractions
  vector<lower=0,upper=1>[N_xi] xi;
}
transformed parameters {
  // pairwise sampling fraction for each strata pair
  vector[L_per_group] lxi_pair[L_group];
  // intermediate terms to evaluate function values
  vector[M_nD] coefficient[L_group];
  // log rates of the Poisson likelihood
  vector[L_per_group] llbd[L_group];
  // baselines for male-to-female and female-to-male
  vector[L_group] muv;

```

```

// male-to-female GP evaluations
vector[L_per_group] f_mf;
// female-to-male GP evaluations
vector[L_per_group] f_fm;
// GP evaluations
vector[L_per_group] f[L_group];
muv = rep_vector(0,L_group);
muv[id_mf] = muv[id_mf] + mu_mf;
coefficient[1] = gpa_coefficient(M_nD, alpha[1], rho[1], beta[1], slambda);
f_mf = Xgp * coefficient[1];
coefficient[2] = gpa_coefficient(M_nD, alpha[2], rho[2], beta[2], slambda);
f_fm = Xgp * coefficient[2];
f[id_mf]=f_mf;
f[id_fm]=f_fm;
for (i in 1:L_group){
  lxi_pair[i] = log(xi[xi_id_src[,i]]) + log(xi[xi_id_rec[,i]]);
  llbd[i] = mu + muv[i] + f[i] + lxi_pair[i];
}
}
model {
  for (i in 1:N_xi){
    xi[i] ~ beta(shape[i,1],shape[i,2]);
  }
  for (i in 1:2){
    beta[i] ~ normal(0,1);
  }
  rho[1][1] ~ inv_gamma(d1_invg_par1_mf, d1_invg_par2_mf);
  rho[1][2] ~ inv_gamma(d2_invg_par1_mf, d2_invg_par2_mf);
  rho[2][1] ~ inv_gamma(d1_invg_par1_fm, d1_invg_par2_fm);
  rho[2][2] ~ inv_gamma(d2_invg_par1_fm, d2_invg_par2_fm);
  alpha ~ normal(0, 10);
  mu_mf ~ normal(0, 10);
  mu ~ normal(0, 10);
  for (i in 1:L_group){
    n[,i] ~ poisson_log(llbd[i]);
  }
}

```

}
}