Imperial College London

Imperial College London
Department of Mathematics

Bayesian reconstruction of epidemic transmission chains from viral deep-sequence data

Mélodie Monod

CID: 01526205

Supervised by Dr Oliver Ratmann

September 5, 2019

Submitted in partial fulfilment of the requirements for the MSc in Statistics of Imperial College London

The work contained in this thesis is my own work unless otherwise stated.

Signed: Date:

Abstract

Transmission chains are the fundamental building blocks of the dynamics of any infections disease. Estimating the time of infection, pairwise linkage, and direction of transmission give a unique insight into the disease dynamic, including likely receiver and spreader groups within a host population. It may inform public health policy and programs of prevention and treatment. We investigate the HIV-1 epidemic using deep-sequence data collected in a large population-based sample of infected individuals in Rakai District, Uganda. These data provide information not only on how closely related individuals are epidemiologically, but also in which direction transmission occurred. To harness this new type of data, we develop a Bayesian model for inference on the transmission chain and use an MCMC to test our approach on small networks involving two to five individuals. Of the reconstructed transmission chains, male were more likely to spread the infection than female. In addition, the younger a person was at time of infection, the more likely he was to transmit the virus.

Acknowledgements

I would like to express my sincere gratitude to my patient and supportive supervisor, Dr Oliver Ratmann, whose insight and knowledge into the subject matter steered me through this research. I would also like to thank Dr Matthew Hall for his participation and engagement in this research. Lastly, I am grateful to Dr Pierre-Marie Grollemund and Mr Wei Pan for their consistent support during the running of this project.

Contents

1 Description of symbols

This section provides the reader with a reference for the notation and description of all variables and parameters used in the thesis.

Symbol	Description
Genetic data, $\boldsymbol{x}$	
D_{ijw}	mean tip-to-tip patristic distance between individuals i and j in window w
R_{ijw}	phylogenetic relationship between individuals i and j in window w
Variables of the transmission network, $\boldsymbol{T}$	
S_i	Source that infected the i-th individual
t_i^I	Day of infection of individual i
Set of Genetic parameters, $\boldsymbol{\psi}$	
μ	Substitution rate per subst/site/year
$\boldsymbol{\pi}$	Set of probabilities of phylogenetic relationship
$\alpha, \beta, \delta, \tilde{\alpha}, \tilde{\beta}, \tilde{\delta}$	Regression coefficients
$\sigma, \sigma_\delta, \tilde{\sigma}_\delta$	Regression variances
Set of Epidemiological parameters, $\boldsymbol{\theta}$	
θ_0	Baseline transmission rate
Other	
$\boldsymbol{F}$	Flow matrix
$W(i,j)$	Set of windows on which phylogenetic data was available for pair ij
t_i^S	Day of sampling of individual i's pathogen
t_{ij}^E	Time elapsed between individual i and j
w(.)	Generation time function
$\lambda_{i \to j}$	Transmission hazard from case i to susceptible j

Table 1: Description of symbols used in the thesis.

2 Introduction

2.1 HIV-1 epidemic in Africa

Human Immunodeficiency Virus (HIV) is an infectious disease attacking the human immune system. Infectious diseases are disorders caused by organisms — such as bacteria, fungi, parasites, and, as in the case of HIV, viruses. The virus is transmitted through direct contact with HIV-infected body fluids, such as blood, semen, and vaginal fluids. It might also be passed from a mother to her child during pregnancy, delivery, or breastfeeding (Rom and Markowitz, 2007). Without treatment, the number of cells fighting infection declines as the number of HIV copies (viral load) increases. HIV infection eventually progresses to AIDS and cause the death of the infected individual (Lui et al., 1988, UNAIDS and WHO, 2007).

The first HIV-1 case in Africa was reported in 1978–79 in Uganda (Nahmias et al., 1986, Buvé et al., 2002). By 1986, it was clear that HIV-1 had spread in the populations of many countries in sub-Saharan Africa and was posing a significant public health problem (Quinn et al., 1986). The introduction of highly effective ART in the West, together with the creation of the Joint United Nations Programme on HIV/AIDS (UNAIDS) in 1996 established a favorable environment to political action on AIDS. In 1997, the World Bank Multi-Country AIDS Program for Africa started to raise new funding to fight against AIDS (Piot et al., 2007).

Nowadays, Africa remains the continent most severely affected by the global HIV burden (including HIV-1 and HIV-2). The rates of infectiousness in Africa are heterogeneous with East and Southern Africa general severely affected than West and Central Africa, as shown in Figure 1 for the 15-49 year-olds population in 2017. According to a dataset from UNAIDS sourced from the World Bank (WorldBank, 2017), in 2017, HIV prevalence among adults aged 15 to 49 was 22.8% in Botswana, 23.8% in Lesotho, 18.8% in South Africa and 5.9% in Uganda.

2.2 HIV-1 epidemic in Rakai District, Uganda

The data used in this thesis are from Rakai District in Uganda. This is a high HIV-1 prevalence region (approximately 40%, according to Chang et al. (2016)), composed of agrarian, trading, and fishing communities along Lake Victoria. This region is indicated in Figure 1 by a red circle.

2.3 Overall motivation

The range of HIV prevalence in Africa indicates that there does not exist a single global HIV epidemic. Differences in prevalence among site and age group suggest that different age groups in different geographical areas need specific interventions. UNAIDS (2010) recommended prevention programs following a 'know your epidemic, know your response' approach. Before deciding on how to target a specific population, an evidence-informed picture of their epidemic is needed. This involves examining factors such as modes of HIV transmission, key affected populations, and critical epidemiological trends. In particular, Wilson and Halperin (2008) highlighted three main questions that need to be addressed in this context: 'How to reach vulnerable groups within the target population?', 'How to change fundamental community

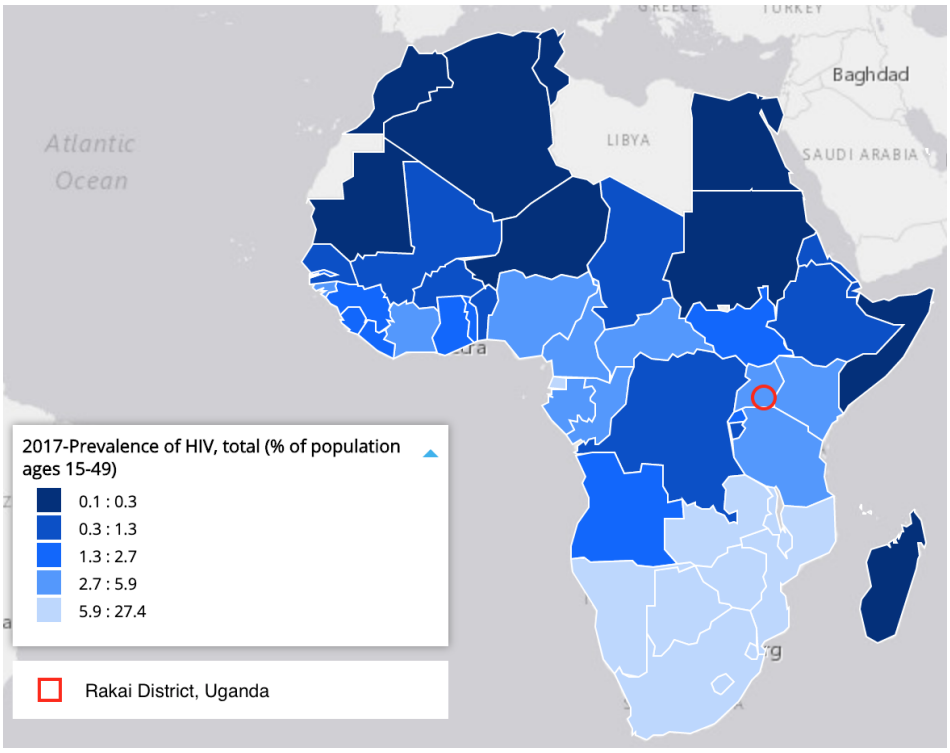


Figure 1: HIV prevalence rate (including HIV-1 and HIV-2) for adults aged 15 to 49 in 2017 in Africa. Data from WorldBank (2017). Prevalence of HIV refers to the percentage of people ages 15-49 who are infected with HIV.

norms?' and 'How to more accurately estimate the relative proportion of infections from different transmission sources?'. From an applied perspective, this thesis focuses on the last of these questions for the HIV-1 epidemic.

2.4 Definition of HIV-1 transmission chains

In order to better understand how HIV-1 infection spreads, we focus in this thesis on the reconstruction of HIV-1 transmission chains from viral deep-sequence data.

A *transmission chain* illustrates the spread of a disease in a subset of the population over time. Mathematically, we define a transmission chain $\boldsymbol{T}$ as follows. We consider individuals $i = 1, ..., N$. The chain of transmission events is encoded in a $N \times N$ *binary flow* matrix $\boldsymbol{F}$ where rows denote source individuals, columns denote recipient individuals, and a value of 1 in cell ij means transmission from i to j. There are particular constraints on the flow matrix. First, an individual can be infected only once, so that there is only one non-zero entry per column. This is referred to as an *indegree* equals to one. Second, an individual can infect none or more than one individuals, so that the number of non-zero entry on each row is between 0 and $N - 1$. This is called the *outdegree*. In this thesis, we assume that all individuals of a chain were sampled and that all individuals in a host-population are connected.

For convenience, we do not consider the binary flow matrix, and instead examine the *vector of source cases*. Suppose individuals are ordered in a known sequence in the rows and

columns, the source vector $\boldsymbol{S}$, of dimension N, encapsulates the corresponding source of every individual. The source of individual j is found by

$$S_j = \underset{i}{\operatorname{argmax}}\, \boldsymbol{F}_{(i,j)}. \tag{2–1}$$

There is a bijection between the row with a positive entry on each column of the flow matrix and the source vector, because the indegree is equal to 1. It is thus equivalent to investigate one or the other. The *index case* is the first individual infected in the host population. He is assumed to have been infected by an external source, such that his entry in $\boldsymbol{S}$ is set to 0. In addition, we include in $\boldsymbol{T}$ *the infection times* of every individual $\boldsymbol{t}^I = \{t_1^I, ..., t_N^I\}$. Thus,

$$\boldsymbol{T} = \{\{t_1^I, ..., t_N^I\}, \{S_1, ..., S_N\}\}. \tag{2–2}$$

Figure 4.A. presents an example of a transmission history involving four individuals. Suppose we order the individuals as {F1, F2, M1, M2}, then $\boldsymbol{S} = \{0, \text{M1}, \text{F1}, \text{F1}\}$.

2.5 Estimating HIV-1 transmission chains

To infer transmission chains, we adopt a Bayesian approach using viral deep-sequence data denoted by $\boldsymbol{x}$. Epidemiologic and phylogenetic model parameters are referred to as $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ respectively. The information contained in those three terms will be made precise later; inference will take place via the joint posterior distribution

$$\begin{aligned} p(\boldsymbol{T}, \boldsymbol{\theta}, \boldsymbol{\psi} | \boldsymbol{x}) &\propto p(\boldsymbol{x} | \boldsymbol{T}, \boldsymbol{\theta}, \boldsymbol{\psi}) p(\boldsymbol{T}, \boldsymbol{\theta}, \boldsymbol{\psi}) \\ &= p(\boldsymbol{x} | \boldsymbol{T}, \boldsymbol{\psi}) p(\boldsymbol{T} | \boldsymbol{\theta}) p(\boldsymbol{\theta}) p(\boldsymbol{\psi}). \end{aligned} \tag{2–3}$$

Considering estimated transmission chains in aggregate will then allow us to go back to our epidemiological motivation, and to obtain insights on the age, geography, gender categories of source populations that disproportionately pass on infections.

2.6 Thesis structure

The literature review in Section 3 presents methods to reconstruct transmission chains from epidemiological data only, and then move to the simultaneous use of epidemiological and phylogenetic data. Section 4 introduces the phylogenetic and epidemiological data used in this thesis, for model construction and inference. Components of the Bayesian model, namely the likelihoods and the prior are developed in Section 5. To sample from the posterior space, we construct a Markov-Chain Monte Carlo algorithm and assess its performance in Section 6. We describe the inferred transmission chains among small subsets of individuals in Section 7. Finally, we discuss methods and findings in Section 8.

3 Literature Review

We present in this section some of the relevant research around the components of the posterior distribution (2–3). In Subsection 3.1 focuses on time-dynamic models that explain disease spread - used to define the prior distribution on transmission chains given epidemiologic parameters, $p(\boldsymbol{T}|\boldsymbol{\theta})$. In Subsection 3.2, we discuss different data types $\boldsymbol{x}$ considered for inference. Finally some likelihood forms $p(\boldsymbol{x}|\boldsymbol{T}, \boldsymbol{\psi})$ are specified in in Subsection 3.3.

3.1 Transmission model

A major component of transmission chain estimation is the description of the infection process with a *transmission model*, also called epidemic model, to find a form for $p(\boldsymbol{T}|\boldsymbol{\theta})$.

In the *compartmental model*, each individual in the population passes through states, in order, during the observation period. Individuals at the same state form a group. Disease dynamics are generated by the flux of individuals from one group to another. The most famous compartmental model is called the *epidemic SIR* model (Becker and Britton, 1999) in which members of the population are classified in one of three states: susceptible, infected or recovered. We denote the number of susceptible members at time t by $S(t)$, infected by $I(t)$ and recovered by $R(t)$ and,

$$S(t) + I(t) + R(t) = N, \tag{3–1}$$

at any point in time. Suppose, susceptible individuals become infected with *transmission rate* β then recover (or die) with *recovery rate* γ. The disease dynamics are expressed by

$$\frac{dS(t)}{dt} = -\frac{\beta I(t) S(t)}{N}, \tag{3–2}$$

$$\frac{dI(t)}{dt} = \frac{\beta I(t) S(t)}{N} - \gamma I(t), \tag{3–3}$$

$$\frac{dR(t)}{dt} = \gamma I(t). \tag{3–4}$$

The *Chain-Binomial model* is a stochastic model used for the propagation of SIR diseases in closed populations in discrete time. Suppose the probability of infection is p at all times for all individuals and the recovery period is fixed to one, then,

$$I(t+1) \sim \text{Binomial}(S(t), p), \tag{3–5}$$

$$S(t+1) = S(t) - I(t+1), \tag{3–6}$$

$$R(t+1) = R(t) + I(t). \tag{3–7}$$

This is the *Greenwood model* (Greenwood, 1931) where

$$p = 1 - exp(-\beta I(t)). \tag{3–8}$$

The transmission chain does not include information of who-infected-who (i.e.,the source

vector) and the probability of particular times of infection is given by,

$$P(\boldsymbol{T}|\beta) = \prod_{t=0}^{T} \binom{S(t)}{I(t+1)} p^{I(t+1)}(1-p)^{S(t)-I(t+1)}, \tag{3–9}$$

$$S(t) = \sum_{j:t_j^I<t} 1, \tag{3–10}$$

$$I(t+1) = \sum_{j:t<t_j^I<t+1} 1. \tag{3–11}$$

A more complex Chain-Binomial is the *Reed–Frost model* which assumes that p depends on the time of infection, such that the probability of infection at time t is $p(t)$ (Abbey, 1952, Allen, 2008). The *stochastic epidemic model* is a stochastic model that is related to the SIR in continuous-time model, with γ a random variable. The model can be defined in terms of Markov transition rates, so in particular the transition to the state $(S(t)-1, I(t)+1, R(t))$ happens at rate $\beta S(t)I(t)$ and transition to the state $(S(t), I(t)-1, R(t)+1)$ happens at rate $\gamma I(t)$. Infections occur at times $t_1^I \leq ... \leq t_N^I$ and removals occur at times $r_1 \leq ... \leq r_N$ included in $\boldsymbol{T}$. Again, the transmission chain does not include information on transmission linkage at individuals level. Assuming that the time distribution to the next event is exponential with mean γ^{-1},

$$P(\boldsymbol{T}|\beta,\gamma) = \Big(\prod_{j=2}^{N} \beta I(t_j^I-)S(t_j^I-)\Big)\Big(\prod_{i=1}^{N} \gamma I(t_j^I-)\Big) \times exp\Big(-\int_{t_1^I}^{r_1} \beta I(t)S(t) + \gamma I(t)dt\Big), \tag{3–12}$$

where $S(t-) = \lim_{u\uparrow t} S(u)$ (O'Neill and Roberts, 1999, Allen, 2008, 2017). There exist many other compartmental models that include other disease compartments. For example, the *SIS* (Susceptible, Infectious, Susceptible) include only two groups as immunization is not given upon infection. Another, for disease with a significant incubation period (the period between infection and being contagious), is the *SEIR* (Susceptible, Exposed, Infected, Recovered). During this period individual is in compartment E (for exposed), with

$$S(t) + E(t) + I(t) + R(t) = N. \tag{3–13}$$

In this case a third epidemiological parameter σ enters $\boldsymbol{\theta}$, where $1/\sigma$ is the average period of incubation. Examples of applications include Cauchemez et al. (2004), Didelot et al. (2014) that uses a SIR to reconstruct an influenza and tuberculosis outbreak, a SIS was used in Cauchemez et al. (2006) to model an S. pneumonia outbreak and a SEIR in Cori et al. (2009) to investigate a SARS outbreak. O'Neill and Roberts (1999) compared results of the Reed-Frost model and a general stochastic epidemic model on a measles propagation.

Jombart et al. (2014) estimated simultaneously the times of infection and the pairwise transmission linkages. The authors assumed that the time of infection of the recipients depended only on the one of their source. They introduced a *generation time function* that indicates the probability of infectiousness after a certain period since infection. Moreover, the

authors allowed for unsampled individuals by introducing a proportion of sampled cases π. To this aim, in addition of the source vector and times of infection, authors incorporated in $\boldsymbol{T}$ a generation vector $\boldsymbol{\kappa}$. Where κ_i denotes the number of generation between i and S_i (i.e. S_i is the sampled individual who was inferred to have transmitted to i defined in (2–1) - his source). The latter being dependent of π, such that if $\pi = 1$, then $\kappa_i = 1$, $\forall i$. By applying the chain rule ordered by increasing time of infection,

$$p(\boldsymbol{T}|\pi) = p(\boldsymbol{t}^I, \boldsymbol{\kappa}, \boldsymbol{S}|\pi) \tag{3–14}$$

$$= \prod_{j=2}^{N} p(t_j^I|S_j, t_{S_j}^I, \kappa_j)p(\kappa_j|\pi)p(S_j)p(t_1^I, \kappa_1, S_1) \tag{3–15}$$

$$= \prod_{j=2}^{N} \mathrm{w}(t_i^I - t_{S_i}^I) \times \mathrm{NB}(1|\kappa_i - 1, \pi) \times p(S_j)p(t_1^I, \kappa_1, S_1), \tag{3–16}$$

for which they assumed that all cases were independent conditional on their ancestries. w(.) is the generation time function, and $\mathrm{NB}(r, p)$ is the negative binomial distribution with a number of failures r and a probability of success p. The first part quantifies the probability of infectiousness of S_i after period $t_i^I - t_{S_i}^I$ since infection. The second part gives the probability of unobserved intermediate cases, modeled with a negative binomial distribution, indicating the probability of obtaining one 'success' (here, sampling a case) after r 'failures' (unobserved cases) with a probability of success π. The third part is assumed to stay constant.

Finally, in the *Cox-type hazard model*, the force of infection depends on specific covariates of the source and recipient in time. In particular, the force of infection can be modeled by the transmission hazard

$$\lambda_{i\to j}(t|\boldsymbol{\theta}) = \mathrm{f}(X_{it}, X_{jt}|\boldsymbol{\theta}), \tag{3–17}$$

for case i and susceptible j on day t, as a function of time-varying covariates of both individuals, X_{it} and X_{jt}, and epidemiological parameters $\boldsymbol{\theta}$. The specific effect of X_{it} and X_{jt} on the transmission hazard is measured by function f(.). The later can, for instance, take the form of a Cox proportional hazards,

$$\mathrm{f}(X_{it}, X_{jt}|\boldsymbol{\theta}) = \exp(\theta_0 + \boldsymbol{\theta}_1^T X_{it} + \boldsymbol{\theta}_2^T X_{jt}) \quad > 0. \tag{3–18}$$

Cauchemez et al. (2011) and Cauchemez and Ferguson (2012) used this approach and formulated the probability of infection with a Cox-type survival model. We will use this approach in the thesis and derive several key equations later.

3.2 From epidemiological to phylogenetic data

Most methods for inferring transmission chains were inspired and developed with epidemiological data collected during an infectious disease outbreak. Depending on the application in question, models may incorporate latent (incubation) periods, removal times, diagnosis times and range of individual-level meta-data such as gender, age, school class, seating arrangement, etc., of observed infected individuals (O'Neill et al., 2000). These epidemiological data are encapsulated in $\boldsymbol{x}$. Likelihoods for these data are not presented in this review.

Two limitations motivated the use of other data. Firstly, the data collected on epidemics consists of surveillance data and tend to be of poor quality, usually suffering from severe under-reporting or rates of reporting that vary over time. Secondly, the observational time handled has to be short to make accurate and robust reconstruction of the transmission history. Indeed, with long generation times, epidemiological data alone are consistent with many different scenarios of who-infected-who. For instance, in the following applications that used only epidemiological data, the authors observed short outbreaks. O'Neill et al. (2000), Cauchemez et al. (2004, 2011) studied household transmission in a period of one year, fifteen days an two months respectively; in Cauchemez et al. (2006), an epidemic among schoolchildren was investigated within a year; in Cori et al. (2009) healthcare practitioners were susceptible under investigation for a period of one year.

The genetic sequence of viruses sampled from infected individuals could provide interesting insights into transmission chain reconstruction. The structure and organization of the genes in the HIV-1 genome are presented in Figure 2. If the virus alternates its nucleotide sequence, we say that it mutates. After applying a phylogenetic analysis, one can identify which pathogens are genetically closer to others. The branching pattern of a phylogenetic tree reflects the history of sampled pathogens evolution (i.e., mutation). The leaves correspond to sampled pathogens. The root node of a phylogenetic tree does not have a parent node and corresponds to the most recent common ancestor of all sampled pathogens. Each node represents a specific event in evolution that separated two sampled pathogens. The length of each branch between one node to the next represents the degree of changes that occurred until the next separation. For many pathogens, including HIV-1, evolutionary processes occur on the same time scale as epidemiological processes (Holmes et al., 1995, Pybus and Rambaut, 2009) so that the phylogenetic information can be used to reconstruct aspects of disease spread. Lam et al. (2010) describes some of the methods available to reconstruct phylogenies from virus sequences.

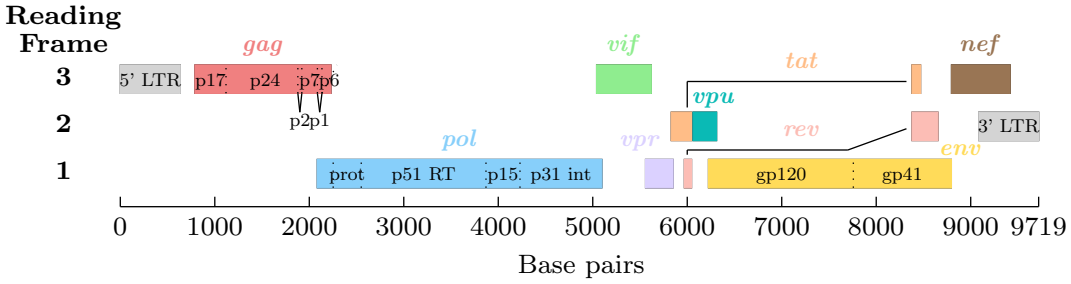


Figure 2: Structure of the RNA genome of HIV-1. The HIV-1 genome contains nine genes and genes product (*gag*, *pol*, *env*, *rev*, *vif*, *vpr*, *vpu*, and *nef*) that encode fifteen viral proteins (Li et al., 2015). Each gene is indicated in color and represented with a rectangle. Long terminal repeats (LTRs) are identical sequences repeated hundreds of times (Allaby, 2012).

3.3 Phylogenetic likelihood models

Several models have been proposed to quantify the probability of genetic or phylogenetic data given a particular transmission chain $\boldsymbol{T}$.

Morelli et al. (2012) described the probability for observing a certain number M of mutations between two sequences from two individuals in a transmission chain, with the idea that individuals with short time distance in a transmission chain should have viruses with few mutations between them. Specifically, they used a *substitution model* to find the probability to observe a number of mutation M given time interval Δ and substitution rate μ per nucleotide per day. Let n_x be the fixed length of sampled sequences, then

$$M|\Delta,\mu \sim \text{Binomial}\big(n_x, \frac{3}{4}1 - exp(-\frac{4}{3}\mu\Delta)\big), \tag{3–19}$$

where Δ is the evolutionary time separating the observation of the pathogen of i and the pathogen of j. It be computed along the transmission chain between pairwise individuals by,

$$\Delta = (t_i^S - t_{div(i,j)}^I) + (t_j^S - t_{div(i,j)}^I), \tag{3–20}$$

where t_i^S is the sampling time of i and $t_{div(i,j)}^I$ denotes the infection time at which the transmission linkages leading to the infection of i and those leading to the infection of j diverged. Here, the *Jukes-Cantor's correction* (Jukes and Cantor, 1969) is used to find the probabilities of a mutation per site per day. The likelihood given the transmission chain was specified by,

$$p(\boldsymbol{x}|\boldsymbol{T},\mu) = \prod_{i=2}^{N}\prod_{j-1}^{i-1} p(M(x_i,x_j)|\Delta = (t_i^S - t_{div(i,j)}^I) + (t_j^S - t_{div(i,j)}^I)) \tag{3–21}$$

Jombart et al. (2014) followed a similar approach, but first introducing the possibility that some individuals were unsampled, and second by considering only likelihood terms among pairs identified as directly linked (i.e., one transmitted to the other) in the transmission chain. The genetic likelihood of case j was defined as the probability of observing the mutations between the sequence x_j and the ancestral sequence x_{S_j} with j and S_j being separated by κ_j generations,

$$p(\boldsymbol{x}|\boldsymbol{T},\mu) = \prod_{j=2}^{N} \mu^{M(x_j,x_{S_j})}(1-\mu)^{(\kappa_j \times n_x) - M(x_j,x_{S_j})}. \tag{3–22}$$

Didelot et al. (2014) took into account that a single individual might transmit to his recipient a viral lineage from a different branch than the one that was sampled in the source individuals, which can frequently occur as HIV-1-infected individuals harbor a large amount of viral diversity (Lemey et al., 2007). The authors introduced a within-host evolutionary parameter $N_e g$, with constant population size N_e and average generation time g. For individual j, authors encapsulated in x_j the genealogy of the phylogenetic tree involving only individual j's branches. This sub-tree corresponds to the within-host evolution of j's pathogen. It possesses n_j leaves, equals to one plus the number of individuals infected by i. The time attributed to them corresponds to the time of infection of j and of j's recipients. The proba-

bility of the timed genealogy x_j given the dates of its leaves and the within-host evolutionary parameter is,

$$p(\boldsymbol{x}|\boldsymbol{T}, N_e g) = \prod_{j=1}^{N} \prod_{k=2}^{n_j} \frac{exp(-A_k/N_e g)}{N_e g(1 - exp(-B_k/N_e g))}, \tag{3–23}$$

where the n_j leaves are in increasing order of age. j is the ancestor of all individuals in $n_j - 1$ leafs such that the time when $k = 1$ is t_j^I. From x_j, two measures of evolutionary distance (sum of branches) are computed iteratively. A_k denotes the evolutionary distance between the time of leaf k and the time where it coalesces with an ancestor of a previously considered leaf (t_j^I if it is the first infected recipient). Moreover, B_j denotes the sum of branch lengths between the time of leaf j and the time t_j^I.

While all presented models used a measure of evolutionary distance of the pathogen to explain the position on the transmission chain, Romero-Severson et al. (2016), Leitner and Romero-Severson (2018) found that the phylogenetic topology derived from many sequences per individuals could explain where individuals are in the transmission chain and, in particular, what the direction of transmission is. These findings motivate new research that combines both evolutionary distances and phylogenetic topology as evidence for reconstructing transmission chains. Viral deep-sequencing data captures multiple sequences per individual in a single sample, and so these data promise a potentially powerful approach to molecular epidemiology (Wymant et al., 2017).

This thesis develops a novel approach for reconstructing transmission chains from deep-sequence data. In Wymant et al. (2017) and Ratmann et al. (2019), similar outputs have already been utilized to make inference on transmission chains. The procedure required using a threshold to define close and distant individuals. Here, we wish to avoid such thresholds and furthermore harness the information in the actual value of the outputs. This motivates the change of approach.

4 Data for reconstructing transmission chains

4.1 Study population and deep-sequencing

The 'Phylogenetics and Networks for generalized HIV Epidemics in Africa' consortium (PANGEA-HIV) constructed extensive deep-sequence data set of HIV-1 on four cohort sites. In this thesis, we analyze data from the Rakai Community Cohort Study (RCCS), where pathogens of $2,652$ individuals, aged 15-49 years, ART-naive before sampling, were deep-sequenced.

4.2 Structure of the deep-sequence phylogenetic data

To treat deep-sequence phylogenetic data we use the Phyloscanner tool (Wymant et al., 2017). Phyloscanner takes as inputs deep-sequence RNA fragments (*reads*) from infected individuals in the target population and reconstructs phylogenies from reads alignments that slide along the HIV-1 genome (Wymant et al., 2017) as shown in Figure 3.A.B.C. For the analysis as specified for the Rakai data, there are for all individuals 335 windows across the genome. On every one of which, a phylogeny was inferred.

Two essential statistics are extracted from the reconstructed deep-sequence phylogenies and used in this thesis. For each pair of individuals i and j, we consider in each phylogeny

- the *phylogenetic relationships* of their viral reads in the phylogeny, as well as
- the *mean tip-to-tip patristic distances* between viral reads from i and those from j.

Topological relationships and distances are defined as follows; see also Table 2 and Figure 3 for illustrations. Consider a viral phylogeny, colored through ancestral state reconstruction. The following statistics are calculated at the pair level. We say that i and j are *adjacent*, ($A_{ij} = 1$) if the shortest path between at least one subgraph u from i and v from j is not attributed to any sampled individual other than i and j; and not adjacent ($A_{ij} = 0$) otherwise. Next, we define the *number of paths* from i to j (P_{ij}) as the number of subgraphs from j which have as ancestor a subgraph from i. The two statistics are used to define the following phylogenetic relationships between any pairs of individuals i and j in any one phylogeny. We say that pathogen of i is *ancestral* to pathogen of j if i and j are adjacent and there are only paths from i to j, i.e.

$$R^{1\to 2} = \begin{cases} 1 & \text{if } A_{ij} = 1,\ P_{ij} \geq 1 \text{ and } P_{ji} = 0 \\ 0 & \text{otherwise} \end{cases}. \tag{4–1}$$

And in the other direction,

$$R^{2\to 1} = \begin{cases} 1 & \text{if } A_{ij} = 1,\ P_{ij} = 0 \text{ and } P_{ji} \geq 1 \\ 0 & \text{otherwise} \end{cases}. \tag{4–2}$$

If the pathogen of i and the one j are adjacent and there exist a subgraph where i is ancestral

to j and others where j is ancestral i, we say that pathogens of i and j are *intermingled*,

$$R^{1\leftrightarrow 2} = \begin{cases} 1 & \text{if } A_{ij} = 1,\ P_{ij} \geq 1 \text{ and } P_{ji} \geq 1 \\ 0 & \text{otherwise} \end{cases}. \tag{4–3}$$

The pathogens of i and j are *sibling* if they are adjacent and there is no subgraph for which one is ancestral to the other,

$$R^{1\sqcup 2} = \begin{cases} 1 & \text{if } A_{ij} = 1,\ P_{ij} = 0 \text{ and } P_{ji} = 0 \\ 0 & \text{otherwise} \end{cases}. \tag{4–4}$$

Lastly, if the pathogens of i and j are not adjacent, they are said to be *disconnected*,

$$R^{1\not\leftrightarrow 2} = \begin{cases} 1 & \text{if } A_{ij} = 0 \\ 0 & \text{otherwise} \end{cases}. \tag{4–5}$$

In what follows, we will consider several phylogenies reconstructed from across sliding genomics windows of the entire HIV-1 genome. We will index these phylogenies by w. The relation formed by the branches of i and j on the phylogeny reconstructed in window w is denoted by R_{ijw}, and can take 5 forms,

$$R_{ijw} \in \{R^{1\rightarrow 2}, R^{2\rightarrow 1}, R^{1\leftrightarrow 2}, R^{1\sqcup 2}, R^{1\not\leftrightarrow 2}\} \tag{4–6}$$

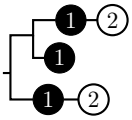
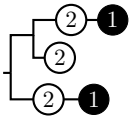
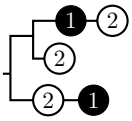
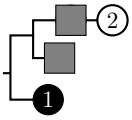
Notation	Meaning	Examples
$R^{1\rightarrow 2}$	pathogen of 1 is **ancestral** to the one of 2	
$R^{2\rightarrow 1}$	pathogen of 2 is **ancestral** to the one of 1	
$R^{1\leftrightarrow 2}$	pathogens of 1 and 2 are **intermingled**	
$R^{1\sqcup 2}$	pathogens 1 and 2 are **siblings**	
$R^{1\not\leftrightarrow 2}$	pathogens of 1 and 2 are **disconnected**	

Table 2: Notation and examples of the five phylogenetic topologies.

Secondly, we extract a measure of viral phylogenetic distance between two individuals.

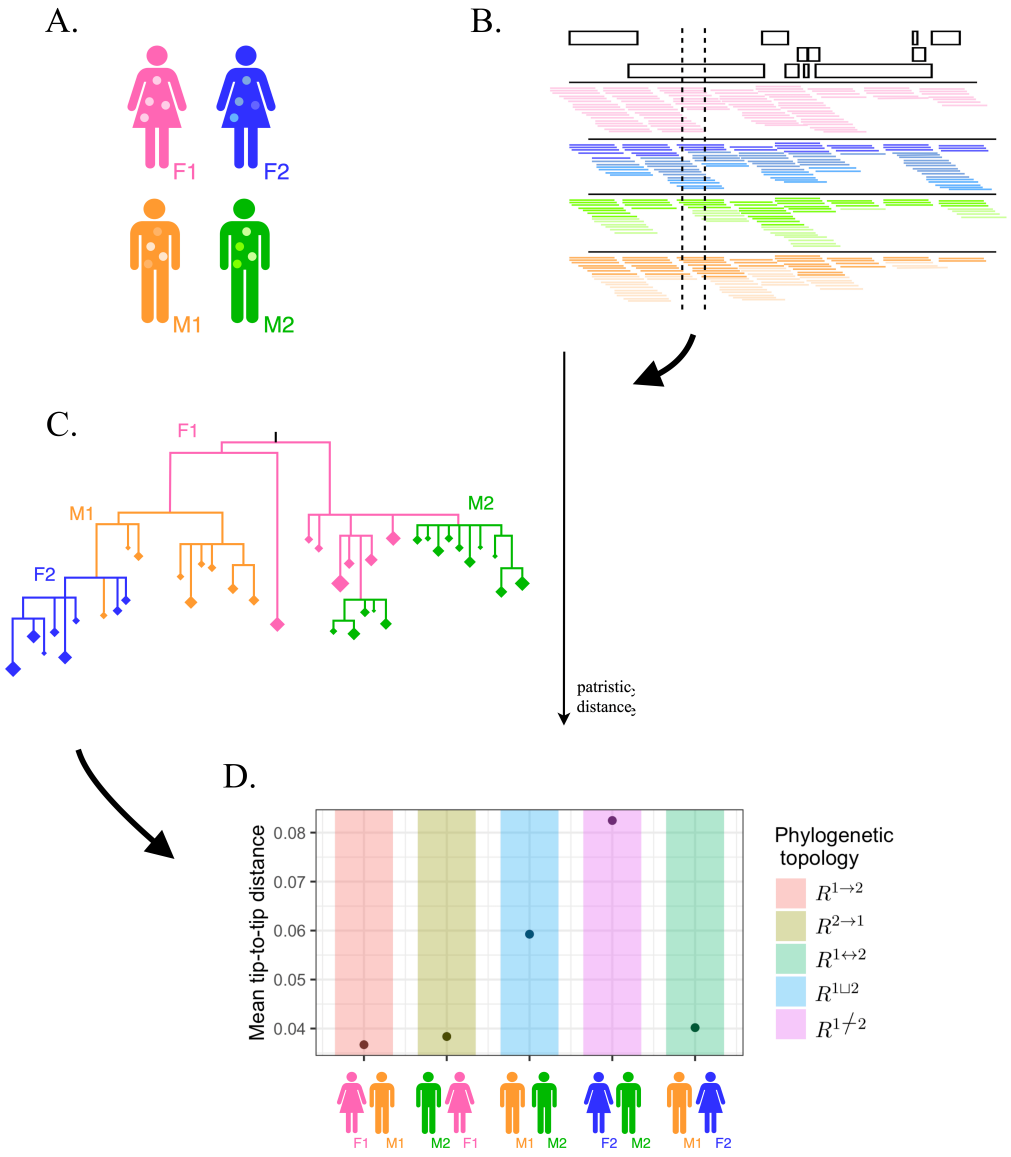


Figure 3: Finding phyloscanner outputs from HIV-1 deep-sequence data. In this example, the host population is composed of two females and two males. (**A**.) The viral diversity within-host is illustrated by dots of different color living inside the four infected individuals. Pathogens are sampled from the population. (**B**.) Deep-sequencing produces reads, which are fragments of the genome sequence of one pathogen. They are mapped against the HIV-1 genome shown in Figure 2. The success rate of sequencing and extraction is variable along the genome. (**C**.) Phylogenies are inferred for every 335 sliding windows along the genome. In one of the windows delimited by dotted lines, the reconstructed phylogeny is shown. The vertical distance of the phylogeny is the patristic distance that quantifies the evolutionary divergence between two nodes. The horizontal distance does not have an interpretation. Colors identify individuals, diamonds indicate unique fragments read, and the size of the diamonds corresponds to the number of copies. (**D**.) From the phylogeny we extract two outputs, the phylogenetic topology defined in (4–6) and the mean tip-to-tip patristic distance defined in (4–8).

The *patristic distance* between two viral reads in a phylogenetic tree is the estimated number of substitutions per site separating the two reads, which is a real number. It quantifies the extent of evolutionary divergence between them. Phylogenetic distance between two individuals can be measured in many ways because each individual is represented by many reads. In this thesis, we use the mean tip-to-tip patristic distance. Let the number of reads belonging to individual i on a phylogeny be n_i. There are $n_i \times n_j$ tip combinations between i and j. We denote by d_{ij}^k the patristic distance between the tips involved in the kth combination and passing through the node of the most recent common ancestor in the phylogeny. The mean tip-to-tip patristic distance between individuals i and j in a phylogeny is then defined by

$$D_{ij} = \frac{1}{n_i \times n_j} \sum_{k \in n_i \times n_j} d_{ij}^k. \tag{4–7}$$

By indexing the reconstructed phylogenies by window, the mean tip-to-tip patristic distance between i and j in the wth phylogeny is,

$$D_{ijw} = \frac{1}{n_{iw} \times n_{jw}} \sum_{k \in n_{iw} \times n_{jw}} d_{ijw}^k. \tag{4–8}$$

In Wymant et al. (2017) and Ratmann et al. (2019), the shortest patristic distance between the subgraphs of two individuals was used. We opt for a different approach because we want to estimate the time of infection and need to find a type of distance that can be specified as a function on the former. The dependence form between the time of infection and the mean tip-to-tip patristic distance will be explain later.

Sequencing success rates were low on this cohort and varied substantially across the genome. For this reason, phylogenetic relationships of individuals were only evaluated when they had at least 30 reads overlapping a particular genomic window. This meant that the number of deep-sequence phylogenies, in which the epidemiologic relationship of a pair of individuals i and j was assessed, varied, and rarely reached the maximum possible number.

4.3 Data sets used in this thesis

4.3.1 Data used to estimate transmission chains

Using slightly different definitions of pairwise phylogenetic relationships, Ratmann et al. (2019) reconstructed 493 distinct sets of phylogenetically closely related individuals, which are called *phylogenetic transmission networks* referred to as 'network' in this thesis. The analyses in Ratmann et al. (2019) were based on the most likely transmission chains that connect individuals in the Rakai population.

In this thesis, we pre-selected 7 of the 493 transmission networks to establish proof of principle in sampling from the posterior distribution of transmission chains associated with each network. We present in Table 3 the characteristics and data summary of the individuals involved in the 7 networks. The first column lists all individuals on all networks. The second and third column precise their gender and time of sampling. The fourth column specified the pairwise linkage according to the most likely transmission chain as estimated in Ratmann et al. (2019). Specifically, it shows the entry of the source vector for every individual. The

index case, who was infected by an external source, is identified by a source value of 0. For pair ij, the fifth column shows the median and 95% credible interval of their mean tip-to-tip patristic distance observations,

$$\mathrm{Q}_{50\%,[2.25\%,97.5\%]}(\{D_{ijw}\}_{w\,\in\,W(i,j)}) \tag{4–9}$$

where $\mathrm{Q}_{q\%}(.)$ is the q quantile of the distance observations. The set $W(i,j)$ contains windows on which the reconstructed phylogeny included pathogen reads of both i and j. That is windows on which we observe phylogenetic data for pair ij. This set may differ from pair to pair but is the same for pair ij and ji. The cardinality of set $W(i,j)$ is n_{ij}. The sixth column shows the count (and proportion) of phylogenies in whom the source i was topologically ancestral to the recipient j,

$$\sum_{w\,\in\,W(i,j)} \mathbb{1}_{\{R_{ijw}\,=\,R^{1\rightarrow 2}\}}, \tag{4–10}$$

where $\mathbb{1}$ is the indicator function. The proportion (indicated in parentheses) is obtained by dividing the count by n_{ij}. And vice versa in the seventh column, where the count is

$$\sum_{w\,\in\,W(i,j)} \mathbb{1}_{\{R_{ijw}\,=\,R^{2\rightarrow 1}\}}, \tag{4–11}$$

and we divide by n_{ij} to obtain the proportion. One might note that,

$$\mathbb{1}_{\{R_{ijw}\,=\,R^{2\rightarrow 1}\}} = \mathbb{1}_{\{R_{jiw}\,=\,R^{1\rightarrow 2}\}}, \forall\; i=1,...,N; j=1,...,N; j\neq i, w\in W(i,j). \tag{4–12}$$

These two proportions may not sum to one because the topological relations R_{ijw} can take up to five categories. We observe that, for each pair, the count (4–10) is higher than (4–11). This is not surprising as Ratmann et al. (2019) used this information to identify the direction of transmission.

4.3.2 Data used for model building

In the next section, we develop a likelihood model for the deep-sequence phylogenetic data, presented in (4–6) and in (4–8), that depends on certain genetic parameters $\boldsymbol{\psi}$. To specify the prior distribution of these parameters, we choose an informative empirical Bayes approach. Specifically, we consider pairs of individuals with almost certainly known epidemiologic relationship, and between whom the deep-sequence phylogenetic statistics in (4–6) and (4–8) are observed. Using this data, we can infer the posterior distribution of $\boldsymbol{\psi}$ and use this distribution as the prior distribution for our purposes, i.e., inference of transmission chains.

Individual	Gender	Time of sampling	Most likely source	Average mean tip-to-tip patristic distance	Number of topology with consistent direction	Number of topology with opposite direction
Network 1: $N = 2$ individuals						
RkA07581M	male	2012.68	0	-	-	-
RkA07578F	female	2012.68	RkA07581M	0.023 [0.0128, 0.0579]	33 (54.1%)	22 (36.07%)
Network 2: $N = 2$ individuals						
RkA07286F	female	2015.50	0	-	-	-
RkA07207M	male	2013.23	RkA07286F	0.0287 [0.0111, 0.1732]	22 (40%)	14 (25.45%)
Network 3: $N = 2$ individuals						
RkA07342F	female	2010.08	0	-	-	-
RkA07150M	male	2008.46	RkA07342F	0.0263 [0.0116, 0.0853]	74 (42.05%)	51 (28.98%)
Network 4: $N = 3$ individuals						
RkA02808M	male	2011.90	0	-	-	-
RkA00505F	female	2011.91	RkA02808M	0.0441 [0.016, 0.1008]	28(50.91%)	11 (20%)
RkA03589F	female	2011.91	RkA02808M	0.0337 [0.0193, 0.1013]	8 (40%)	6 (30%)
Network 5: $N = 4$ individuals						
RkA06111M	male	2013.33	0	-	-	-
RkA05669F	female	2011.72	RkA06111M	0.0275 [0.0158, 0.0673]	39 (54.93%)	11 (15.49%)
RkA03794M	male	2013.36	RkA06111M	0.1933 [0.1115, 0.4242]	42 (63.64%)	6 (9.09%)
RkA04533F	female	2013.29	RkA03794M	0.0496 [0.0239, 0.1886]	24 (100%)	0 (0%)
Network 6: $N = 5$ individuals						
RkA05079F	female	2012.50	0	-	-	-
RkA04621M	male	2012.53	RkA05079F	0.0252 [0.0111, 0.0681]	45 (42.45%)	22 (20.75%)
RkA01459F	female	2013.02	RkA04621M	0.0408 [0.0128,0.1538]	24 (22.43%)	8 (7.48%)
RkA01733M	male	2012.56	RkA01459F	0.0376 [0.014, 0.1401]	22 (22.45%)	16 (16.33%)
RkA05257M	male	2012.35	RkA01459F	0.0212 [0.0076, 0.0717]	54 (35.06%)	44 (28.57%)
Network 7: $N = 5$ individuals						
RkA00569M	male	2012.07	0	-	-	-
RkA01258F	female	2011.89	RkA00569M	0.0511 [0.0273, 0.0983]	24 (96%)	1 (4%)
RkA04356F	female	2011.93	RkA00569M	0.0317 [0.0185, 0.0557]	38 (69.09%)	4 (7.27%)
RkA01351M	male	2012.22	RkA00569M	0.0688 [0.0225, 0.132]	17 (30.91%)	5 (9.09%)
RkA02116F	female	2012.25	RkA01351M	0.0224 [0.0121,0.044]	24 (43.64%)	14(25.45%)

Table 3: Characteristics of the networks. The first column presents all individuals included in the networks. Their gender and sampling time are shown in the second and third column, respectively. Ratmann et al. (2019) reconstructed transmission chains by finding the most likely source to every individual in the fourth column. The index case (i.e., first individual infected in the network) is assumed to have been infected by an external individual denoted by 0. For linked pairs, summary statistics of phylogenetic data, defined in (4–9), (4–10) and (4–11), are presented in the fifth, sixth and seventh columns.

5 Construction of a Bayesian model for reconstructing transmission chains from deep-sequence data

5.1 Overall model

This project aims to estimate transmission chains from HIV-1 deep-sequence phylogenies, reconstructed with phyloscanner. To estimate $\boldsymbol{T}$, we adopt a Bayesian approach. Let us denote the data, which comprise the pairwise viral phylogenetic relationships and the pairwise mean tip-to-tip patristic distances defined in (4–6) and (4–8) respectively, by

$$\boldsymbol{x} = \{R_{ijw}, D_{ijw}\}_{i=1,\ldots,N;\ j=1,\ldots,N;\ j\neq i;\ w\in W(i,j)}. \tag{5–1}$$

This set includes the observable topological relations, and distances of all $N \times (N-1)$ pairs ij (order matters) in the host population. The set $W(i,j)$ contains windows on which phylogenetic data was available for pair ij. Assuming that the phylogenetic relationships do not depend on the patristic distances, we consider

$$\begin{aligned} p(\boldsymbol{T},\boldsymbol{\theta},\boldsymbol{\psi}|\boldsymbol{x}) &\propto p(\boldsymbol{x}|\boldsymbol{T},\boldsymbol{\theta},\boldsymbol{\psi})p(\boldsymbol{T},\boldsymbol{\theta},\boldsymbol{\psi}) \\ &= \prod_{ij}\prod_{w\in W(i,j)} p(D_{ijw},R_{ijw}|\boldsymbol{T},\boldsymbol{\psi})p(\boldsymbol{T}|\boldsymbol{\theta})p(\boldsymbol{\theta})p(\boldsymbol{\psi}) \\ &= \prod_{ij}\prod_{w\in W(i,j)} p(D_{ijw}|\boldsymbol{T},\boldsymbol{\psi})p(R_{ijw}|\boldsymbol{T},\boldsymbol{\psi})p(\boldsymbol{T}|\boldsymbol{\theta})p(\boldsymbol{\theta})p(\boldsymbol{\psi}), \end{aligned} \tag{5–2}$$

where $p(D_{ijw}|\boldsymbol{T},\boldsymbol{\psi})$ is the likelihood of the observed mean tip-to-tip patristic distance conditional on the transmission chain $\boldsymbol{T}$ and latent parameters $\boldsymbol{\psi}$, $p(R_{ijw}|\boldsymbol{T},\boldsymbol{\psi})$ is the likelihood of the observed phylogenetic relationships conditional on the transmission chain $\boldsymbol{T}$ and latent parameters $\boldsymbol{\psi}$, $p(\boldsymbol{T}|\boldsymbol{\theta})$ is the prior probability of the transmission chain given epidemiological parameters $\boldsymbol{\theta}$, and $p(\boldsymbol{\theta})$, $p(\boldsymbol{\psi})$ are prior densities on $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ respectively. The latent parameters $\boldsymbol{\psi}$ include, among others, a standardised HIV-1 substitution rate as we discuss below. The epidemiologic parameters $\boldsymbol{\theta}$ include a parameter which quantify the transmission rate.

The product in (5–2) is taken over pairs ij with direct transmission, with a left-to-right direction. This is $ij \in \{S_k, k\}_{k=2,\ldots,N}$, where, without loss of generality, 1 is the index case, associated with an external source. This restriction is made for scalability as illustrated in Figure 4. A second product is then taken over all windows $w \in W(i,j)$, where ij have observable phylogenetic data. Therefore we assume firstly that all connected pairs are independent, and secondly that, for each of these pair, phylogenetic data are independent across windows.

In Section 5.2 and 5.3, we specify the likelihood components of (5–2). We also use an empirical Bayes approach and specify the priors of the corresponding nuisance parameters $\boldsymbol{\psi}$, based on available data from couples in known long-term sexual relationships, from who deep-sequence data was previously extracted. The prior on $\boldsymbol{T}$ is developed in Section 5.4.

5.2 Mean tip-to-tip patristic distance likelihood

Our specification of the distance likelihood $p(D_{ijw}|\boldsymbol{T},\boldsymbol{\psi})$ is motivated by prior work of Leitner and Albert (1999), Morelli et al. (2012), Vrancken et al. (2014), Ratmann et al. (2016, 2019), which develop an HIV-1 empirical clock model based on the time elapsed since divergence of

the virus in the founder individual. The main idea is that then, conditional on the evolutionary time elapsed, it is possible to specify the probability of the tip-to-tip patristic distance as a function of the HIV-1 substitution rate and the time elapsed.

5.2.1 Time elapsed

A problem with the approach is that the time of divergence of the virus in the founder individual between the viral lineages of individuals i and j is not known, and hard to elicit from $\boldsymbol{T}$. To circumvent this problem, we assume immediate divergence of viral lineages within hosts, so that the time at which divergence starts coincides with the infection time of the founder individual. Consider coloring the branches of the phylogeny with a unique color for each host. Intuitively, we assume that nodes of the phylogeny connecting two branches with different color coincide with transmission events. This is illustrated in Figure 4. The same assumption has been made in Morelli et al. (2012), Ypma et al. (2012) and Didelot et al. (2014). Under this assumption, we can calculate the time since the start of divergence until the time of sequence sampling of i and j by

$$t_{ij}^E = (t_i^S - t_{\text{MRCA}_{ij}}^I) + (t_j^S - t_{\text{MRCA}_{ij}}^I), \tag{5–3}$$

where t_i^S is the sampling time of individual i and $t_{\text{MRCA}_{ij}}^I$ denotes the infection time of the *most recent recent ancestor* (MRCA) of individuals i and j. It is the most recent individual from which i and j are directly descended in the transmission chain. One can note that MRCA_{ij} and MRCA_{ji} refer to the same individual. Moreover, if individual i transmitted the virus to individual j, then MRCA_{ij} is i. Because we consider only directly linked pair in (5–2), the MRCA of ij designated one of i or j. Our assumption requires that the time of infection of MRCA_{ij} corresponds to the node regrouping all pathogens of i and j before they started to mutate separately. We refer to (5–3) as the *time elapsed* between sampled virus from i and sampled virus from j. The time elapsed can be directly calculated from the variables contained in $\boldsymbol{T}$. Notice further that (5–3) is similar to the 'evolutionary time' used by Morelli et al. (2012), presented in 3–21.

The time elapsed corresponds to the period of the evolutionary distance formation, as illustrated in Figure 5. The most straightforward possible approach for modeling patristic distances is to assume a strict molecular clock, which is constant across hosts. In this case,

$$\text{E}[D_{ijw}|t_{ij}^E, \mu] = t_{ij}^E \times \mu, \tag{5–4}$$

where μ denotes the HIV-1 substitution rate. Unfortunately, as we show now, this model does not capture the empirical relationship between patristic distance and time elapsed well.

5.2.2 Data for exploratory analysis

In general, it is challenging to know that one individual transmitted HIV-1 to another individual, and thus to characterize the empirical relationship between patristic distance and time elapsed in transmission pairs. Among participants of the RCCS, 415 indicated that they were in a long-term stable relationship. For 133 of them, phylogenetic analysis from Ratmann et al. (2019) indicated linkage and resolved direction. Besides, for 27 pairs, an estimate of the

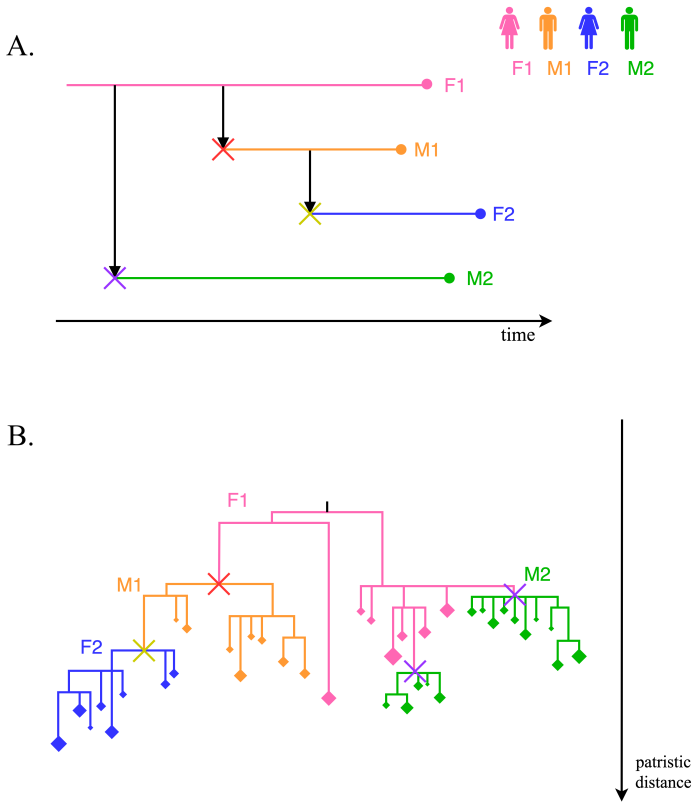


Figure 4: From transmission history to phylogenetic tree. The infected population is composed of two females and two males referred to as F1, F2, and M1, M2, respectively. There is in total 4×3 pairs (the order matters). (**A**.) The history of transmission shows how the pairs are related. Selecting only directly linked pairs (i.e., with right-to-left and left-to-right direction) reduces the number to eight pairs, and in one direction to only four pairs. This illustrates why choosing only directly linked pairs with left-to-right direction is a gain in term of scalability. Next, we are interested in the transmission linkage and direction between M1 and F2. We observe from the transmission history that M1 infected F2. (**B**.) The phylogeny reconstructed on window w that respects our hypothesis illustrates the diversity of individuals' sampled pathogens. The virus of M1 is phylogenetically ancestral to that of F2. Our assumption requires that, as soon as transmission occurred, the pathogen infecting F2 replicates and mutates to create diversity. Therefore, the first root of the subgraphs of F2 corresponds to her time of infection. An individual might transmit more than one type of pathogen to the recipient as in the transmission between F1 and M2. In this case, our hypothesis requires that all transmitted pathogens create their diversity instantly at the time of infection.

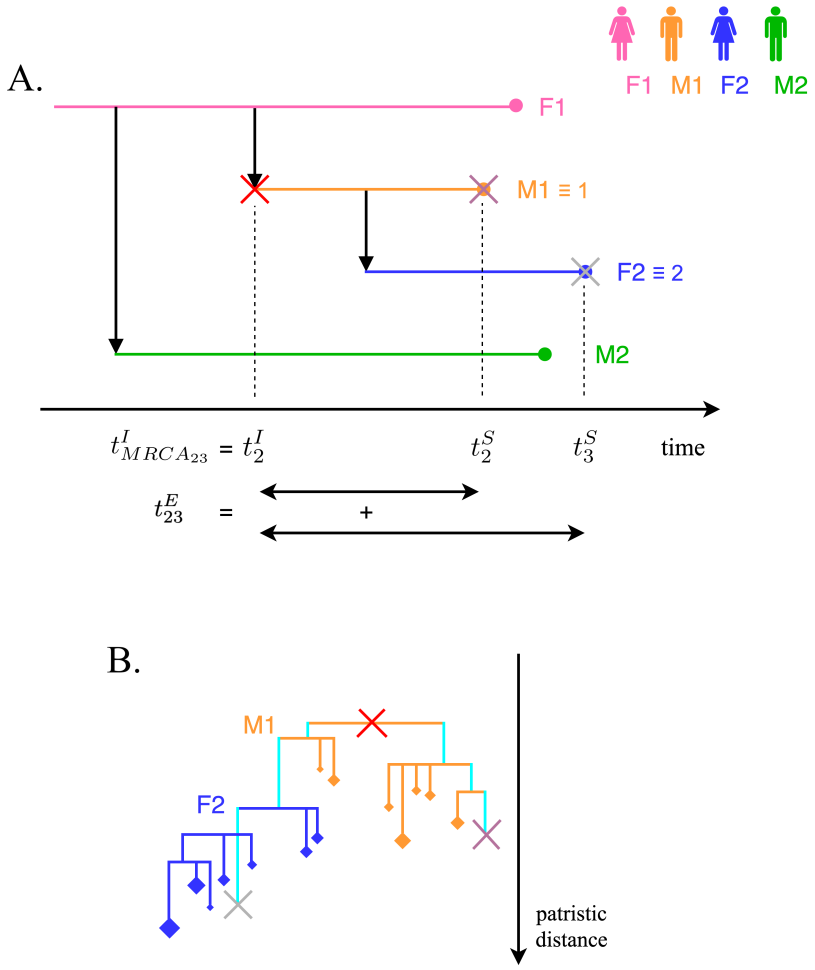


Figure 5: Correspondence between the timed transmission chain and the phylogeny. From this phylogeny, 8×8 tip-to-tip patristic distances are extracted for individuals M1 and F2 (refers as 1 and 2 in the time notation for visibility). One of them is equal to the sum of the turquoise branches. By identifying their sampling times and the infection time of M1, we observe that the tip-to-tip patristic distance has developed over the time elapsed.

time of infection is found for the source case. Specifically, in these pairs, the source case has a last negative and first positive HIV-1 test, and the infection time could be estimated as the midpoint between the them, which is commonly done this way (Skar et al., 2013, Ratmann et al., 2016, Hanson et al., 2016). We refer to these 27 selected pairs as 'related couples.'

5.2.3 Exploratory analysis

For these 27 related couples, 2,566 mean tip-to-tip patristic distances from multiple phylogenies across the genome are obtainable for analysis. The genomic windows start at position 800nt of the HXB HIV-1 reference genome shown in Figure 2, and slide in increments of 25 base pairs up to position 9400nt giving 335 windows in total. Figure 6 illustrates the regions

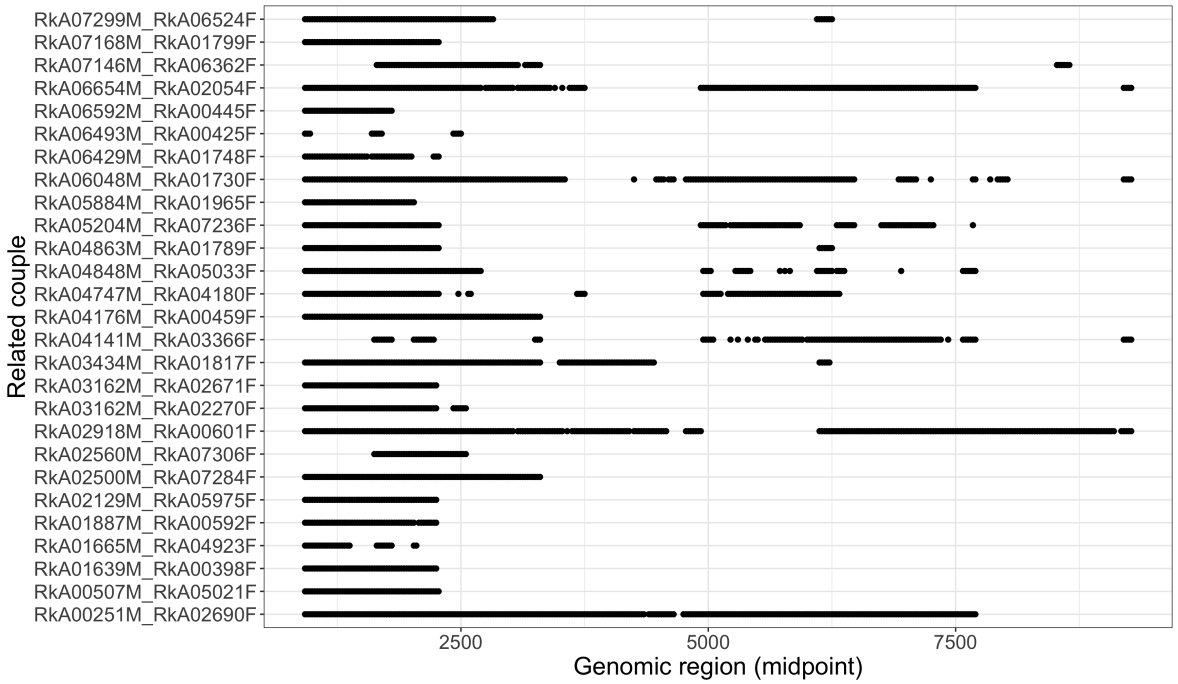
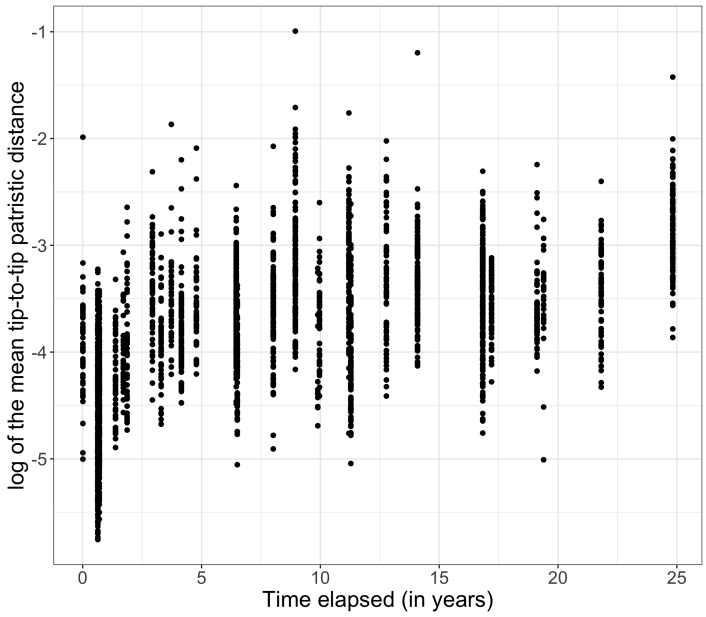


Figure 6: Observations of the mean tip-to-tip patristic distance along the genome by related couple. This Figure shows the windows $W(i, j)$ with observable phylogenetic data along the genome for each pair ij. A point represents an observation. The genomic region is the midpoint of the window coordinate. Individuals are designated with an encrypted identification number of the form 'RkxxxxxF,' where the first two letters indicate the Cohort study and the last later designated the gender, 'F' for female and 'M' for male. Source and recipient are separated with '_' in the related couple name.

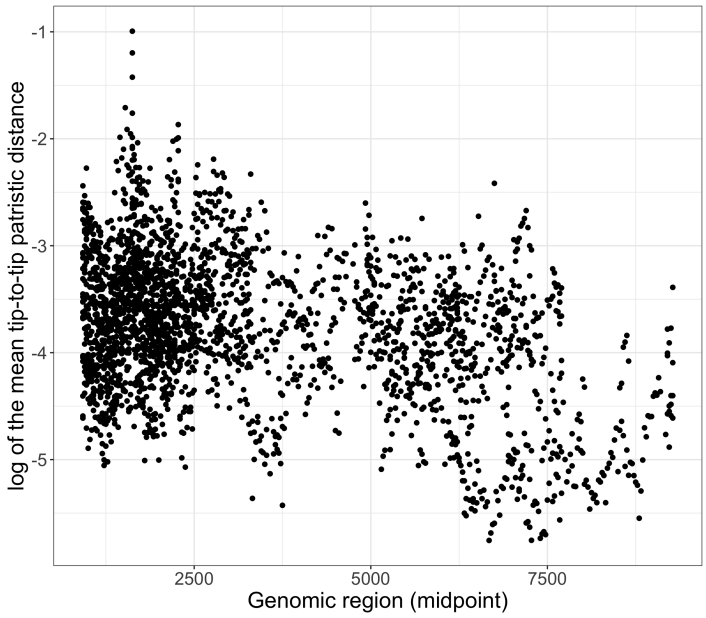
of the genome for which data are available. That is, it presents elements of set $W(i, j)$ for each related couples ij. We transform distances to the log scale for two reasons. First, there are many outliers on the original scale. The logarithmic transformation reduces the influence of these observations. Second, the variance of the distances on the original scale increases with time elapsed. Taking the log eliminates this heteroscedasticity. The transformation is further explored in Appendix A.

Dot plots reveal, as expected, trends in the mean of the log tip-to-tip patristic distances by time elapsed (Figure 7a), and by genomic region (Figure 7b). Some regions of the genome seem to be more favorable to mutation (i.e., creating distance) than others.

Lastly, we investigate the lag-1 correlation in the log distances across genomic windows for each related couple. Because related couples do not have an observation on all 335 windows, we consider blocks of windows and compute the lag-1 correlation for each of them. The median lag-1 correlation is 0.4713, with 95% credible interval [-0.6258, 0.8219].



(a) Log of the mean patristic distance against time elapsed



(b) Log of the mean tip-to-tip patristic distance against window coordinate

Figure 7: Log of the mean tip-to-tip patristic distance against time elapsed and window coordinate. For each related couple, mean tip-to-tip patristic distances are calculated in observed phylogenies across the genome. We compare the distance distribution of each related couple to the genomics region in which the phylogeny was reconstructed (7a) and to the time elapsed (7b).

5.2.4 Model

We model the log distances in a hierarchical mixed-effects model, in which we consider dependence on time elapsed as a fixed effect and dependence on window coordinates as a random effect. The model is,

$$logD_{ijw}|\mu_{ijw}, \sigma \sim t_\nu(\mu_{ijw}, \sigma), \tag{5–5a}$$

$$\mu_{ijw} = \alpha + \beta \times t_{ij}^E + \sum_{l=1}^{w} \delta_l, \tag{5–5b}$$

$$\alpha \sim \mathcal{N}(0, 100), \tag{5–5c}$$

$$\beta \sim \mathcal{N}(0, 10), \tag{5–5d}$$

$$\delta_1 \sim \mathcal{N}(0, 100), \tag{5–5e}$$

$$\delta_w|\delta_{w-1} \sim \mathcal{N}(\delta_{w-1}, \sigma_\delta^2), \quad w \in \{2, ..., 335\}, \tag{5–5f}$$

$$\sigma_\delta, \sigma \sim C^+(0, 1), \tag{5–5g}$$

$$\nu \sim Gamma(2, 0.1), \tag{5–5h}$$

where t_ν denotes the Student t-distribution with ν degrees of freedom, and $C^+(x_0, \gamma)$ denotes the half-Cauchy distribution with density,

$$\mathrm{f}(x|x_0, \gamma) = \frac{2}{\pi\gamma[1 + (\frac{x-x_0}{\gamma})^2]}, \quad x \geq 0, \gamma > 0. \tag{5–6}$$

The average mutation rate of the log tip-to-tip patristic distance mean, μ_{ijw}, is expressed per unit time and window. To model it, a linear effect β captures the variation in the average log distance for an additional year elapsed. The increment does not vary with time, corresponding to a strict molecular clock model. α represents the overall mean across windows. Heterogeneity in the genome is captured for each genomic window on which phylogenies were reconstructed. Specifically, we model such heterogeneities by specifying random increments δ_w at each genomic window coordinate, indexed by w. The δ_w are given an AR1 prior (Havard and Leonhard, 2005) to allow for arbitrary functional forms of heterogeneity, subject to AR1 smoothing specified by the parameter σ_δ^2. Several outliers on the response are observed (Appendix A). They are likely due to contamination of sequencing readings or errors in phylogeny reconstruction. To account for them, we use a Student-t distribution, with ν degrees of freedom, on the response. Finally, the variance in the log distances, σ is assumed to be constant across the genome and time elapsed.

The prior densities on the baseline and the fixed effect are chosen to be vague. The prior density on the random effect of the first window was specified to be less restrained than the following, to avoid over-constraining subsequent increments (Ghitza, 2014). The degree of freedom follows a gamma distribution that was found to work well in practice in Juárez and Steel (2010). Similarly, the priors on the standard deviations are chosen for practical reasons (Polson and Scott, 2012, Gelman, 2006).

5.2.5 Model fit

Model 5–5 is fit with Stan version 2.18.1, using 3 Hamiltonian Monte Carlo chains with 15,000 iterations each, of which the first 2,000 iterations are considered as a burn-in. There are no problems in convergence and mixing; the minimum and maximum effective sample sizes are respectively 981 and 58,745.

Table 4 shows the empirical median and 95% credible intervals for the marginal posterior distributions of all model parameters except the windows' random effects. The mean of the log tip-to-tip patristic distance between reads of two partners was estimated to increase by 0.0463 [0.0435, 0.0491] log subst/site for every additional year elapsed. Very similar estimates are obtained in alternative models fitted in Appendix B. The median of the degrees of freedom ν was relatively small, indicating significant discrepancies between the tails of the Student-t and those of the normal. The posterior distribution of the δ_w describe changes in the expected log distance from window $w-1$ to window w for time elapsed held fixed. It is more useful to look at the posterior distribution of

$$\Delta_w = \sum_{l=1}^{w} \delta_l, \tag{5–7}$$

which quantifies the difference in the expected log distance compared to the average response in window w, for time elapsed held fixed. Figure 8 presents the estimated marginal posterior distributions of Δ_w for every window. There is significant heterogeneity in the expected log distance across the genome.

Lastly, we investigate if our model was able to explain the serial auto-correlation in log distances across windows. For related couple ij, we compute the lag-1 correlation in the posterior expected distance and posterior errors on every block of observed windows. The posterior errors are taken to be the difference between the observed distance and the expected distance,

$$\epsilon^k_{ijw} = D_{ijw} - \mu^k_{ijw}, \tag{5–8}$$

where μ^k_{ijw} is the estimated mean of the log distance of related couple ij in window w at the kth iteration and $k \in \{1, ..., K\}$ is the Monte Carlo sample index. Table 5 presents the median and 95% credible interval of the empirical lag-1 correlation (computed in 5.2.3), the posterior lag-1 correlation of the expected distance μ_{ijw}, and the lag-1 autocorrelation posterior residuals ϵ_{ijw}.

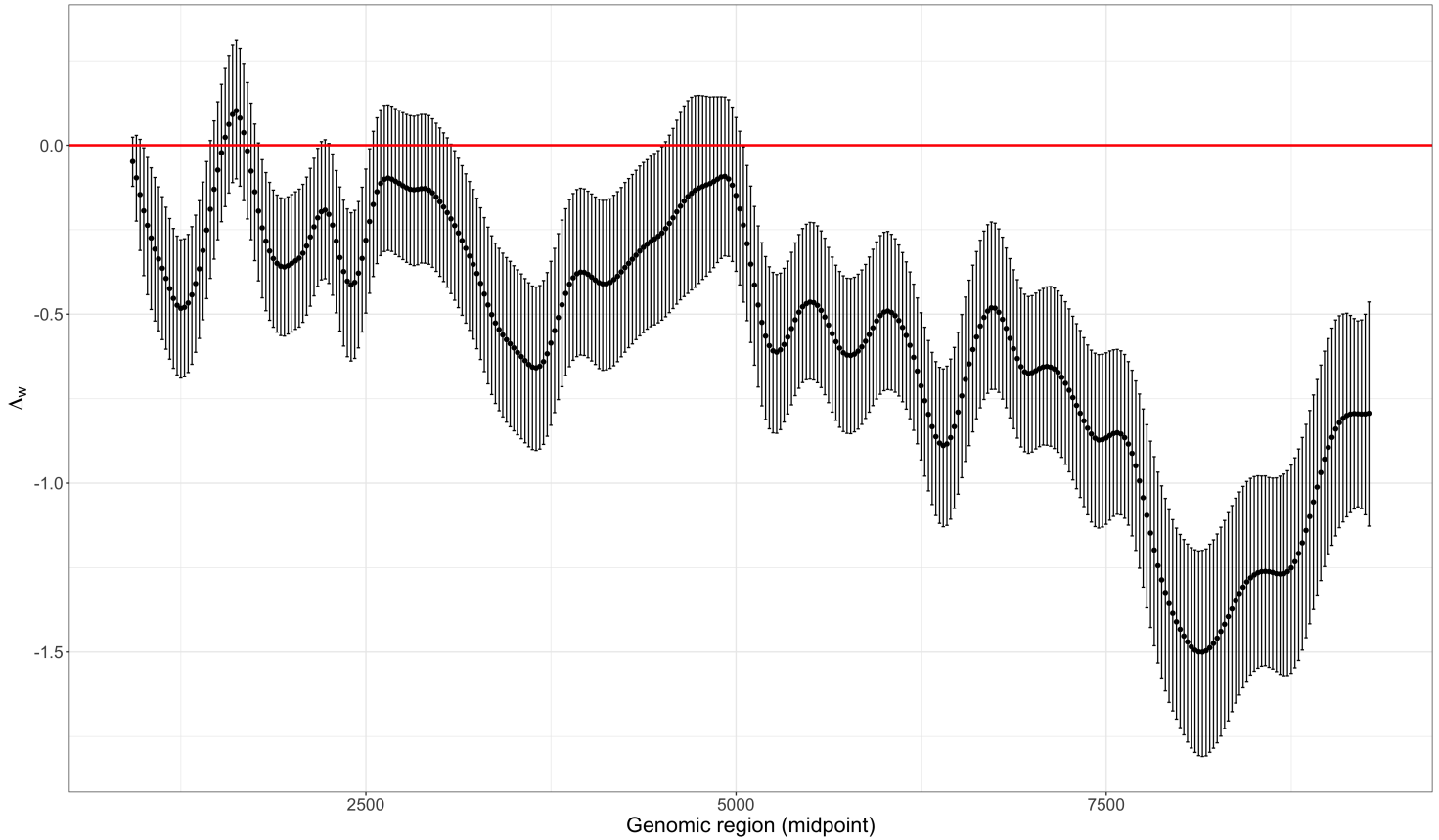


Figure 8: Marginal posterior distribution of Δ_w We fit model 5–5 using 3 Hamiltonian Monte Carlo chains. Each iteration (excluding burn-in) is considered as observation of the posterior distribution for each parameter. Taking into account all these observations, a marginal empirical posterior distribution can be formed for Δ_w defined in (5–7). This is the difference in the expected average of the log mean tip-to-tip patristic distance in window w compared to the baseline. Genomics region is the midpoint of the window coordinate. Shown are the 95% credibility intervals (lines) and median (dots). If a credible interval does not lie on the red line, the average log distance on the associated window is significantly different from the baseline response at the 5 % level.

Parameter	Median	Credible interval
α	-3.7710	[-3.9590, -3.5820]
β	0.0463	[0.0435, 0.0491]
σ	0.4655	[0.4468, 0.4831]
σ_δ	0.0199	[0.0136, 0.0287]
ν	24.2066	[13.8557, 50.9355]

Table 4: Median and credible interval of model 5–5 parameters. We fit model 5–5 using 3 Hamiltonian Monte Carlo chains. Each iteration (excluding burn-in) is considered as observation of the posterior distribution for each parameter. Taking into account all these observations, a marginal empirical posterior distribution can be formed for each parameter. We report the empirical median and 95% confidence interval of all model parameters except the window random effects.

Lag-1 correlation	Median	Credible interval
Empirical correlation (on D_{ijw})	0.4713	[-0.6258, 0.8219]
Posterior correlation (on μ_{ijw})	0.7278	[-0.5000, 0.9714]
Posterior residuals autocorrelation (on ϵ_{ijw})	0.4207	[-0.6258, 0.8219]

Table 5: The median and credible interval of the lag-1 correlation of D_{ijw}, μ_{ijw} and ϵ_{ijw} by observed blocks of windows. We fit the model 5–5 using 3 Hamiltonian Monte Carlo chains. For each related couple, we identify the window blocks on which distances are observed. The lag-1 correlation was taken on these observations, and summary statistics are calculated among the window blocks and related couples (first line). Then, the lag-1 correlation was calculated for each of the posterior draws of μ_{ijw} and the median and credible interval was taken among the window blocks, related couples, and Monte Carlo iterations (second line). Similarly for the posterior residuals ϵ_{ijw} (third line).

5.3 Phylogenetic relationship likelihood

In this section, we wish to develop a likelihood model for the phylogenetic topological relationship data between pairs of source-recipient. Let R_{ijw} be the observed topology of pair ij in the reconstructed phylogeny of window w. As described in Table 2, R_{ijw} can take one of five possible values. We model the probability distribution of R_{ijw} through

$$R_{ijw}|\boldsymbol{\pi}_{ijw} \sim Categorical(\boldsymbol{\pi}_{ijw}), \tag{5–9}$$

where $\boldsymbol{\pi}_{ijw}$ denote the probabilities for a pathogen reads of ij to form each category in genomics region w, such that

$$\sum_r \pi^r_{ijw} = 1, \tag{5–10}$$

with $r \in \{R^{1\to 2}, R^{2\to 1}, R^{1\leftrightarrow 2}, R^{1\sqcup 2}, R^{1\not\to 2}\}$. We build a hierarchical model on the relationships by finding a model on these probabilities and specify the linear predictor in terms of the genomic region and time elapsed as explanatory variables.

5.3.1 Data for exploratory analysis

Among the 415 RCCS participants in long-term relationships, 35 of them presented epidemiological and phylogenetic evidence of transmission and direction. Considering both the epidemiological and phylogenetic data, it is thus very likely that transmission occurred within these couples. Moreover, for 9 of them, the source case has an estimate of the time of infection. We refer to them as 'connected couples' in the following. 634 phylogenetic relationships are available for these connected couples across the genomics windows. Because we consider only pairs with left-to-right direction, the phylogenetic relation $R^{1\to 2}$ has a direction coherent with the epidemiological direction of transmission. We refer to it as 'correct direction'. Similarly, $R^{2\to 1}$ is the 'incorrect direction'.

5.3.2 Exploratory analysis

Table 6 presents the count of observations by relation category and Figure 9 shows there distribution for every connected couples ij along observable windows $W(i,j)$.

We first look at the empirical proportion of each category within connected couples. This is the count of each observed phylogenetic topology divided by the total number of observation for the same pair. Figure 10 presents the median and credible intervals across topological data. The x-axis is the empirical proportion of the category within couples. The median of the empirical proportions within couples for the correct category is 50%. At least 97.5% of the proportion are bellow this value for all other categories. From this Figure, we observe which category is more likely to appear between pathogens of connected couples. Pathogen's reads form, the most often, a topology coherent with the epidemiological direction of transmission, and, the less often, they are 'intermingled'.

Phylogenetic relationship	$R^{1\to 2}$	$R^{2\to 1}$	$R^{1\leftrightarrow 2}$	$R^{1\sqcup 2}$	$R^{1\not\to 2}$
Count of observation	309	157	110	50	8

Table 6: Count of observed phylogenetic topologies for connected couples. We observe 634 topologies among 9 connected couples.

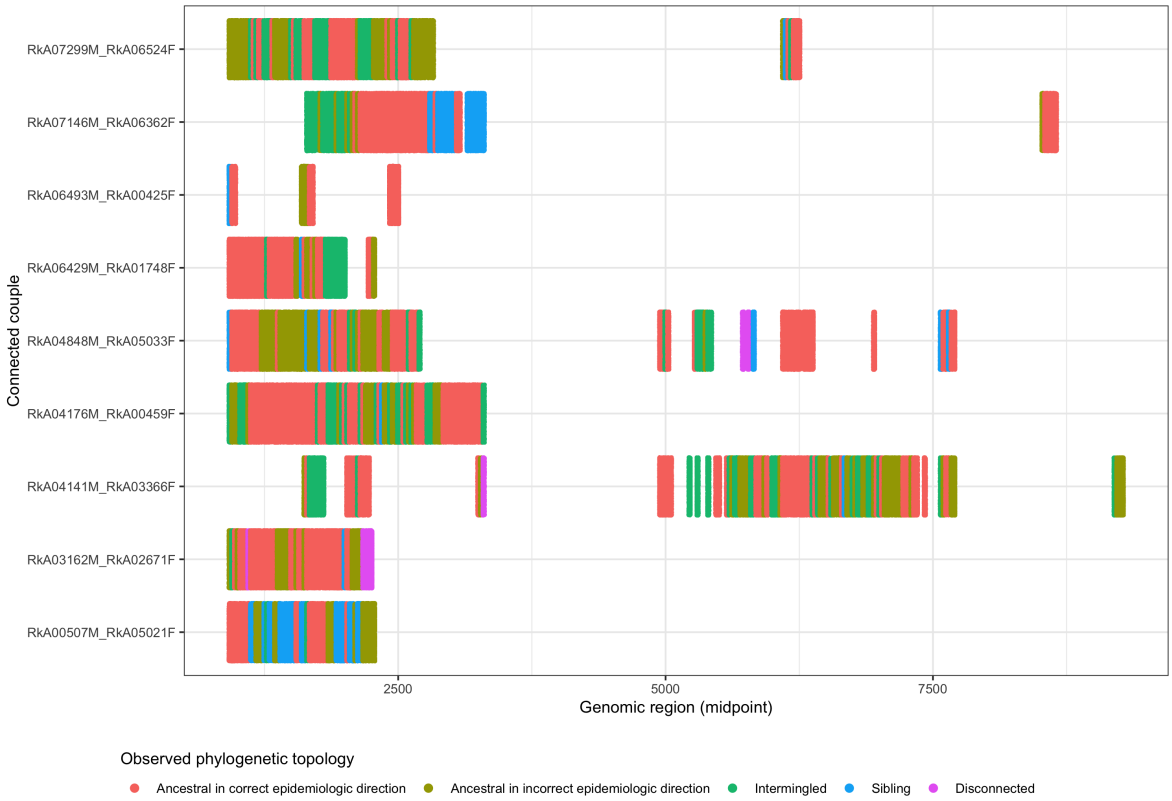


Figure 9: Observations of the phylogenetic topologies for connected couples. The five possible topological relationships between subgraphs of two individuals are presented in Table 2. We observe their distribution along the genomic windows $W(i, j)$ for each connected couples ij. Individuals are designated with an encrypted identification number of the form 'RkxxxxxF,' where the first two letters indicate the Cohort study and the last later designated the gender, 'F' for female and 'M' for male. Source and recipient are separated with '_' in the pair name. A bar represents an observation. Genomics region is the midpoint of the window coordinates.

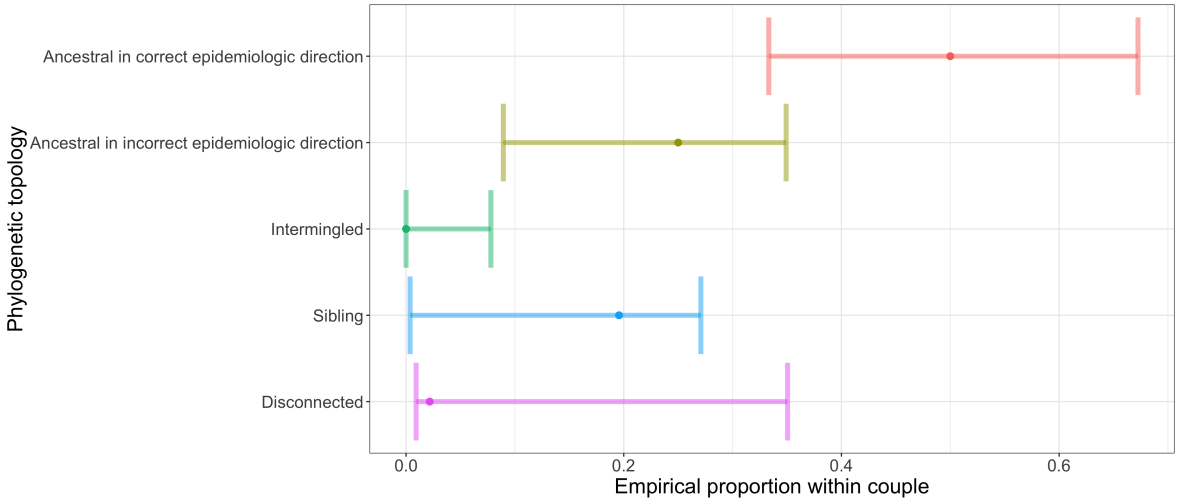


Figure 10: Median and 95% credible interval of the empirical proportions of phylogenetic relationship at couple level. For every connected couple, we extract the count of each phylogenetic relation and normalize them with the number of observation within that couple. We show the median (dots) and 95% credible interval (lines) of the empirical proportions for each category.

There might be heterogeneity in the classification among window coordinates. In other words, the type of relationships formed by the connected couple's branches in a phylogeny may depend on the genomics region on which this phylogeny was reconstructed. Figure 11 displays the empirical proportions distribution on every observed window. Relationships may also depend on the couple's time elapsed. Figure 12 shows Jeffrey's credible interval and probability point estimates by relations against time elapsed. These plots indicate heterogeneity at window and time elapsed level. For the windows, there is further uncertainty at the right of the genomics region, likely due to few observations. For the time elapsed, it seems that the probability of correct direction increases with time at the expense of the wrong direction. We want to capture this heterogeneousness by using these two levels as explanatory variables.

5.3.3 Model

We use a *categorical logistic regression* to describe the likelihood of observing a particular topological deep-sequence relationship in pair ij on genomic window w. Categorical logistic regression is often an attractive option for data such as ours because of its limited number of assumptions in comparison with the *discriminant function analysis* or the *categorical probit regression* (Pohar Perme et al., 2004, Starkweather and Moske, 2011).

We start with a few preliminary remarks to explain the model structure. Consider for any connected couple ij in window w a set of explanatory variables $\boldsymbol{z}_{ijw}$ of length m, and let the coefficients $\boldsymbol{\phi}_r$ take into account the effect of the explanatory variables on the probability that an observation of any couple falls into category r denoted by $\pi^r_{ijw} = p(R_{ijw} = r|\boldsymbol{z}_{ijw}, \boldsymbol{\phi}_{R^{1\to 2}}, ..., \boldsymbol{\phi}_{R^1 \not\to 2})$. These probabilities take the form of the soft-

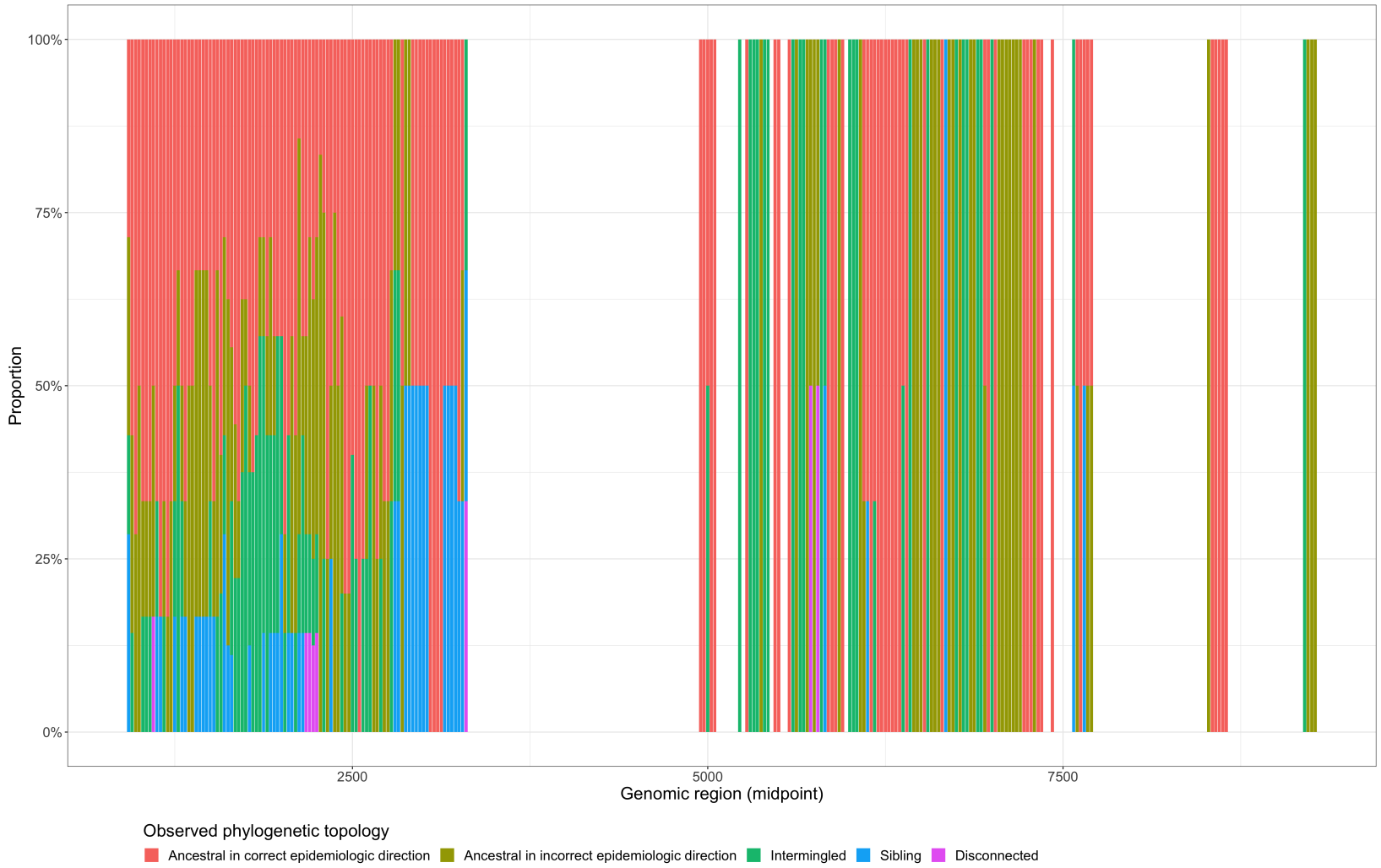


Figure 11: Proportion of each phylogenetic topology along genomics regions for connected couples. On every window, we extract the count of each phylogenetic relation and normalize them with the number of observation within that window. These proportions are stacked on each other.

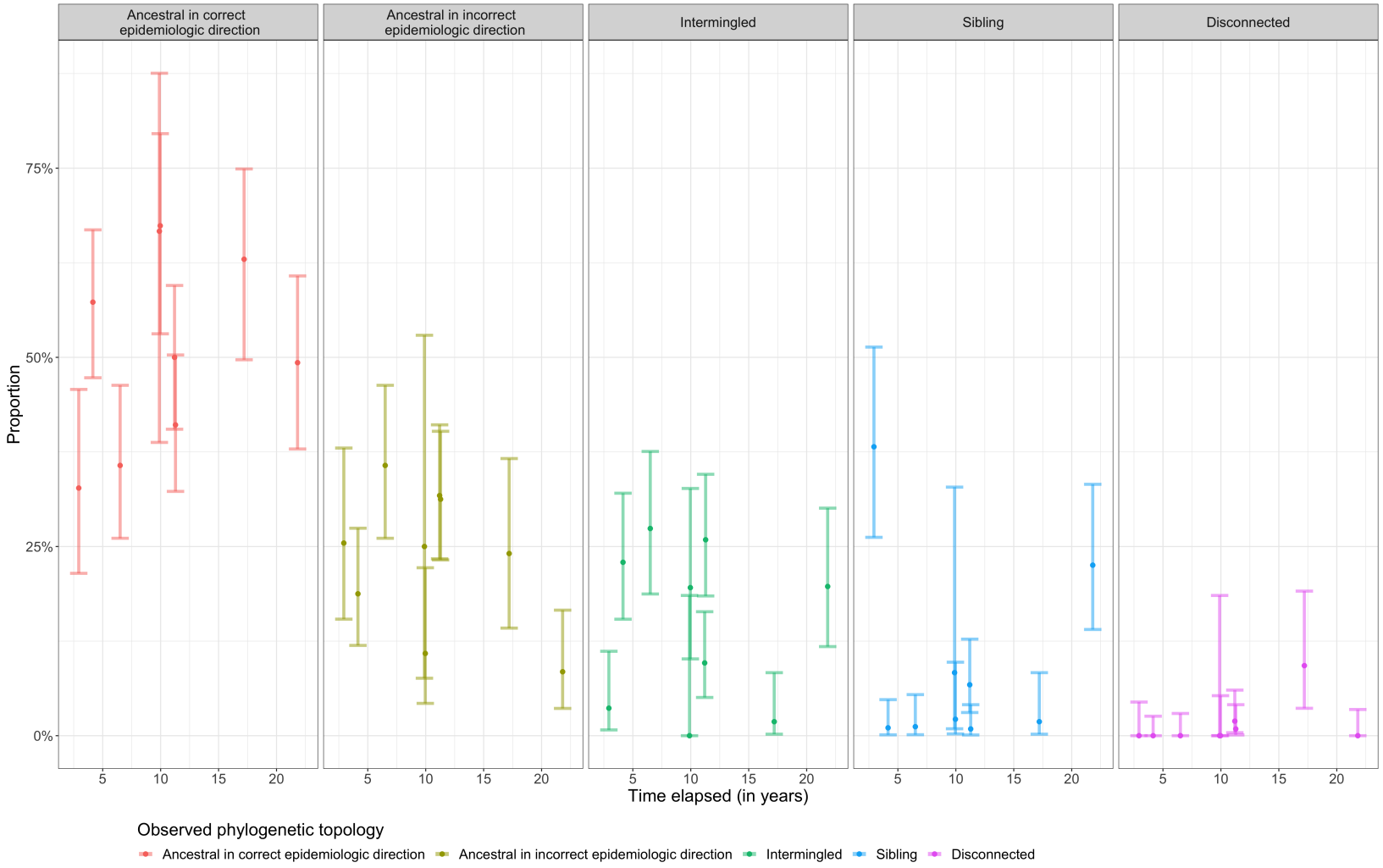


Figure 12: Proportion and confidence interval of each phylogenetic topology along time elapsed for connected couples. For every time elapsed we extract the count of each observed phylogenetic topology and normalize them with the number of observation within that time. The point estimate is shown with a dot. Jeffrey's confidence intervals are also computed and are displayed with a line.

max function,

$$\boldsymbol{\pi}_{ijw} = \begin{bmatrix} \pi_{ijw}^{R^{1\to 2}} \\ \pi_{ijw}^{R^{2\to 1}} \\ \pi_{ijw}^{R^{1\leftrightarrow 2}} \\ \pi_{ijw}^{R^{1\sqcup 2}} \\ \pi_{ijw}^{R^{1\not\leftrightarrow 2}} \end{bmatrix} = \frac{1}{\sum_r e^{\boldsymbol{z}_{ijw}^T \boldsymbol{\phi}_r}} \begin{bmatrix} e^{\boldsymbol{z}_{ijw}^T \boldsymbol{\phi}_{R^{1\to 2}}} \\ e^{\boldsymbol{z}_{ijw}^T \boldsymbol{\phi}_{R^{2\to 1}}} \\ e^{\boldsymbol{z}_{ijw}^T \boldsymbol{\phi}_{R^{1\leftrightarrow 2}}} \\ e^{\boldsymbol{z}_{ijw}^T \boldsymbol{\phi}_{R^{1\sqcup 2}}} \\ e^{\boldsymbol{z}_{ijw}^T \boldsymbol{\phi}_{R^{1\not\leftrightarrow 2}}} \end{bmatrix}, \tag{5–11}$$

where ij is any directly connected couple with an observation in window w. Notice that $1/\sum_r e^{\boldsymbol{z}_w^T \boldsymbol{\phi}_r}$ normalises the distribution, so that it sums to one. The softmax regression has the property of having a set of 'redundant' parameters. This implies that adding $\boldsymbol{a} \in \mathbb{R}^m$ to all coefficient vectors $\boldsymbol{\phi}$ gives the same probabilities. To identify model parameters, some restrictions must be made, usually of the type $\boldsymbol{\phi}_k = \mathbf{0}$ for a category k, referred to as baseline. In this case, the probabilities in the Categorical model are given by:

$$\pi_{ijw}^r = \frac{e^{\boldsymbol{z}_{ijw}^T \boldsymbol{\phi}_r^T}}{1 + \sum_{l \neq k} e^{\boldsymbol{z}_{ijw}^T \boldsymbol{\phi}_l^T}}, \tag{5–12}$$

$$\pi_{ijw}^k = \frac{1}{1 + \sum_{l \neq k} e^{\boldsymbol{z}_{ijw}^T \boldsymbol{\phi}_l^T}}. \tag{5–13}$$

With these preliminaries in mind, we define our model relative to the reference category $k = R^{1\to 2}$. The full model is specified by

$$R_{ijw} | \boldsymbol{\pi}_{ijw} \sim Categorical(\boldsymbol{\pi}_{ijw}), \tag{5–14a}$$

$$\boldsymbol{\pi}_{ijw} = (\pi_{ijw}^{R^{1\to 2}}, ..., \pi_{ijw}^{R^{1\not\leftrightarrow 2}}), \quad \sum_l \pi_{ijw}^l = 1, \tag{5–14b}$$

$$log \frac{\pi_{ijw}^r}{\pi_{ijw}^{R^{1\to 2}}} = \tilde{\alpha}^r + \tilde{\beta}^r \times t_{ij}^E + \sum_{l=1}^{w} \tilde{\delta}_l, \tag{5–14c}$$

$$\tilde{\alpha}^r \sim \mathcal{N}(0, 100), \tag{5–14d}$$

$$\tilde{\beta}^r \sim \mathcal{N}(0, 10),\ r \in \{R^{2\to 1}, R^{1\leftrightarrow 2}, R^{1\sqcup 2}, R^{1\not\leftrightarrow 2}\}, \tag{5–14e}$$

$$\tilde{\delta}_1 \sim \mathcal{N}(0, 100), \tag{5–14f}$$

$$\tilde{\delta}_w | \tilde{\delta}_{w-1} \sim \mathcal{N}(\tilde{\delta}_{w-1}, (\tilde{\sigma}_\delta)^2), \quad w \in \{2, ..., 335\} \tag{5–14g}$$

$$\tilde{\sigma}_\delta \sim C^+(0, 1). \tag{5–14h}$$

The probability for any phylogeny branches of ij to form category r depends on which category is considered, the genomics region where the phylogeny was reconstructed and the couple's time elapsed. A constant $\tilde{\alpha}^r$ represents the intercept of the log odds of category r relative to the correct category, referred to as the baseline category, for all windows and time elapsed. The log odds dependence of category r on time elapsed is assumed to be linear, proportional to

$\tilde{\beta}^r$. We choose the same autocorrelation structure between windows as in (5–5), by using 335 increments $\tilde{\delta}_w$, for each window $w \in \{1, ..., 335\}$. From Figure 11, we assume that the window effects on the log odds are similar for all categories relative to the baseline. While, from Figure 12, every category relative to the correct one, behave differently. The prior densities are chosen for the same reason that those in the distance model.

5.3.4 Model fit

Model (5–14) is fit with Stan version 2.18.1, using 3 Hamiltonian Monte Carlo chains with 30,000 iterations each, of which the first 2,000 iterations are considered as burn-in. There are no problems in convergence and mixing; the minimum and maximum effective sample sizes are respectively 516 and 129,233. Table 4 shows the median and 95% credible intervals for the marginal posterior distributions of all model parameters except the window random effects. The categories that include significant coefficients can be interpreted in terms of how much they increase or decrease the odds ratios with respect to the correct direction ($R^{1\to 2}$), which was taken as the baseline category. If the intercept $\tilde{\alpha}^r$ is no significantly different from zero, the probability for pathogens of two connected couples with equal time elapsed on the same window, to form category r or the correct topology is the same. While the median of categories 'ancestral in incorrect direction' and 'intermingled' are negative, their regression coefficient is not significantly different from zero. This result is not coherent with the theory and is likely due to incorrect phylogeny reconstruction on windows with few reads observations. The interpretation of the 'sibling' intercept is that, the relative probability of the phylogenetic branches of a connected couples to form a 'sibling' rather than a the correct topology is expected to be 14.21% [4.12%, 52.91%] ($exp(\tilde{\alpha}^r)$) lower, with window coordinate and time elapsed held fixed. The time elapsed effect, $\tilde{\beta}^r$, corresponds to the marginal rate of change of the relative log odds between r against baseline category for every additional year elapsed, at the same window coordinate. As we expected from Figure 12, only the coefficient of the category 'ancestral in incorrect direction' is significant, and negative. Specifically, the relative odds ratio of the incorrect direction for a one-year increase of time elapsed is expected to decrease by a factor of 0.9590 [0.9234, 0.9949] ($exp(\tilde{\beta}_r)$), with window coordinate held fixed. Similar estimates are obtained in alternative models in Appendix C. The posterior distribution of δ_w describes the changes in the log odds from window $w-1$ to window w for all category relative to the baseline with fixed time elapsed. We look at the posterior distribution of

$$\tilde{\Delta}_w = \sum_{l=1}^{w} \tilde{\delta}_l, \tag{5–15}$$

which quantifies the difference in the log odds of all categories compared to the baseline in window w, for time elapsed held fixed. Figure 13 presents the estimated marginal posterior distributions of $\tilde{\Delta}_w$. We observe wider confidence intervals for windows without observation. A positive $\tilde{\Delta}_w$ indicates that, at window w, the odds for connected couple's branches to fall in an incorrect category rather than the correct one is greater than the baseline odds. This is true at genomics coordinates [2100 - 2275], [5250 - 5900], [6525 - 7200] and [8925 - 9275].

r	$R^{2\to 1}$		$R^{1\leftrightarrow 2}$		$R^{1\sqcup 2}$		$R^{1\not\leftarrow 2}$	
Parameter	Median	Credible interval	Median	Credible interval	Median	Credible interval	Median	Credible interval
$\tilde{\alpha}^r$	-0.3394	[-1.4575, 0.8751]	-0.8232	[-1.9724, 0.4090]	-1.9510	[-3.1896, -0.6367]	-5.3447	[-7.7865, -3.2624]
$exp(\tilde{\alpha}^r)$	0.7122	[0.2328, 2.3992]	0.4390	[0.1391, 1.5053]	0.1421	[0.0412, 0.5291]	0.0048	[0.0004, 0.0383]
$\tilde{\beta}^r$	-0.0419	[-0.0797, -0.0051]	-0.0287	[-0.0706, 0.0118]	0.0030	[-0.0519, 0.0564]	0.1184	[-0.0020, 0.2449]
$exp(\tilde{\beta}^r)$	0.9590	[0.9234, 0.9949]	0.9717	[0.9319, 1.0119]	1.0030	[0.9495, 1.0580]	1.1256	[0.9980, 1.2775]
$\tilde{\sigma}_\delta$	0.0567 [0.0237, 0.1115]							

Table 7: Median and credible interval of model (5–14) **parameters.** We fit model (5–14) using 3 Hamiltonian Monte Carlo chains. Each iteration (excluding burn-in) is considered as observation of the posterior distribution for each parameter. Taking into account all these observations, a marginal empirical posterior distribution can be formed for each parameter. We report the median and 95% confidence interval of all model parameters except the window random effects. Additionally, we report the median and 95% confidence interval of the exponential parameter observations.

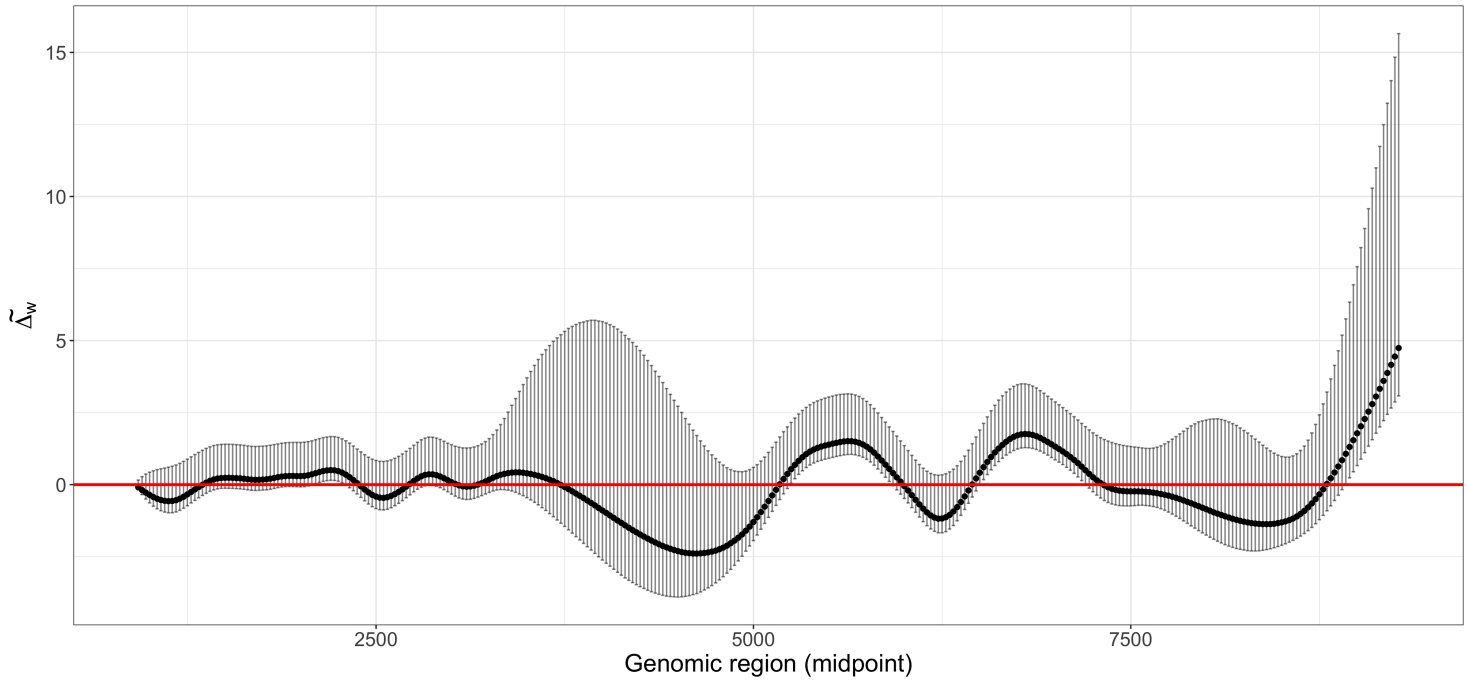


Figure 13: Marginal posterior distribution of $\tilde{\Delta}_w$. We fit model 5–14 using 3 Hamiltonian Monte Carlo chains. Each iteration (excluding burn-in) is considered as observation of the posterior distribution. Taking into account all these observations, a marginal empirical posterior distribution can be formed for $\tilde{\Delta}_w$ defined in (5–15). This is the difference in the expected log-odds in window w, compared to the average response. Shown are the 95% credibility intervals (lines) and median (dots). If a credible interval does not lie on the red line, the expected log-odds on the associated window is significantly different from the baseline response at the 5 % level.

5.4 Transmission chain prior

We now develop the prior distribution that we associate to transmission chain $\boldsymbol{T}$. Throughout, we consider a transmission chain comprising N individuals. Also recall from (2–2) that $\boldsymbol{T}$ is composed of the N-dimensional vector of sources $\boldsymbol{S}$ and the N-dimensional vector of infection times, $\boldsymbol{t}^I$. We will use daily units for computational ease, so that computations involve summation rather than integration. Without loss of generality, the first day on which any individual in the chain could have been infected is set to 0. It is found by subtracting eight years to the earliest time of sampling. Indeed, Anderson and Medley (1988) estimated to eight years the period between infection and death without ART. The end of the period corresponds to the last sampling time among the N individuals of the transmission chain. Finally, recall from (5–2) that all parameters relating to transmission to the next individual are subsumed in the vector $\boldsymbol{\theta}$.

5.4.1 Iterative construction

The prior distribution is constructed iteratively, by starting with one individual that is infected at the beginning, who initiates spread to $N-1$ further individuals. Time progresses in daily steps from the infection day of the index case to the sampling day of the last individual sampled. The jth individual in the chain can only have been infected by the $j-1$ individuals that were previously infected. Thus

$$\begin{aligned} p(\boldsymbol{T}|\boldsymbol{\theta}) &= p(\boldsymbol{t}^I, \boldsymbol{S}|\boldsymbol{\theta}) \\ &= \prod_{j=2}^{N} p(t_j^I, S_j | t_{1:j-1}^I, S_{1:j-1}, \boldsymbol{\theta}) p(t_1^I, S_1) \\ &= \prod_{j=2}^{N} p(S_j | t_{1:j}^I, \boldsymbol{\theta}) p(t_j^I | t_{1:j-1}^I, \boldsymbol{\theta}) p(t_1^I, S_1), \end{aligned} \tag{5–16}$$

where the index j is sorted with respect to increasing infection day, $t_{j-1}^I < t_j^I$.

5.4.2 Hazard model

To specify the probabilities of the next infection date $p(t_j^I | t_{1:j-1}^I, \boldsymbol{\theta})$ and the next source case $p(S_j | t_{1:j}^I, \boldsymbol{\theta})$ in (5–16), we adopt a standard approach that is based on transmission hazards, defined as the instantaneous force of infection from infected person i to uninfected person j given that no infection occurred up to that time (Cauchemez and Ferguson, 2012). It is standard to specify transmission hazards in continuous time. Let us denote by t_j^{cont} the continuous time of infection of individual j. At any continuous time point u during day t (i.e. $u \in [t, t+1)$), we denote the instantaneous hazard exerted by i on individual j by $\lambda_{i\to j}^*(t_j^{cont} = u | \boldsymbol{\theta}, t_i^I)$. The transmission hazard is assumed to be constant for any $u \in [t, t+1)$:

$$\lambda_{i\to j}^*(t_j^{cont} = u | \boldsymbol{\theta}, t_i^I) = \lambda_{i\to j}(t_j^I = t | \boldsymbol{\theta}, t_i^I) \text{ for } t \le u < t+1, \tag{5–17}$$

so the continuous-time transmission hazard in our model is just a step function with daily steps, as in Cauchemez and Ferguson (2012).

Transmission hazards can be specified in a variety of ways, though the following Cox proportional hazards decomposition is quite common. We consider

$$\lambda_{i\to j}(t_j^I = t|\boldsymbol{\theta}, t_i^I) = \mathrm{w}(\underbrace{t - t_i^I}_{\Delta_i})exp(\theta_0), \tag{5–18}$$

where w is the so-called generation time distribution and Δ_i is the time since infection of source case i. The generation time distribution w(.) specifies the degree of infectiousness of case i after $\Delta_i = t - t_i^I$ days since infection. We assume that an individual cannot infect another individual on the day of his own infection. Moreover, we use the generation time distribution derived in Hollingsworth et al. (2008), which is based on an in-depth analysis of HIV-1-infected couples in the same study population. In Hollingsworth et al. (2008), the generation time distribution is a step function of four time periods. The first period is called primary infection and is associated with a high rate of transmission. The second period is called the asymptotic period during which the rate of transmission is approximately constant at a lower level. The third period is called the AIDS phase of systematic immune deficiency during which the rate of transmission is again high. The fourth period is immediately before death, during which no transmission occurs because the infected individual is so unhealthy that sexual activity is not considered to occur. The relative transmission rates and durations of each of the three infectious periods were estimated by Hollingsworth et al. (2008). Then, Bellan et al. (2015) revised the estimates associated with the acute phase. Lastly, the entire period from initial infection with HIV-1, to the development of symptoms of AIDS, until death is estimated to be around eight years (Anderson and Medley, 1988). The generation time function we use is fully specified through

$$\mathrm{w}(\Delta t) = \begin{cases} 0 & \text{if } \Delta t = 0, \\ 0.53/365 & \text{if } \Delta t \leq 87, \\ 0.106/365 & \text{if } 87 < \Delta t \leq 2,350, \\ 0.760/365 & \text{if } 2,350 < \Delta t \leq 2,620, \\ 0 & \text{if } \Delta t > 2,620. \end{cases} \tag{5–19}$$

The free parameters of our hazard model are thus just the baseline transmission rate θ_0, $\boldsymbol{\theta} = \{\theta_0\} \in \mathbb{R}$. To avoid too short disease free survival periods, we incorporate informative prior knowledge on the transmission hazard through the baseline transmission rate. Specifically, we choose a prior centered at -1 on the θ_0 and a tight variance,

$$\theta_0 \sim \mathcal{N}(-1, 1) \tag{5–20}$$

5.4.3 Probability of infection of the jth individual on day t_j^I

We first describe the probability that j gets infected at any continuous time point between t_j^I and $t_j^I + 1$ (24 hours),

$$p(t_j^I | t_{1:j-1}^I, \boldsymbol{\theta}), \tag{5–21}$$

as also presented in Cauchemez and Ferguson (2012). It is convenient to consider the force of infection exerted on individual j on day t by summing the transmission hazards of the $j-1$ previously infected individuals:

$$\lambda_j(t|\boldsymbol{\theta}, t^I_{1:j-1}) = \sum_{i:t_i^I<t} \lambda_{i\to j}(t|\boldsymbol{\theta}, t_i^I). \tag{5–22}$$

In continuous time, the force of infection on individual j at time u is given by:

$$\lambda^*_j(u|\boldsymbol{\theta}, t^I_{1:j-1}) = \sum_{i:t_i^I<\lfloor u\rfloor} \lambda^*_{i\to j}(u|\boldsymbol{\theta}, t_i^I). \tag{5–23}$$

where $f(x) = \lfloor x \rfloor$ is the floor function. Indices are sorted with respect to increasing infection time, such that $\{i : t_i^I < \lfloor u \rfloor\} = \{1, ..., j-1\}$. For readability, we drop the condition on $\boldsymbol{\theta}$ and cases' day of infection $t^I_{1:j-1}$ in the following.

The continuous-time transmission hazard is the instantaneous probability of infection at time u conditional on healthiness until time u. Considering a small time interval du,

$$\begin{aligned}\lambda^*_j(u) &= \lim_{du\to 0} \frac{\Pr(u < t_j^{cont} \le u + du | t_j^{cont} > u)}{du} = \lim_{du\to 0} \frac{\Pr(u < t_j^{cont} \le u + du)}{du \times S(u)} = \frac{f(u)}{S(u)} \\ &= -\frac{S'(u)}{S(u)} = -\frac{d}{du}\ln(S(u)),\end{aligned} \tag{5–24}$$

where $S(u) = p(t_j^{cont} > u)$ is the survival function. We thus obtain directly from (5–24)

$$S(u) = exp\Big(-\int_0^u \lambda^*_j(x)dx\Big). \tag{5–25}$$

Then, the probability that j gets infected between t and $t+1$ (24 hours) is

$$\begin{aligned}Pr(t \le t_j^{cont} < t+1) &= Pr(t_j^{cont} < t+1) - Pr(t_j^{cont} \le t) \\ &= \left\{1 - exp\left(-\int_0^{t+1} \lambda^*_j(u)du\right)\right\} - \left\{1 - exp\left(-\int_0^{t} \lambda^*_j(u)du\right)\right\} \\ &= -exp\left(-\int_0^{t} \lambda^*_j(u)du - \int_t^{t+1} \lambda^*_j(u)du\right) + exp\left(-\int_0^{t} \lambda^*_j(u)du\right) \\ &= \left\{1 - exp\left(-\int_t^{t+1} \lambda^*_j(u)du\right)\right\} \times exp\left(-\int_0^{t} \lambda^*_j(u)du\right) \\ &= \left\{1 - exp\left(-\int_t^{t+1} \lambda^*_j(u)du\right)\right\} \times exp\left(-\sum_{d=0}^{t-1}\int_d^{d+1} \lambda^*_j(u)du\right) \\ &= \{1 - exp\left(-\lambda_j(t)\right)\} \times exp\left(-\sum_{d=0}^{t-1} \lambda_j(d)\right).\end{aligned} \tag{5–26}$$

In (5–26), we use condition (5–17). Lastly, under our assumptions, infection must occur

between the day of sampling and $8 \times 365 = 2,920$ days before. We obtain

$$p(t_j^I = t | t_{1:j-1}^I, \boldsymbol{\theta}) = \begin{cases} \Big\{1 - exp\big(-\lambda_j(t|t_{1:j-1}^I, \boldsymbol{\theta})\big)\Big\} & \text{if } t_j^S - 2,920 < t < t_j^S, \\ \qquad \times exp\Big(-\sum_{d=0}^{t-1} \lambda_j(d|t_{1:j-1}^I, \boldsymbol{\theta})\Big) & \\ 0 & \text{otherwise.} \end{cases} \tag{5–27}$$

5.4.4 Probability that i is the source of j on infection day t_j^I

We next describe the probability that case i is the source of j conditional on the day of infection, as in Cauchemez and Ferguson (2012). Conditional on the (continuous) time of infection t_j^{cont} of case j, we have:

$$p(S_j = i | t_{1:j-1}^I, t_j^{cont}, \boldsymbol{\theta}) = \frac{\lambda_{i \to j}(t_j^{cont}|\boldsymbol{\theta}, t_i^I)}{\sum_{k=1}^{j-1} \lambda_{k \to j}(t_j^{cont}|\boldsymbol{\theta}, t_k^I)}, \tag{5–28}$$

if $\sum_{k=1}^{j-1} \lambda_{k \to j}(t_j^{cont}|\boldsymbol{\theta}, t_k^I) > 0$, and 0 otherwise. This probability is constant for any $t_j^{cont} \in [t_j^I, t_j^I + 1)$ because of (5–17).

We can now find the probability that i is the source of j conditional on the day of infection:

$$p(S_j = i | t_{1:j-1}^I, t_j^I, \boldsymbol{\theta}) = \int_{t_j^{cont}} p(S_j = i | \boldsymbol{\theta}, t_{1:j-1}^I, t_j^{cont}) p(t_j^{cont} | t_j^I) dt_j^{cont} \tag{5–29}$$

$$= \int_{t_j^I}^{t_j^I + 1} \frac{\lambda_{i \to j}(t_j^{cont}|\boldsymbol{\theta}, t_i^I)}{\sum_{k=1}^{j-1} \lambda_{k \to j}(t_j^{cont}|\boldsymbol{\theta}, t_k^I)} p(t_j^{cont} | t_j^I) dt_j^{cont} \tag{5–30}$$

$$= \frac{\lambda_{i \to j}(t_j^I|\boldsymbol{\theta}, t_i^I)}{\sum_{k=1}^{j-1} \lambda_{k \to j}(t_j^I|\boldsymbol{\theta}, t_k^I)} \int_{t_j^I}^{t_j^I + 1} p(t_j^{cont} | t_j^I) dt_j^{cont} \tag{5–31}$$

$$= \frac{\lambda_{i \to j}(t_j|\boldsymbol{\theta}, t_i^I)}{\sum_{k=1}^{j-1} \lambda_{k \to j}(t_j^I|\boldsymbol{\theta}, t_k^I)}. \tag{5–32}$$

if $\sum_{k=1}^{j-1} \lambda_{k \to j}(t_j^I|\boldsymbol{\theta}, t_k^I) > 0$, and 0 otherwise. In (5–31) we use the fact that probability (5–28) is constant for any $t_j^{cont} \in [t_j^I, t_j^I + 1)$. In (5–30), t_j^{cont}, given t_j^I, is defined only on $[t_j^I, t_j^I + 1)$, therefore the density on this region must integrate to one in (5–32).

5.4.5 Probability that j is the index case

The source of the first case is external, and so we fix S_1 to 0, such that $p(t_1^I, S_1) = p(t_1^I)$. Moreover, we do not have particular knowledge on the day of infection of the first infected individual and place a discrete uniform prior distribution from the beginning to the end of the observation period. Therefore $p(t_1^I)$ is a constant.

6 Numerical inference

6.1 Strategy

We built a Metropolis-Hasting Markov Chain Monte Carlo (MCMC) algorithm (Metropolis et al., 1953, Hastings, 1970) for obtaining a sequence of random samples from the posterior distribution (5–2) that for ease of reference we give here again,

$$p(\boldsymbol{T}, \boldsymbol{\theta}, \boldsymbol{\psi}^M | \boldsymbol{x}) \propto p(\boldsymbol{x}|\boldsymbol{T}, \boldsymbol{\psi}^M) p(\boldsymbol{T}|\boldsymbol{\theta}) p(\boldsymbol{\theta}) p(\boldsymbol{\psi}^M). \tag{6–1}$$

The likelihood $p(\boldsymbol{x}|\boldsymbol{T}, \boldsymbol{\psi}^M)$ was specified in (5–5) and (5–14), the chain prior $p(\boldsymbol{T}|\boldsymbol{\theta})$ in (5–16) and the epidemiological prior $p(\boldsymbol{\theta})$ in (5–20). For now, we are fixing $\boldsymbol{\psi}$ at the empirical median obtained by the informative empirical Bayes analysis, denoted by $\boldsymbol{\psi}^M$. Therefore $p(\boldsymbol{\psi}^M)$ stays constant. The time elapsed, on which the likelihood depends on, is found by dividing by 365 the days elapsed. The MCMC presented below is run on networks presented in Subsection 3. In Subsection 6.2, we present the parameters initialization for the MCMC, and in Subsection 6.3, we describe the MCMC proposal distributions for all parameters, and corresponding Metropolis-Hastings ratios.

6.2 Starting values

6.2.1 Starting values for $\boldsymbol{\theta}$

The starting value of the epidemiological parameter is set to its prior mean, i.e. -1.

6.2.2 Starting values for $\boldsymbol{S}$ and $\boldsymbol{t}^I$

The source variable is initialized as the most likely transmission chain within a given network, according to the calculations in Ratmann et al. (2019).

To simplify book-keeping, we associate each individual in the chain with an order that corresponds conceptually to his generation in the transmission chain. That is, a source is associated with a smaller order than his recipient. Note that the order does not exactly correspond to a generation, but interval of orders do. The index case has the first order ($= 1$) (i.e., $\{1\}$ is the set of individual of the first generation). Descendants of the next generation are sorted as follow. The descendant with the earliest ordered source is considered first and assign an order. His order is equal to an increment of one from the last order attributed to any individual. Then, descendants are iteratively considered with respect to the increasing order of their sources, until the descendant of the latest ordered source. Thereafter, we change generation. If several individuals share the same source, subsequent orders are randomly assigned. Let element $L_j \in \{1, ..., N\}$ denotes the order assigned to individual j,

$$L_j = \begin{cases} L_{S_j} + l + 1 & \text{if } j \text{ is the only recipient of } S_j, \\ L_{S_j} + l + r + 1 & \text{if } S_j \text{ has } r > 1 \text{ recipients that have already been assigned an order,} \end{cases} \tag{6–2}$$

where l is the last order assigned. With this book-keeping, it is more straightforward to initialize the day of infection. Indeed, to this aim, we iteratively consider individuals based

on this order, such that $t^I_{S_j}$ will always have been initialized before t^I_j.

The day of infection of individual j must be subsequent to that of his source S_j and, under our assumptions, less than eight years (2,920 days) after the source's infection. Moreover individuals cannot transmit on the same day as their infection i.e.,

$$t^I_{S_j} + 1 \leq t^I_j \leq t^I_{S_j} + 2,920. \tag{6–3}$$

Further, it cannot have happened 2,920 days before his sampling day and is no later than his sampling day, i.e.,

$$t^S_j - 2,920 \leq t^I_j \leq t^S_j. \tag{6–4}$$

Every day in between is assumed to have the same probability mass, so that the initialized variable,

$$t^I_j \sim \mathcal{U}\Big(\max\big\{t^I_{S_j} + 1, t^S_j - 2,920\big\},\ \min\big\{t^I_{S_j} + 2,920, t^S_j\big\}\Big). \tag{6–5}$$

$\mathcal{U}(a, b)$ is the Discrete Uniform distribution with density

$$\mathcal{U}(x; a, b) = \frac{1}{b - a + 1}. \tag{6–6}$$

6.3 MCMC updates

6.3.1 MCMC updates for $\boldsymbol{\theta}$

We use a Gaussian random walk to update the component of $\boldsymbol{\theta}$, i.e.

$$q(\theta^*_0|\theta_0) \sim \mathcal{N}\big(\theta_0, \sigma^2_{\theta_0}\big), \tag{6–7}$$

where * denotes a proposed value. The proposal's variances is adapted at every iteration using the Robust adaptive Metropolis algorithm update proposed in Vihola (2012). Cancelling appropriate term in (6–1), the corresponding Metropolis-Hastings ratio is

$$\min\left\{1, \frac{p(\boldsymbol{T}|\theta^*_0)p(\theta^*_0)}{p(\boldsymbol{T}|\theta_0)p(\theta_0)}\right\},, \tag{6–8}$$

because the ratio of the proposal densities cancels.

6.3.2 Joint updates of t^I_j and $\boldsymbol{S}$

To update the transmission tree, we consider in turn each individual ordered by increasing time of infection found at the previous iteration. For every individual considered, the algorithm updates his day of infection and the entire source vector. Suppose we updating t^I_j and $\boldsymbol{S}$ of the jth individual in the chain. Firstly a new day of infection is proposed, and depending on the later, the source vector is changed. For ease of notation define the truncated normal

distribution

$$\mathrm{TN}(x;\mu,\sigma^2,a,b) = \frac{\phi(\frac{x-\mu}{\sigma})}{\sigma\left(\Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})\right)}, \tag{6–9}$$

for $x \in [a,b]$. Here $\phi(\cdot)$ is the probability density function of the standard normal distribution and $\Phi(\cdot)$ is its cumulative distribution function. The proposal on the infection day of individual j is the floor function of a truncated Gaussian random walk,

$$q(t_j^{I*}|t_j^I) = \int_{t_j^{I*}}^{t_j^{I*}+1} q(u|t_j^I)du \tag{6–10}$$

$$= \int_{t_*^{I*}}^{t_j^{I*}+1} \mathrm{TN}\Big(u; t_j^I,\ \sigma_j^2,\ t_j^S - 2,920,\ t_j^S\Big)du, \tag{6–11}$$

where the left-hand constraint represents the first day of infection for which j would have been alive to get his pathogen sampled, and the right-hand constraint indicates that j cannot have been infected after being sampled. This proposal allows to reach any state in the variable space. There are distinct proposal variances σ_j^2 for each individual, and we will adapt them using the Robust adaptive Metropolis algorithm proposed in Vihola (2012).

The source vector is then modified according to the proposed day of infection. Specifically, the transmission events may be adapted if the proposed day of infection of j is not coherent with the day of infection of his ancestors and recipients (i.e., proposed infection before the infection of an ancestor, or after a recipient). Ancestors of j comprise his source, his great source (the source of his source - S_{S_j}), his great-grand-source (the source of the source of his source - $S_{S_{S_j}}$), etc. until the index case. We say that an individual is 'swapped' with another in the transmission linkage when their respective source and recipients are exchanged. If the proposed day of infection of j is later than the infection of his first recipient (i.e., with earliest day of infection), we swap j with him. Only descendants of this recipient that are infected after the proposed day of infection will become recipients of j. If the proposed day of infection of j is earlier than that of his source, we swap him with the ancestor that has the earliest larger time of infection than the proposed day of j. That is, the ancestor of this ancestor is infected before j's proposed day. Otherwise, the order remains the same. We present the moves only on the source of j depending on the proposed day of infection,

$$S_j^* = \begin{cases} S_j & \text{if } t_{S_j}^I \le t_j^{I*} \le t_{r_j^1}^I, \\ r_j^1 & \text{if } t_j^{I*} > t_{r_j^1}^I, \\ S_{S_j} & \text{if } t_{S_{S_j}}^I \le t_j^{I*} < t_{S_j}^I, \\ S_{S_{S_j}} & \text{if } t_{S_{S_{S_j}}}^I \le t_j^{I*} < t_{S_{S_j}}^I, \\ \quad\vdots & \\ 0 & \text{if } t_j^{I*} < t_1^I. \end{cases} \tag{6–12}$$

Here r_j^1 denotes the first recipient of j (i.e., with the earliest day of infection), 1 is the

index case and 0 is the external source. One should note that, when j is swapped with another individual, their recipients and sources are exchanged such that several entries on the source vector are changed. An example of these moves according to an illustrated timeline is presented in Figure 14. Note that in case of equality between the proposed day of infection and the day of infection of the source (or proposed source), the move will have a null probability to be accepted because of the assumptions in (5–19).

The joint proposal on the transmission tree for individual j takes the form

$$\begin{aligned} q(t_j^{I*}, \boldsymbol{S}^* | \boldsymbol{t}^I, \boldsymbol{S}) &= q(\boldsymbol{S}^* | t_j^{I*}, \boldsymbol{t}^I, \boldsymbol{S}) q(t_j^{I*} | t_j^I) \\ &= \begin{cases} q(t_j^{I*} | t_j^I) & \text{if } S_j^* \in \text{(6–12)}, \\ 0 & \text{otherwise.} \end{cases} \end{aligned} \tag{6–13}$$

Cancelling appropriate term in (6–1), the corresponding Metropolis-Hastings ratio is

$$\min\left\{1, \frac{p(\boldsymbol{x}|\boldsymbol{t}^I_{1:N\setminus j}, t_j^{I*}, \boldsymbol{S}^*, \boldsymbol{\psi}^M) p(\boldsymbol{t}^I_{1:N\setminus j}, t_j^{I*}, \boldsymbol{S}^* | \boldsymbol{\theta}) q(t_j^I, \boldsymbol{S} | \boldsymbol{t}^I_{1:N\setminus j}, t_j^{I*}, \boldsymbol{S}^*)}{p(\boldsymbol{x}|\boldsymbol{t}^I_{1:N}, \boldsymbol{S}, \boldsymbol{\psi}^M) p(\boldsymbol{t}^I_{1:N}, \boldsymbol{S} | \boldsymbol{\theta}) q(t_j^{I*}, \boldsymbol{S}^* | \boldsymbol{t}^I_{1:N}, \boldsymbol{S})}\right\}. \tag{6–14}$$

6.4 Structure of the MCMC algorithm

We consider each variable in turn such that one sweep through all variables is done in $1 + N$ updates. There is 1 update for θ_0. Consecutively, there are N updates on $\{t_1^I, \boldsymbol{S}\}$, ..., $\{t_N^I, \boldsymbol{S}\}$. For this, individuals are ordered by increasing day of infection from the last day of infection vector completely updated (i.e., from the previous iteration). The Metroplis-Hastings MCMC move takes the form:

Structure of the MCMC algorithm

1. Move $\boldsymbol{\theta}$

Propose θ_0^* using (6–7).

Accept θ_0^* with probability (6–8).

If accepted, set $\theta_0 = \theta_0^*$

2. Move $\boldsymbol{T}$

Find indices by infection order $\{\sigma_1, ..., \sigma_N \mid t^I(\sigma_1) < ... < t^I(\sigma_N), \boldsymbol{\sigma} \in \mathcal{S}_N\}$, where $\mathcal{S}_N$ is the symmetric group.

For $j = \{\sigma_1, ..., \sigma_N\}$

Propose $\{t_j^{I^*}, \boldsymbol{S}^*\}$ using (6–13).

Accept $\{t_j^{I^*}, \boldsymbol{S}^*\}$ with probability (6–14).

If accepted, set $\{t_j^I, \boldsymbol{S}\} = \{t_j^{I^*}, \boldsymbol{S}^*\}$

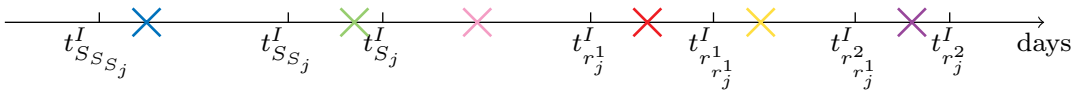


Figure 14: Moves on the source vector depending on the proposed day of infection. In this example, we consider five proposed days of infection for individual j and present the corresponding displacements applied to the source vector. Suppose that individual j has, according to the previously accepted source vector, two recipients r^1_j and r^2_j. The former himself has two descendants $r^1_{r^1_j}$ and $r^2_{r^1_j}$. Moreover j has three ancestors, his source S_j, his great source S_{S_j} and his great-grand-source $S_{S_{S_j}}$. The later is the index case. We write t^I_i, the previously accepted day of infection of individual i. At the red, gold and purple crosses, the proposed day of infection is later than the day of infection of the first recipient r^1_j. j is exchanged with r^1_j. It implies that the source of j becomes r^1_j, the source of r^1_j becomes S_j and the source of r^2_j becomes r^1_j. At the red cross, all descendants of r^1_j are infected after the proposed day, so the source of $r^2_{r^1_j}$ and $r^1_{r^1_j}$ becomes j. At the gold cross, one of the descendants of r^1_j is infected before the proposed day, so the source of $r^1_{r^1_j}$ stays r^1_j and only $r^2_{r^1_j}$ becomes recipient of j. At the purple cross, all r^1_j's descendants are infected before the proposed day, therefore j is not assigned to be the source of anyone. At the green and blue crosses, the proposed day of infection is earlier than that of the source of j. We exchange it with the nearest ancestor with a later infection time. At the green cross, we exchange j with S_j, which implies that the source of j becomes S_{S_j}, the source of S_j becomes j and the source of r^1_j and r^2_j becomes S_j. Also, if S_j had other recipients than j, their source would become j. At the blue cross, we exchange j with S_{S_j}, which implies that the source of j becomes $S_{S_{S_j}}$, the source of S_{S_j} becomes j, the source of r^1_j and r^2_j becomes S_{S_j} and the source of S_j becomes j. Finally, at the pink cross, the day of the proposed infection is later than that of the source and before that of the first recipient. The proposed source vector remains the same.

6.5 Performance

6.5.1 Convergence and mixing

We use our MCMC algorithm of Subsection 6.4 to sample from the transmission chain posterior for each network presented in Section 3. For every network, we run 3 Metropolis-Hasting Markov Chains of 6,000 iterations with 1,000 iterations considered as burn-in.

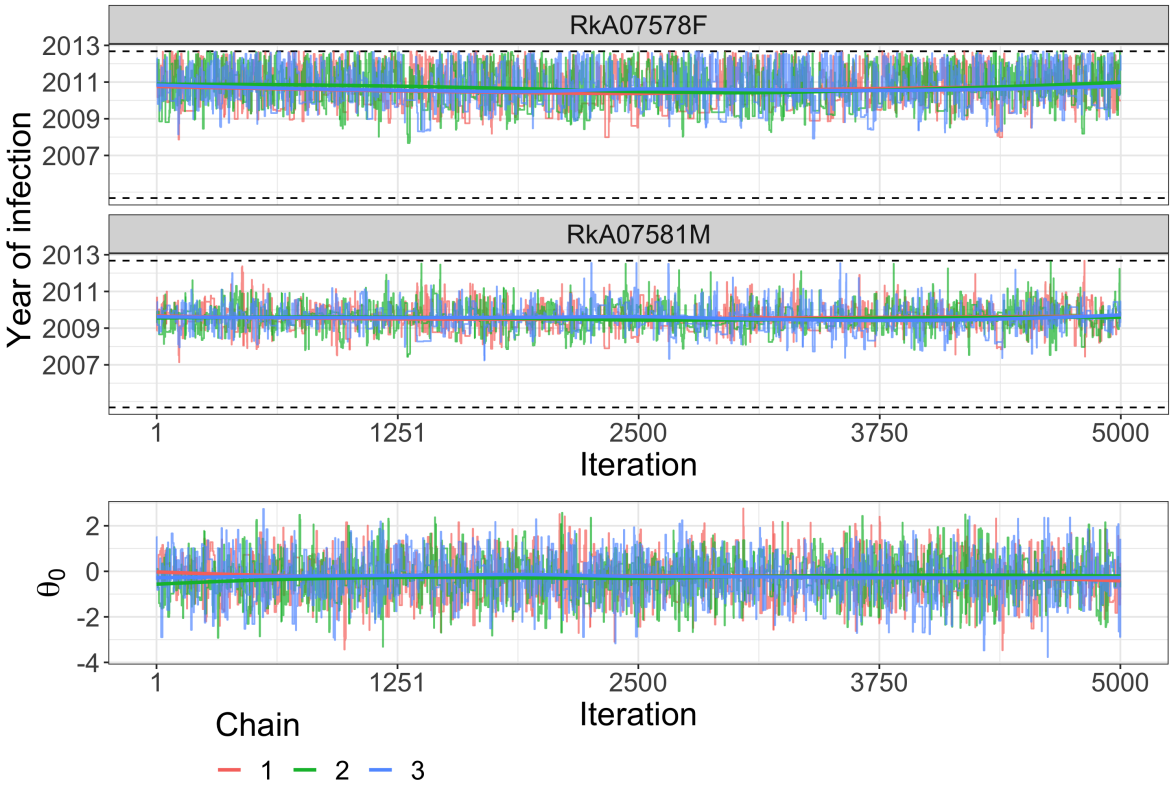
Figure 15 shows trace plots of the times of infection and the baseline transmission rate θ_0. The dashed line on the time of infection trace plot represents the limits of the latent variable (i.e., the lower limit is 8 years before time of sampling and the upper limit is the time of sampling). While we did inference on the day of infection, we map back the posterior observations to obtain the time of infection in units of year. There was no problem of convergence and mixing. The *potential scale reduction factors* (Gelman and Rubin, 1992) were between 1 and 1.01. Across all variables of all transmission chains, the minimum and maximum effective sample sizes were respectively 519 and 4,265. These two diagnostics were calculated with the coda package (Plummer et al., 2005).

6.5.2 Simulation analysis

We focus on this part on elements that influence the inferred direction of transmission (i.e., source vector variable). When we presented the summary statistics of the networks in Table 3, we noticed that all initialized pairs had a count of topology in one direction greater than in the other, which could favor the former direction in inference. Therefore, before discussing the results, we decide to simulate phylogenetic relationship data to observe if the updates proposed by the MCMC allowed individuals to move in the transmission chain.

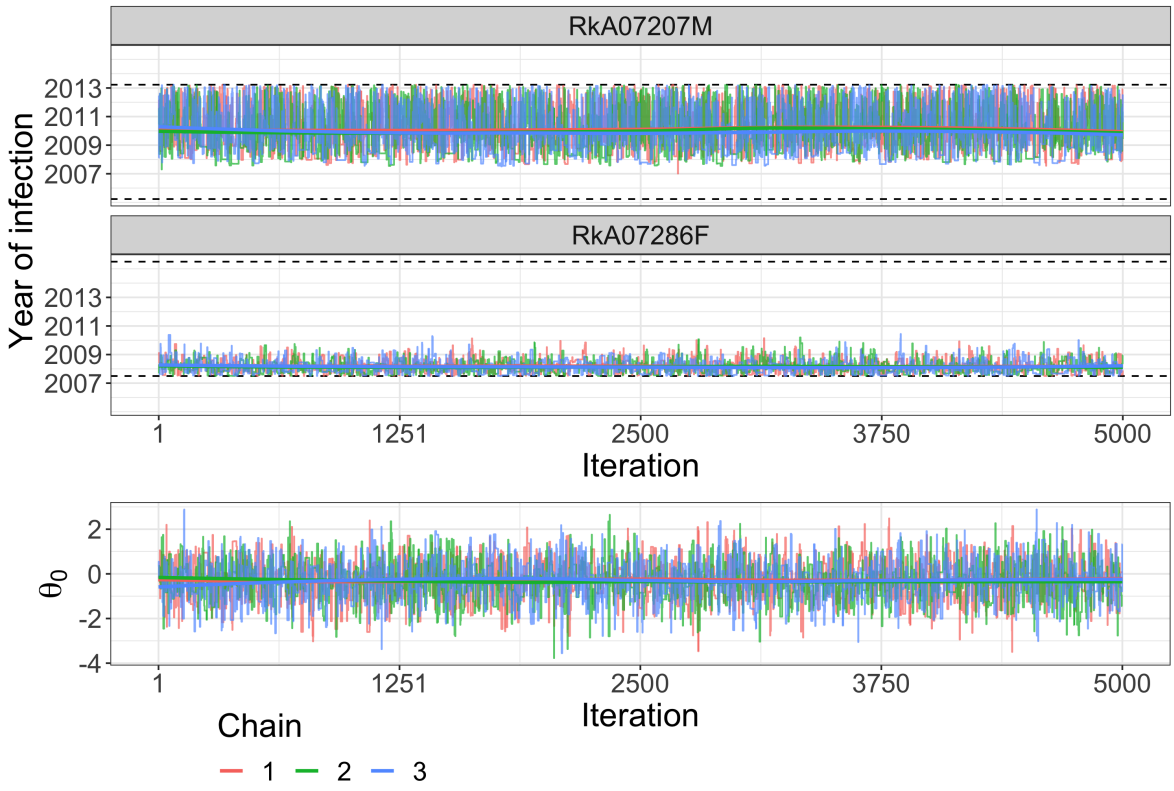
Let us remember that the proposed epidemiological direction from $\boldsymbol{T}$ is weighted by the likelihood depending on the topology observed in the same direction, the ones on the other direction, and the location of the latter (Table 7 and Figure 13). Specifically, observing a phylogenetic direction opposite to the given epidemiological one leads to a smaller likelihood than observing one in the same direction. Concurrently, observing the coherent direction on window w for which $\tilde{\Delta}_w^M > 0$ is associated to a smaller likelihood than on regions with $\tilde{\Delta}_w^M < 0$, where $\tilde{\Delta}_w^M$ is the empirical median of $\tilde{\Delta}_w$ defined in (5–15). This parameter is difference of the log odds of any categories relative to the proposed epidemiological direction, between window w and the baseline, for time elapsed held fixed.

It is intuitive to consider a network of $N = 2$ individuals for this simulation. Indeed, in this setting and by our assumptions, one individual will inevitably transmit to the other. The change between the source and recipient is convenient for observing if our latent variable $\boldsymbol{S}$ changes along iterations. Let us investigate Network 1 involving individuals $i, j \in$ {RkA07581M, RkA07578F}. To create our simulated data-set we let the number of observation, denoted by n_x as it is (in this case, $n_x = 60$). Among them, we fix a proportion p of ancestral topology with a direction coherent to the one found by Ratmann et al. (2019) (RkA07581M $\rightarrow$ RkA07578F), and a proportion $1 - p$ in the opposite direction. We refer to the initialized direction as $M \rightarrow F$ ($M \equiv$ RkA07581M and $F \equiv$ RkA07578F). We do not need to take into account sibling, intermingled, or disconnected because they will not favor any of the direction. To avoid mixing the heterogeneity sources, we fix all the windows index to the first window for now. We start by setting $p = 0.5$. In this case, the number of ancestral phylo-



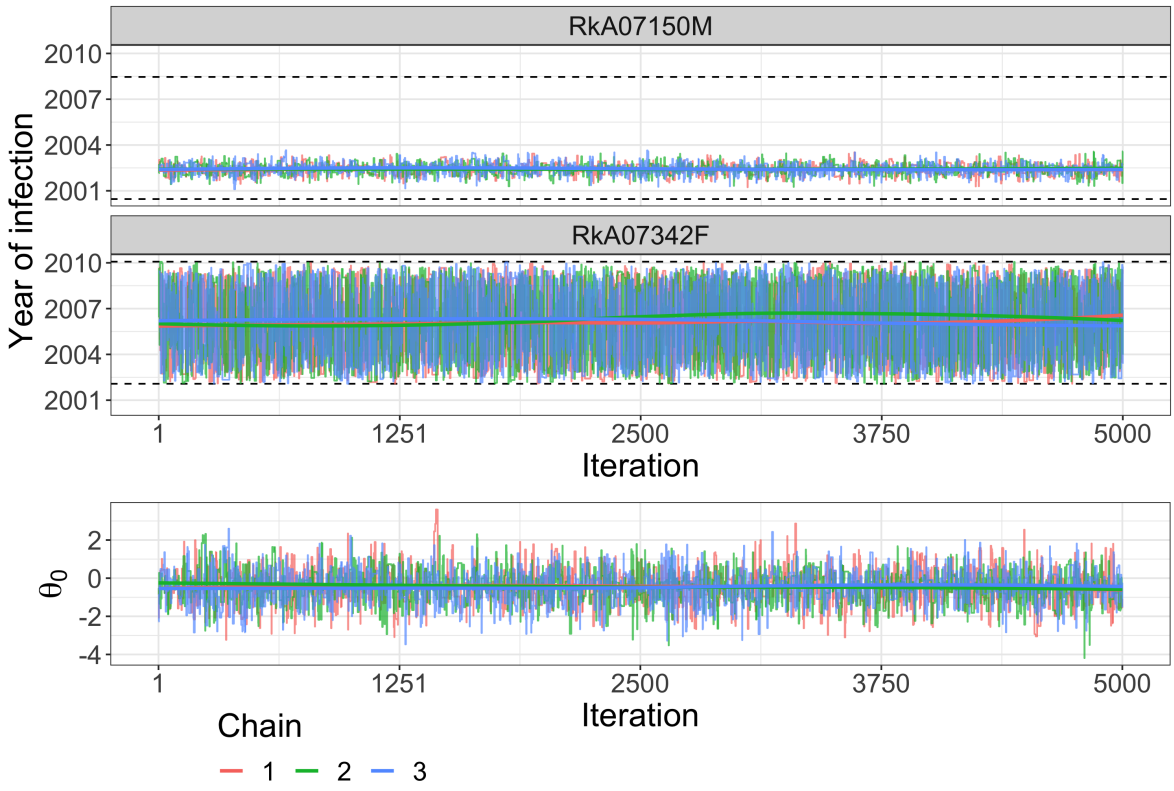
(a) Network 1, $N = 2$ individuals

Figure 15: Trace plots of t_i^I and θ_0 for 7 networks involving individuals $i \in \{1, ..., N\}$. 3 Metropolis-Hasting Markov Chains of 6,000 iterations with 1,000 iterations considered as burn-in were used to sample from the posterior distribution of the transmission chain of every network. We present the trace plot of the day of infection of every individual and the epidemiological parameter.



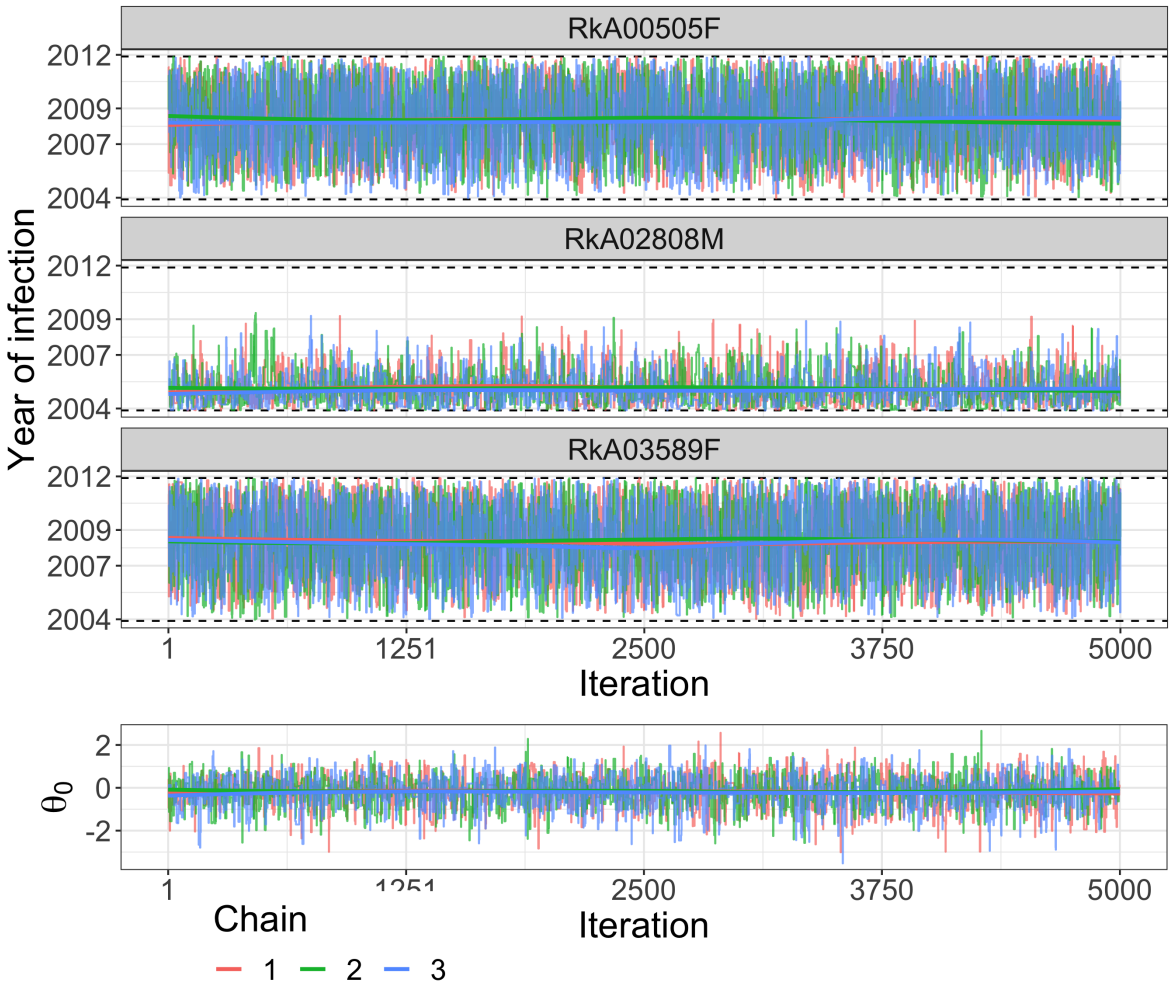
(b) Network 2, $N = 2$ individuals

Figure 15: Trace plots of t_i^I and θ_0 for 7 networks involving individuals $i \in \{1, ..., N\}$(cont.).



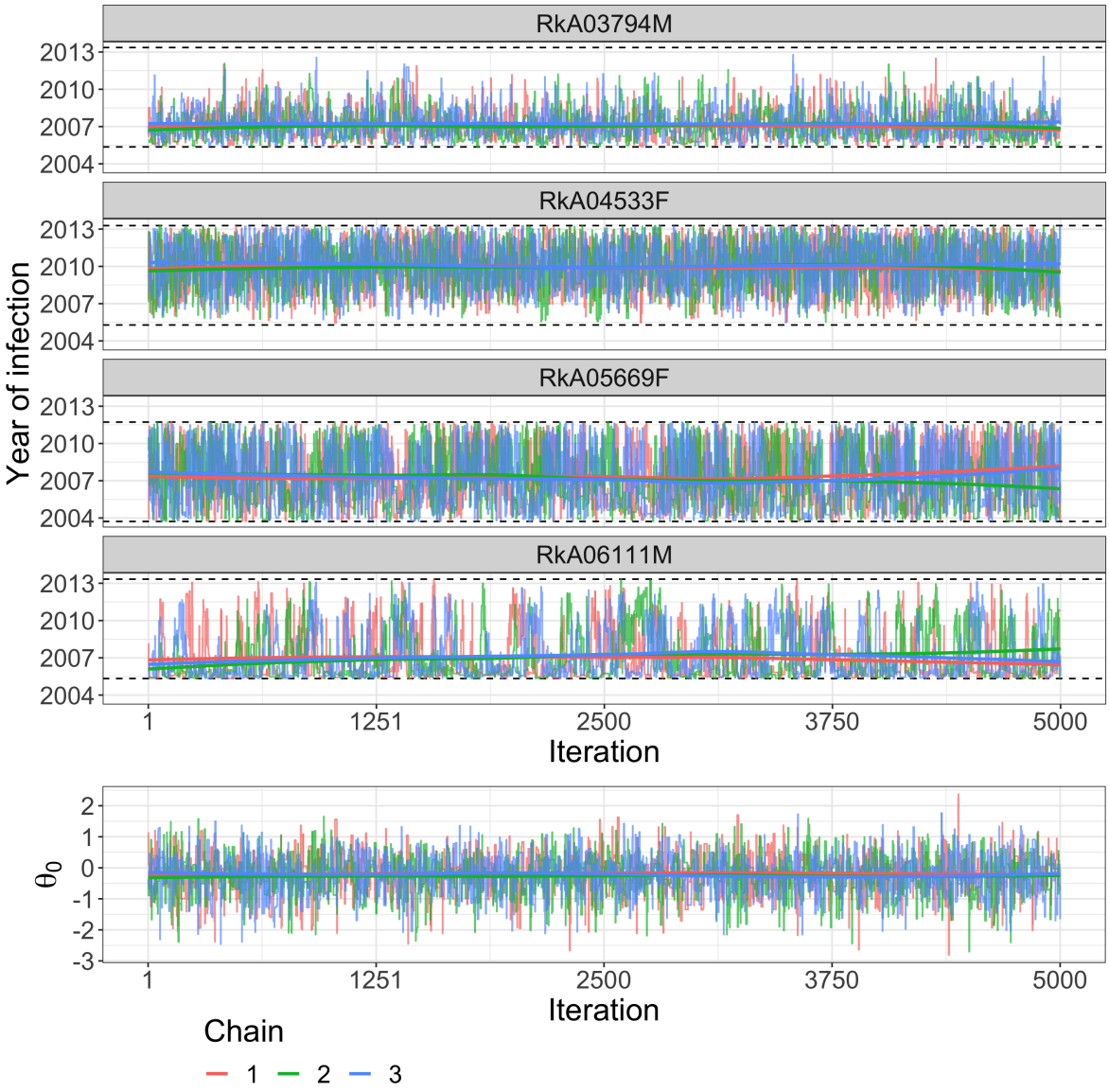
(c) Network 3, $N = 2$ individuals

Figure 15: Trace plots of t_i^I and θ_0 for 7 networks involving individuals $i \in \{1, ..., N\}$(cont.).



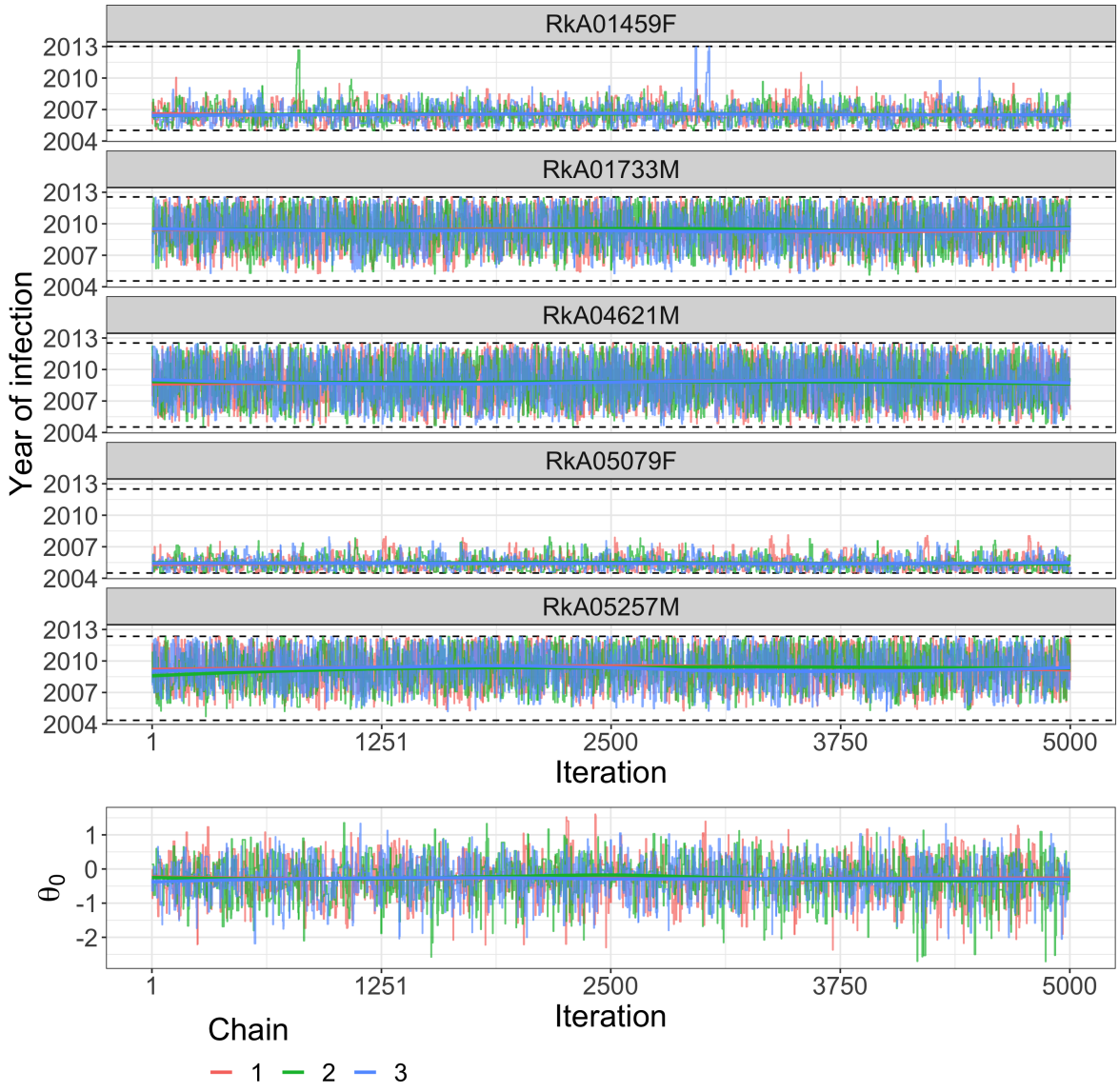
(d) Network 4, $N = 3$ individuals

Figure 15: Trace plots of t_i^I and θ_0 for 7 networks involving individuals $i \in \{1, ..., N\}$(cont.).



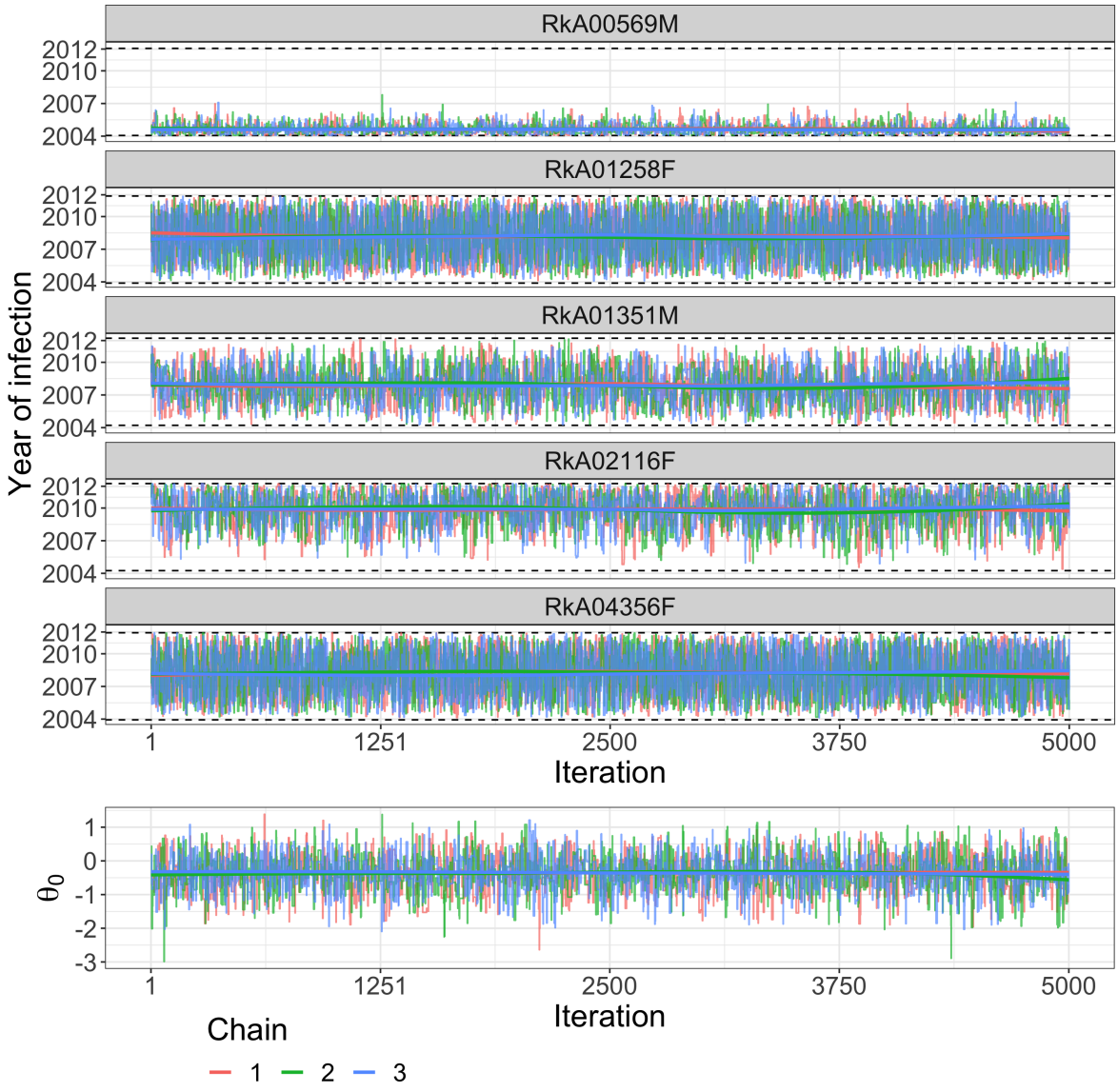
(e) Network 5, $N = 4$ individuals

Figure 15: Trace plots of t_i^I and θ_0 for 7 networks involving individuals $i \in \{1, ..., N\}$(cont.).



(f) Network 6, $N = 5$ individuals

Figure 15: Trace plots of t_i^I and θ_0 for 7 networks involving individuals $i \in \{1, ..., N\}$(cont.).



(g) Network 7, $N = 5$ individuals

Figure 15: Trace plots of t_i^I and θ_0 for 7 networks involving individuals $i \in \{1, ..., N\}$(cont.).

genetic topology in both directions is the same. Let the proportion of iteration on which M is the source be $p^{M\to F}$. The latter should, in principle, be equal to 0.5. Then we increase p with step of 0.02 until 0.7. We use 10 Metropolis-Hastings Chains, with 550 iterations including 50 of burn-in, from the algorithm presented in Subsection 6.4, for every p. The median and 95% credible interval of $p^{M\to F}$ is shown in Table 8. We see that with uniformative topological data ($p = 0.5$), our algorithm is not able to determine the direction of transmission (median of $p^{M\to F}$ is 0.51). This alternation between source and recipient is illustrated by a trace plot of both individuals' time of infection in Figure 16. The order of infection keeps changing, and the day of infection are intermingled. Then, the more phylogenetic relationship data support one direction, the greater the posterior proportion of the similar epidemiological direction. The latter increasing faster than the former. Finally, with 68% of ancestral topology in one direction, the algorithm stays stuck on the corresponding epidemiological direction.

The proportion $p^{M\to F}$ increases until settling to 1 at $p = 68\%$, when there is no move on the source vector anymore. It means that the likelihood favored the $M \to F$ direction repeatedly, by a great extend. In other words, the other direction was always rejected. Let us explore this further and relax one layer of heterogeneity, the location of observed phylogenetic topology. We create three regions on the genome. The first region with windows $W_1 = \{w :$

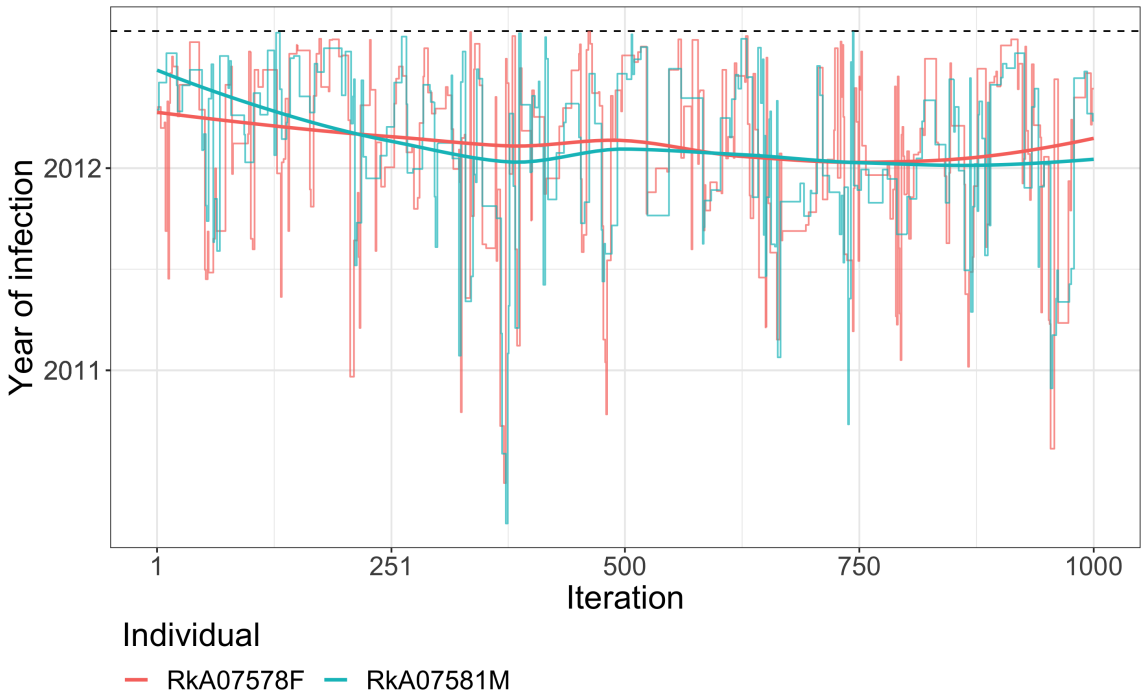


Figure 16: Trace plots of t_i^I for $i \in$ {RkA07528F, RkA07581M} in Network 1 with uniformative phylogenetic relationship data. We simulate phylogenetic relationship data for Network 1. Among the 60 observations, we attribute half of them to an ancestral topology with direction $M \to F$ and the other half to the opposite direction. All windows indexes are fixed to the first window. We use our Metropolis-Hastings algorithm presented in Subsection 6.4 with 1100 iterations, 100 of those considered as burn-in.

p	50%	52%	54%	56%	58%
$p^{M\to F}$	50.7% [45.22, 56.82]	75.0% [68.05, 78.93]	87.6% [81.97, 90.78]	95.5% [93.14, 96.71]	98.3% [96.34, 99.51]

p	60%	62%	64%	66%	68%	70%
$p^{M\to F}$	99.8% [98.89, 100]	99.9% [99.60, 100]	100 % [99.64, 100]	100 % [99.80, 100]	100 % [100, 100]	100 % [100, 100]

Table 8: Inferred epidemic direction given simulated phylogenetic data. We simulate phylogenetic relationship data for Network 1. Among the 60 observations, $p \times 100\%$ of them are simulated to have an ancestral topology with direction $M \to F$ and $(1-p) \times 100\%$ of them with the opposite direction. All windows indexes are fixed to the first window. We use our Metropolis-Hastings algorithm presented in Subsection 6.4 with 550 iterations, 50 of those considered as burn-in and observe the proportion of iteration for which the direction $M \to F$ was sampled, $p^{M\to F}$. We repeat this 10 times and obtain the median and 95% credible interval of $p^{M\to F}$

$\tilde{\Delta}_w^M < -0.76\}$, the second region includes $W_2 = \{w : \tilde{\Delta}_w^M > 0.41\}$ and the third region considers windows in between, $W_3 = \{w : -0.76 < \tilde{\Delta}_w^M < 0.41\}$. These limits were taken as the 25% and 75% quantiles of $\tilde{\Delta}_w^M$. The number of windows in each of these group is 84, 84, and 167, respectively. From the topology model in (5–14), the likelihood will be greater for the ancestral topology with similar direction as the epidemiological direction, observed in the first window group than in the second window group. We say that there is no notable difference on the likelihood in the third window group. We create three simulated data set of 35 phylogenetic topology. We attribute to each of them $(1-p) \times 35$ ancestral topologies with direction $F \to M$ and associate to them randomly chosen window indexes from Group 3 (neutral). We refer to these windows as $W^{F\to M}$. Then, for the three groups $k \in \{1, 2, 3\}$, we attribute $p \times 35$ ancestral topologies with a direction similar to the initialized epidemiologic one ($M \to F$). We assign to them randomly selected windows from the corresponding group k and refer to these windows as $W_k^{M\to F}$. Lastly, we fix time elapsed to 5 years. Note that when we propose epidemiologic direction $M \to F$, $R_{ijw} = R^{1\to 2}, \forall w \in W_k^{M\to F}$, and when the epidemiological direction is reconstructed as $F \to M$, $R_{ijw} = R^{2\to 1}, \forall w \in W_k^{M\to F}$. The difference of the log-likelihoods given opposite epidemiological direction for group k is,

$$\sum_{w\in W_k^{M\to F}} logp(R_{ijw} = R^{2\to 1}|\boldsymbol{\psi}^M) - \sum_{w\in W^{F\to M}} logp(R_{ijw} = R^{1\to 2}|\boldsymbol{\psi}^M)$$
$$- \sum_{w\in W_k^{M\to F}} logp(R_{ijw} = R^{1\to 2}|\boldsymbol{\psi}^M) + \sum_{w\in W^{F\to M}} logp(R_{ijw} = R^{2\to 1}|\boldsymbol{\psi}^M). \qquad (6\text{–}15)$$

This difference simulates the likelihood influence on the acceptance ratio of the Metropolis-Hastings algorithm in (6–14) between direction $F \to M$ and $M \to F$. If this difference is great, the move from direction $M \to F$ to $F \to M$ will be likely accepted. On the contrary, if the difference is small, it means that epidemiological direction $M \to F$ is more likely given the data at hand, and the likelihood will be in disfavor of the move. Figure 17 shows the difference (6–15) against a grid of p from 0.5 tp 0.7 with a step of 0.01, for the three groups. In the third group, that does not favor any of the direction based on the genomic region, the difference decreases with p. This corresponds to the settings of the previous simulation study with results in Table 8. At $p = 68\%$, the likelihood difference is -7.3 leading the move to always be rejected and the source vector to stay stuck in one direction. The first group possesses ancestral topologies with $M \to F$ direction in regions that favor the similar epidemiologic direction and therefore presents an even sharpen decrease with p. Interestingly, the second group, with observations on windows in regions that disfavor the corresponding epidemiological direction presents a positive, and steady, log-likelihood difference with p. It means that even though the proportion of topology with direction $M \to F$ increases relative to direction $F \to M$, the location of these observations take precedent over the likelihood and disfavor the corresponding epidemic direction.

6.5.3 Limitations

We observe in Figure 15 that some times of infection are supported on tighter space than others. This is because the likelihood of phylogenetic data of a pair depends on the time of infection of the source (i.e., the individual who transmitted to the other in the pair). The

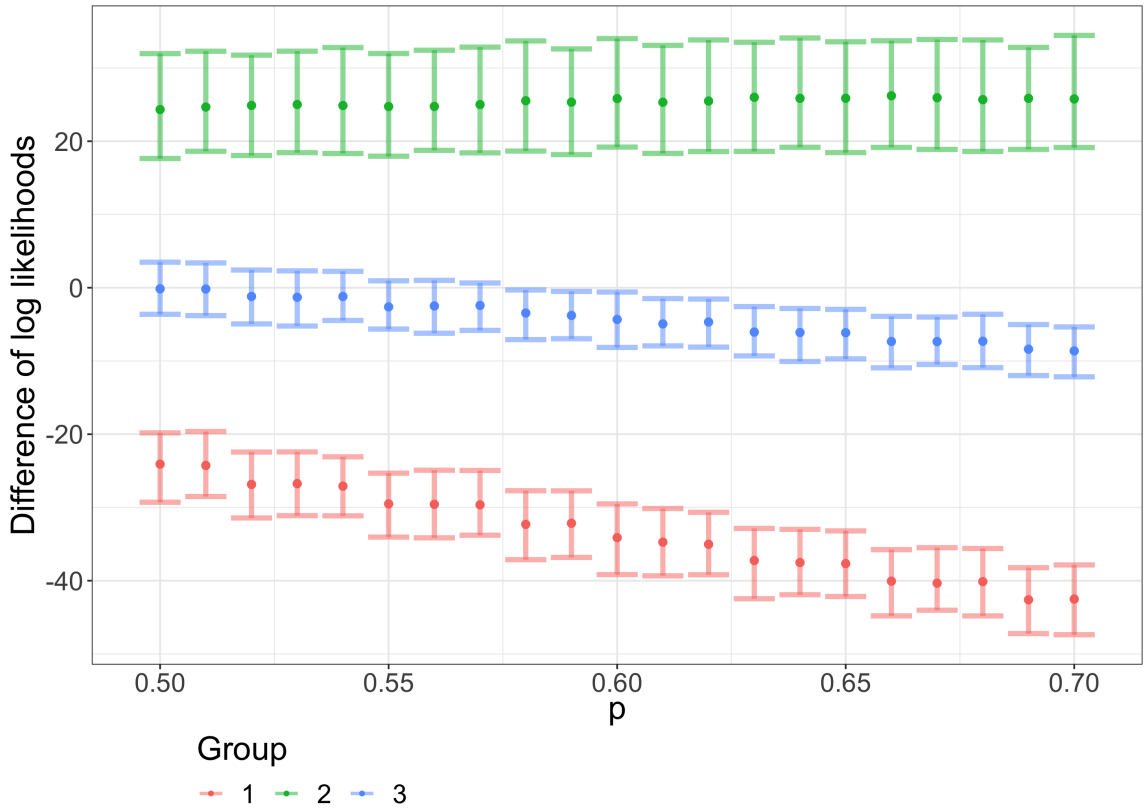


Figure 17: Difference of log likelihoods with opposite epidemiological direction. We simulate three data sets containing 35 phylogenetic relationships. Each of them have $p \times 35$ ancestral topology in the $M \to F$ direction, and $(1-p) \times 35$ in the other direction. Ancestral topologies with direction $F \to M$ from group 1 are associated with windows with $\tilde{\Delta}_w < -0.76$, from group 2 with $\tilde{\Delta}_w > 0.41\}$ and from group 3 with $-0.76 < \tilde{\Delta}_w < 0.41$. Ancestral topologies with $F \to M$ are associated with windows satisfying $-0.76 < \tilde{\Delta}_w < 0.41$ for all three data sets. The difference of log likelihoods defined in 6–15 is map against the proportion p of the topology with $M \to F$ direction.

only constraint on the recipient's time of infection is on the prior of the transmission chain through the generation time function. This is very well illustrated in networks involving only two individuals, in which we can identify the recipient by the wider posterior support of their infection time.

From the simulation analysis, we remark two things. Firstly, if the phylogenetic data are even slightly informative, the likelihood will always favor one epidemiological direction between two individuals. The preference for one direction depends on two criteria. The first one is the proportion of observations with this particular direction relative to the other direction. The second one is the location of these observations. Specifically, if the proportion is higher, the likelihood will be greater for a similar epidemiological direction, with window location held fixed. Moreover, for the same number of observations in each direction, the likelihood will be higher for the epidemiological direction correspondent to the phylogenetic

one observed on windows with a smaller difference in the log odds of all categories relative to the baseline ($\tilde{\Delta}_w^M$) than the other direction. Secondly, this preference, which can be quite strict (i.e., large log-likelihood difference), will stay constant at every iteration because we fixed the genetic parameter $\boldsymbol{\psi}$ to the empirical median. So, many of our pairs will likely stay stuck in one direction as we experienced when p passes the 0.68 thresholds in Table 8.

7 Epidemic results

There are two latent variables encapsulated in the transmission chain $\boldsymbol{T}$ sampled by the MCMC. We investigate the first one, the pairwise source-recipient linkage in Subsection 7.1. In Subsection 7.2, the times of infection, are observed. Lastly, we explore the characteristics of receivers and spreaders identified by the algorithm, among all networks.

7.1 Inferred transmission linkages

The algorithm samples entire transmission chains as a latent variable. We enumerate all possible chains for small networks ($N \leq 3$) and associate their posterior support in Table 9. To study larger transmission chain ($N > 3$), the likely sources, for each individual, are presented in Table 10. Let us focus on two interesting results. Network 2 ($N = 2$) have an initialized direction RkA07286F $\rightarrow$ RkA07207M. Approximately 61% of the ancestral phylogenetic topology are in this direction, as shown in Table 3, which satisfy the first criteria for the direction to have a greater likelihood. Moreover, the windows on which these directed topologies are observed have a smaller coefficient $\tilde{\Delta}$ with median and credible intervals being 0.2304 [0.1689, 0.2919], than in the other direction (0.3789 [0.2185, 0.4723]). Therefore, the likelihood will always favor this direction and, as expected, our algorithm stays stuck most of the time in that order. Consider a second example with Network 3 ($N = 2$). The initialized direction is RkA07342F $\rightarrow$ RkA07150M, coherent with 59% of ancestral topology. However, this direction is observed on windows with greater $\tilde{\Delta}_w^M$ (0.4998 [0.1708,1.3743]), than the other direction (0.1747 [-0.5242 ,0.2580]). Because the absolute proportion is not much greater, the epidemiological direction favored by the likelihood ratio is the opposite of the initialized one.

7.2 Inferred times of infection

Table 11 presents the median and 95% credible intervals of the time of infection for every individuals in each network.

7.3 Risk factors of spreaders and receivers

One of the motivations behind reconstructing transmission chain is to identify receiver and spreader groups within the study population and study their characteristics. Let us consider the most likely transmission chain in each network (i.e., chains with greater support) defined by the corresponding source vector $\bar{\boldsymbol{S}}$. We identify among all networks, spreaders, and receivers. We say that i is a spreader if he transmitted the disease to at least one individual, such that $\exists j$ such that $\bar{S}_j = i$. On the opposite, i is receiver if he did not transmit the disease $\nexists j$ such that $\bar{S}_j = i$. Table 12 presents two characteristics, age at infection and gender for individuals in these groups. We find that individuals in the spreader group were infected at a younger age than in the ones in the receiver group. Intuitively it makes sense because the younger an individual is infected, the more opportunities he has to infect others. We need to be careful about this result. Indeed, as we explained in the limitations, the receiver time of infection is not weighted by the likelihood and is much more variable. Another interesting finding is that spreaders are composed of more men, while a receiver is more likely to be a female.

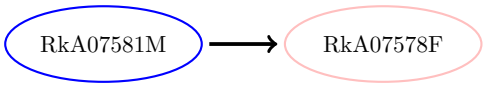
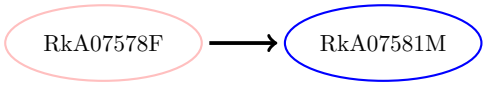
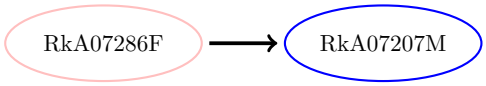
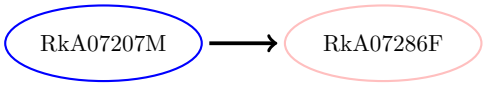
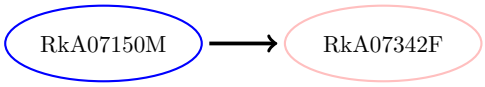
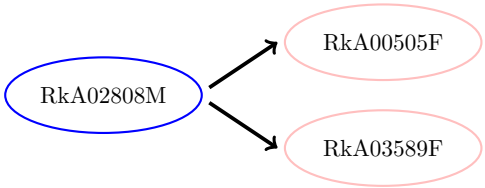
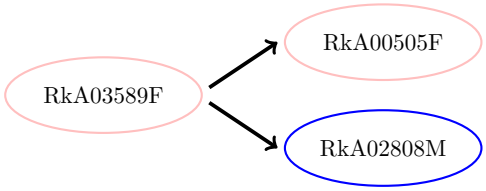
Transmission chain	Posterior support
Network 1, $N = 2$	
	0.9892
	0.0108
Network 2, $N = 2$	
	0.9997
	0.0003
Network 3, $N = 2$	
	1
Network 4, $N = 3$	
	0.9995
	0.0005

Table 9: Transmission chain for small networks ($N \leq 3$). 3 Metropolis-Hastings Chains from our MCMC algorithm presented in Subsection 6.4 sample in the posterior distribution of the transmission chain for each network. The MCMC algorithm samples entire transmission chain as a latent variable. We enumerate sampled chains and associate the posterior support. Female are identified in pink and Male in blue.

Recipient	Source	Posterior support
RkA06111M	External (0)	0.6059
	RkA05669F	0.3941
RkA05669F	RkA06111M	0.6059
	External (0)	0.3941
RkA03794M	RkA06111M	0.6059
	RkA05669F	0.3941
RkA04533F	RkA03794M	1

(a) Network 5, $N = 4$.

Recipient	Source	Posterior support
RkA05079F	External (0)	0.9999
	RkA04621M	0.0001
RkA04621M	RkA05079F	0.9999
	External (0)	0.0001
RkA01459F	RkA05079F	0.9951
	RkA05257M	0.0049
RkA01733M	RkA01459F	0.9951
	RkA05257M	0.0049
RkA05257M	RkA01459F	0.9951†
	RkA05079F	0.0049†
	RkA04621M	0.0001†

(b) Network 6, $N = 5$.

Recipient	Source	Posterior support
RkA00569M	External (0)	1
RkA01258F	RkA00569M	1
RkA04356F	RkA00569M	1
RkA01351M	RkA00569M	0.9986
	RkA02116F	0.0014
RkA02116F	RkA01351M	0.9986
	RkA00569M	0.0014

(c) Network 7, $N = 5$.

†: These sum to 1 without rounding.

Table 10: Pairwise source-recipient linkage for large networks ($N > 3$). 3 Metropolis-Hastings Chains from our MCMC algorithm presented in Subsection 6.4 sample in the posterior distribution of the transmission chain for each network. We present the values of the sampled source vector $\boldsymbol{S}$ by the MCMC algorithm and associate their posterior support.

Parameter	Median	Credible interval
Network 1, $N = 2$		
$t^I_{\text{RkA07578F}}$	2010.54	[2008.77, 2012.55]
$t^I_{\text{RkA07581M}}$	2009.54	[2008.25, 2010.89]
Network 2, $N = 2$		
$t^I_{\text{RkA07207M}}$	2009.73	[2007.70, 2013.05]
$t^I_{\text{RkA07286F}}$	2008.06	[2007.53, 2009.24]
Network 5, $N = 2$		
$t^I_{\text{RkA07150M}}$	2002.43	[2001.74, 2003.10]
$t^I_{\text{RkA07342F}}$	2006.26	[2002.23, 2009.74]
Network 4, $N = 3$		
$t^I_{\text{RkA00505F}}$	2008.40	[2004.69, 2011.66]
$t^I_{\text{RkA02808M}}$	2004.87	[2003.94, 2007.26]
$t^I_{\text{RkA03589F}}$	2008.35	[2004.61, 2011.68]

(a) Small networks ($N \leq 3$).

Parameter	Median	Credible interval
Network 5, $N = 4$		
$t^I_{\text{RkA03794M}}$	2006.95	[2005.53, 2009.95]
$t^I_{\text{RkA04533F}}$	2010.06	[2006.45, 2013.08]
$t^I_{\text{RkA05669F}}$	2006.95	[2003.86, 2011.41]
$t^I_{\text{RkA06111M}}$	2006.08	[2005.35, 2011.72]
Network 6, $N = 5$		
$t^I_{\text{RkA01459F}}$	2006.45	[2005.18, 2008.35]
$t^I_{\text{RkA01733M}}$	2009.41	[2006.08, 2012.36]
$t^I_{\text{RkA04621M}}$	2008.74	[2005.32, 2012.20]
$t^I_{\text{RkA05079F}}$	2005.33	[2004.54, 2006.94]
$t^I_{\text{RkA05257M}}$	2009.29	[2006.06, 2012.17]
Network 7, $N = 5$		
$t^I_{\text{RkA00569M}}$	2004.51	[2004.08, 2005.81]
$t^I_{\text{RkA01258F}}$	2008.18	[2004.62, 2011.56]
$t^I_{\text{RkA01351M}}$	2007.87	[2005.02, 2011.03]
$t^I_{\text{RkA02116F}}$	2010.08	[2006.49, 2012.14]
$t^I_{\text{RkA04356F}}$	2008.24	[2004.59, 2011.60]

(b) Large networks ($N > 3$).

Table 11: Median and 95% credible interval of the time of infection. 3 Metropolis-Hastings Chains from our MCMC algorithm presented in Subsection 6.4 sample in the posterior distribution of the transmission chain for each network. Taking into account all these observations, a marginal empirical posterior distribution can be formed on the time elapsed. We report the empirical median and 95% confidence interval.

Group	Total	Gender		Age at infection
		Female	Male	
Spreaders	10	30%	70%	22.9915 [17.5407, 38.4078]
Receivers	13	69.23%	30.77%	29.6160 [16.1322, 37.6356]

Table 12: Risk factors of identified spreaders and receivers. From the reconstructed transmission chains, we identify spreader individuals (i.e., who transmitted to at least another individual) and receiver (i.e., who did no transmit). Within these two groups, we present two epidemiological characteristics.

8 Discussion

The motivation of this thesis is to identify, in a host population infected with HIV-1, pairwise linkage and direction of transmission as well as the time of transmission. To this aim, we use deep-sequence data that permit a unique insight into the within-host diversity as well as offer better perspective for transmission tree reconstruction. We develop a likelihood for these data, a prior for the transmission chain and we create an MCMC to make inference on 7 networks.

Two data type are considered, the mean tip-to-tip distance and the pairwise phylogenetic relationships. The former identifies related pairs, and the latter determines the direction of transmission. We develop novel likelihoods for both data sources. After finding a parametric form, we use an informative empirical Bayes analysis to estimate the parameters values. We include in this step only pairs that were found to be likely linked to infer the parameters values under this relationship. For the distance likelihood, we introduce an evolutionary time referred to as the time elapsed, that depend on the source time of infection. We find that the distance between pathogen's reads increases with time elapsed and that specific genomics region are more favorable to mutation than others. For the phylogenetic relationship likelihood, we investigate which topologies (i.e., phylogenetic branches pattern) are likely to form depending on the time elapsed and the genetic window. We find that the ancestral topology in the incorrect direction and the intermingled pattern are not significantly less likely to appear than the ancestral topology in the correct epidemiological direction. This result is likely due to incorrect phylogenetic reconstruction on windows with few reads. We also find that the incorrect direction is significantly less likely to appear relative to the correct one, with increasing time elapsed. Then, we derive the prior of the transmission chain that included times of infection and the source vector. We use transmission hazard that quantifies the force of infection on a particular individual at a specific time. This prior depends on the generation time function, which determines the infectiousness of the case throughout his infection, and on one epidemiological parameter, the baseline of the transmission hazard.

We create a Metropolis-Hastings algorithm, presenting the initial values, updates, and acceptance ratio for the transmission chain and epidemiological latent variables. To assert if our model is working, we investigate 7 networks involving individuals that were found to be related by previous work. In each of these networks, we use our MCMC to sample from the posterior of the times of infection, the source vector, and the epidemiological parameter. We have good convergence and mixing. However, two limitations were found. First, the transmission linkage between pairs is likely to stay stuck. Second, the receiver time of infection (i.e., individuals that did not infect anyone) is not included in the likelihood and has wider posterior support than spreader (i.e., individuals that did infect at least one individual).

Finally, we investigate some epidemiological results from the inferred transmission chains. We find that the spreaders group were infected at a younger age than individuals in the receivers group. Moreover, we identify that male were more likely to spread the infection than female.

Several assumptions made in this thesis are not realistic. Firstly, we assume that all individuals in the host population were sampled and connected. Secondly, we assume independence of phylogenetic data from one window to another. This not true because the windows are overlapping. We also assume independence between the two data types, the dis-

tance and the topological relationship, which is also not true. Indeed, it is straightforward to find examples, such as; with greater distance, pathogens are more likely to be disconnected. Lastly, to compute the time elapsed, we consider that the coalescent time is equal to the transmission time.

Further work should create a systematic framework to cluster related individuals in a population before applying our MCMC. In this thesis, we use identified networks from previous work, but we could not have applied our method on a large population. Then, the MCMC could be developed to sample from the genetic parameter posterior rather than fixing it to the empirical median. That would leave more freedom to the likelihood to support, or not, an epidemiological direction and might avoid obtaining a sticky source vector. Finally, the prior, that determines the probability of transmission between two individuals, could include risk factors characterizing the individuals in question, such as age, geographical location, gender, etc.

References

Abbey, H. (1952), 'An examination of the reed-frost theory of epidemics', *Human Biology.* **24**(201).

Allaby, M. (2012), *A Dictionary of Plant Sciences*, Vol. 9780199600571, 3rd edn, Oxford University Press.

Allen, L. (2008), *An introduction to stochastic epidemic models*, Vol. 1945 of *p. 81–130.*, berlin: springer. edn, Brauer F, van den Driessche P, Wu J, editors.

Allen, L. J. S. (2017), 'A primer on stochastic epidemic models: Formulation, numerical simulation, and analysis', *Infectious Disease Modelling* **2**(2), 128–142.

Anderson, R. M. and Medley, G. F. (1988), 'Epidemiology of hiv infection and aids: incubation and infectious periods, survival and vertical transmission', *AIDS* **2**.

Becker, N. G. and Britton, T. (1999), 'Statistical studies of infectious disease incidence', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **61**(2), 287–307.

Bellan, S. E., Dushoff, J., Galvani, A. P. and Meyers, L. A. (2015), 'Reassessment of hiv-1 acute phase infectivity: accounting for heterogeneity and study design with simulated cohorts', *PLoS medicine* **12**(3), e1001801–e1001801.

Buvé, A., Bishikwabo-Nsarhaza, K. and Mutangadura, G. (2002), 'The spread and effect of hiv-1 infection in sub-saharan africa', *The Lancet* **359**(9322), 2011–2017.

Cauchemez, S., Bhattarai, A., Marchbanks, T. L., Fagan, R. P., Ostroff, S., Ferguson, N. M., Swerdlow, D. and (2011), 'Role of social networks in shaping disease transmission during a community outbreak of 2009 h1n1 pandemic influenza', *Proceedings of the National Academy of Sciences* **108**(7), 2825.

Cauchemez, S., Carrat, F., Viboud, C., Valleron, A.-J. and Boelle, P.-Y. (2004), 'A bayesian mcmc approach to study transmission of influenza: application to household longitudinal data', *Statistics in medicine* **23**, 3469–87.

Cauchemez, S. and Ferguson, N. M. (2012), 'Methods to infer transmission risk factors in complex outbreak data', *Journal of the Royal Society, Interface* **9**(68), 456–469.

Cauchemez, S., Temime, L., Valleron, A.-J., Varon, E., Thomas, G., Guillemot, D. and Boëlle, P.-Y. (2006), 'S. pneumoniaetransmission according to inclusion in conjugate vaccines: Bayesian analysis of a longitudinal follow-up in schools', *BMC Infectious Diseases* **6**(1), 14.

Chang, L. W., Grabowski, M. K., Ssekubugu, R., Nalugoda, F., Kigozi, G., Nantume, B., Lessler, J., Moore, S. M., Quinn, T. C., Reynolds, S. J., Gray, R. H., Serwadda, D. and Wawer, M. J. (2016), 'Heterogeneity of the hiv epidemic in agrarian, trading, and fishing communities in rakai, uganda: an observational epidemiological study', *The Lancet HIV* **3**(8), e388–e396.

Cori, A., Boëlle, P.-Y., Thomas, G., Leung, G. M. and Valleron, A.-J. (2009), 'Temporal variability and social heterogeneity in disease transmission: The case of sars in hong kong', *PLOS Computational Biology* **5**(8), e1000471–.

Didelot, X., Gardy, J. and Colijn, C. (2014), 'Bayesian inference of infectious disease transmission from whole-genome sequence data', *Molecular Biology and Evolution* **31**(7), 1869–1879.

Gelman, A. (2006), 'Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper)', *Bayesian Anal.* **1**(3), 515–534.

Gelman, A. and Rubin, D. B. (1992), 'Inference from iterative simulation using multiple sequences', *Statist. Sci.* **7**(4), 457–472.

Ghitza, Y., G. A. (2014), 'The great society, reagan's revolution, and generations of presidential voting', *To be submitted.* .

Greenwood, M. (1931), 'On the statistical measure of infectiousness', *The Journal of hygiene* **31**(3), 336–351.

Hanson, D. L., Song, R., Masciotra, S., Hernandez, A., Dobbs, T. L., Parekh, B. S., Owen, S. M. and Green, T. A. (2016), 'Mean recency period for estimation of hiv-1 incidence with the bed-capture eia and bio-rad avidity in persons diagnosed in the united states with subtype b', *PLOS ONE* **11**(4), e0152327–.

Hastings, W. K. (1970), 'Monte carlo sampling methods using markov chains and their applications', **57**(1), 97–109.

Havard, R. and Leonhard, H. (2005), *Gaussian Markov Random Fields: Theory and Applications*, number 9781584884323 *in* 'Chapman and Hall/CRC Monographs on Statistics and Applied Probability', 1st edition edn, Chapman and Hall/CRC.

Hollingsworth, T. D., Anderson, R. M. and Fraser, C. (2008), 'Hiv-1 transmission, by stage of infection', *The Journal of Infectious Diseases* **198**(5), 687–693.

Holmes, E. C., Nee, S., Rambaut, A., Garnett, G. P., Harvey, P. H., Leigh, B. A. J. and Smith, J. M. (1995), 'Revealing the history of infectious disease epidemics through phylogenetic trees', *Philos Trans R Soc Lond B Biol Sci.* **349**(1327), 33–40.

Jombart, T., Cori, A., Didelot, X., Cauchemez, S., Fraser, C. and Ferguson, N. (2014), 'Bayesian reconstruction of disease outbreaks by combining epidemiologic and genomic data', *PLOS Computational Biology* **10**(1), e1003457–.

Juárez, M. A. and Steel, M. F. J. (2010), 'Model-based clustering of non-gaussian panel data based on skew-t distributions', *Journal of Business & Economic Statistics* **28**(1), 52–66.

Jukes, T. and Cantor, C. (1969), 'Evolution of protein molecules', *New York: Academic Press* pp. pp. 21–132.

Lam, T. T.-Y., Hon, C.-C. and Tang, J. W. (2010), 'Use of phylogenetics in the molecular epidemiology and evolutionary studies of viral infections', *Critical Reviews in Clinical Laboratory Sciences* **47**(1), 5–49.

Leitner, T. and Albert, J. (1999), 'The molecular clock of hiv-1 unveiled through analysis of a known transmission history', *Proceedings of the National Academy of Sciences* **96**(19), 10752.

Leitner, T. and Romero-Severson, E. (2018), 'Phylogenetic patterns recover known hiv epidemiological relationships and reveal common transmission of multiple variants', *Nature Microbiology* **3**(9), 983–988.

Lemey, P., Kosakovsky Pond, S. L., Drummond, A. J., Pybus, O. G., Shapiro, B., Barroso, H., Taveira, N. and Rambaut, A. (2007), 'Synonymous substitution rates predict hiv disease progression as a result of underlying replication dynamics', *PLOS Computational Biology* **3**(2), e29–.

Li, G., Piampongsant, S., Faria, N. R., Voet, A., Pineda-Peña, A.-C., Khouri, R., Lemey, P., Vandamme, A.-M. and Theys, K. (2015), 'An integrated map of hiv genome-wide variation from a population perspective', *Retrovirology* **12**, 18–18.

Lui, K., Darrow, W. and Rutherford, G. (1988), 'A model-based estimate of the mean incubation period for aids in homosexual men', *Science* **240**(4857), 1333.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. (1953), 'Equation of state calculations by fast computing machines', *The Journal of Chemical Physics* **21**(6), 1087–1092.

Morelli, M. J., Thébaud, G., Chadœuf, J., King, D. P., Haydon, D. T. and Soubeyrand, S. (2012), 'A bayesian inference framework to reconstruct transmission trees using epidemiological and genetic data', *PLOS Computational Biology* **8**(11), e1002768–.

Nahmias, A. J., Weiss, J., Yao, X., Lee, F., Kodsi, R., Schanfield, M., Matthews, T., Bolognesi, D., Durack, D., Motulsky, A., Kanki, P. and Essex, M. (1986), 'Evidence for human infection with an htlv iii/lav-like virus in central africa, 1959', *The Lancet* **327**(8492), 1279–1280.

O'Neill, P. D., Balding, D. J., Becker, N. G., Eerola, M. and Mollison, D. (2000), 'Analyses of infectious disease data from household outbreaks by markov chain monte carlo methods', **49**(4), 517–542.

O'Neill, P. D. and Roberts, G. O. (1999), 'Bayesian inference for partially observed stochastic epidemics', **162**(1), 121–129.

Piot, P., Russell, S. and Larson, H. (2007), 'Good politics, bad politics: The experience of aids', *Am J Public Health* **97(11)**, 1934–1936.

Plummer, M., Best, N., Cowles, K. and Vines, K. (2005), 'Coda: Convergence diagnosis and output analysis for mcmc', *R News* **6**.

Pohar Perme, M., Blas, M. and Turk, S. (2004), 'Comparison of logistic regression and linear discriminant analysis : a simulation study', **1**(1), 143–161.

Polson, N. G. and Scott, J. G. (2012), 'On the half-cauchy prior for a global scale parameter', *Bayesian Anal.* **7**(4), 887–902.

Pybus, O. G. and Rambaut, A. (2009), 'Evolutionary analysis of the dynamics of viral infectious disease', *Nature Reviews Genetics* **10**, 540 EP –.

Quinn, T., Mann, J., Curran, J. and Piot, P. (1986), 'Aids in africa: an epidemiologic paradigm', *Science* **234**(4779), 955.

Ratmann, O., Grabowski, M. K., Hall, M., Golubchik, T., Wymant, C., Abeler-Dörner, L., Bonsall, D., Hoppe, A., Brown, A. L., de Oliveira, T., Gall, A., Kellam, P., Pillay, D., Kagaayi, J., Kigozi, G., Quinn, T. C., Wawer, M. J., Laeyendecker, O., Serwadda, D., Gray, R. H., Fraser, C., Consortium, P. and Program, R. H. S. (2019), 'Inferring hiv-1 transmission networks and sources of epidemic spread in africa with deep-sequence phylogenetic analysis', *Nature Communications* **10**(1), 1411.

Ratmann, O., van Sighem, A., Bezemer, D., Gavryushkina, A., Jurriaans, S., Wensing, A., de Wolf, F., Reiss, P., Fraser, C. and (2016), 'Sources of hiv infection among men having sex with men and implications for prevention', *Science Translational Medicine* **8**(320), 320ra2.

Rom, W. and Markowitz, S. (2007), *Environmental and occupational medicine*, number 978-0-7817-6299-1, 4th edn, Philadelphia: Wolters Kluwer/Lippincott Williams and Wilkins.

Romero-Severson, E. O., Bulla, I. and Leitner, T. (2016), 'Phylogenetically resolving epidemiologic linkage', *Proceedings of the National Academy of Sciences* **113**(10), 2690.

Skar, H., Albert, J. and Leitner, T. (2013), 'Towards estimation of hiv-1 date of infection: A time-continuous igg-model shows that seroconversion does not occur at the midpoint between negative and positive tests', *PLOS ONE* **8**(4), e60906–.

Starkweather, J. and Moske, K. (2011), 'Multinomial logistic regression', *Unpublished manuscript* .

UNAIDS (2010), 'Combination hiv prevention: Tailoring and coordinating biomedical, behavioural and structural strategies to reduce new hiv infections', *Last checked: 10/08/2019* .
URL: *http://files.unaids.org/en/media/unaids/contentassets/documents/unaidspublication/2011/20111110_JC2007_Combination_Prevention_paper_en.pdf*

UNAIDS and WHO (2007), '2007 aids epidemic update', *Last checked: 23/08/2019* .
URL: *http://data.unaids.org/pub/epislides/2007/2007_epiupdate_en.pdf*

Vihola, M. (2012), 'Robust adaptive metropolis algorithm with coerced acceptance rate', *Statistics and Computing* **22**(5), 997–1008.

Vrancken, B., Rambaut, A., Suchard, M. A., Drummond, A., Baele, G., Derdelinckx, I., Van Wijngaerden, E., Vandamme, A.-M., Van Laethem, K. and Lemey, P. (2014), 'The genealogical population dynamics of hiv-1 in a large transmission chain: Bridging within and among host evolutionary rates', *PLOS Computational Biology* **10**(4), e1003505–.

Wilson, D. and Halperin, D. T. (2008), '"know your epidemic, know your response": a useful approach, if we get it right', *The Lancet* **372**(9637), 423–426.

WorldBank (2017), 'World development indicators - prevalence of hiv on popoulation ages 15-49 in 2017', *Last checked: 10/08/2019* .
URL: *https://databank.worldbank.org/reports.aspx?source=2&series=SH.DYN.AIDS.ZS&country=ZAF#*

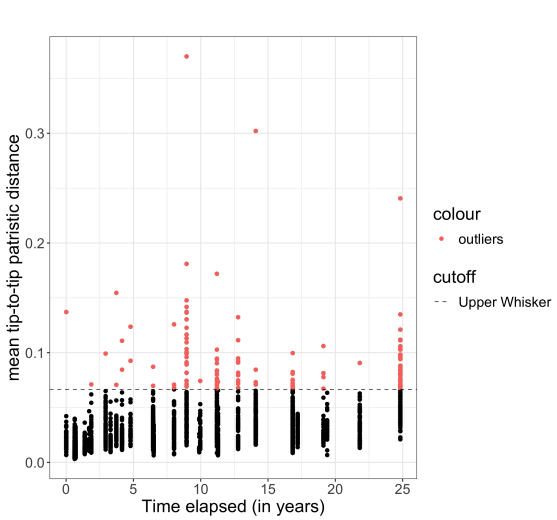
Wymant, C., Hall, M., Ratmann, O., Bonsall, D., Golubchik, T., de Cesare, M., Gall, A., Cornelissen, M., Fraser, C., STOP-HCV Consortium, T. M. P. C. and Collaboration, T. B. (2017), 'Phyloscanner: Inferring transmission from within- and between-host pathogen genetic diversity', *Molecular Biology and Evolution* **35**(3), 719–733.

Ypma, R. J. F., Bataille, A. M. A., Stegeman, A., Koch, G., Wallinga, J. and van Ballegooijen, W. M. (2012), 'Unravelling transmission trees of infectious diseases by combining genetic and epidemiological data', *Proceedings. Biological sciences* **279**(1728), 444–450.

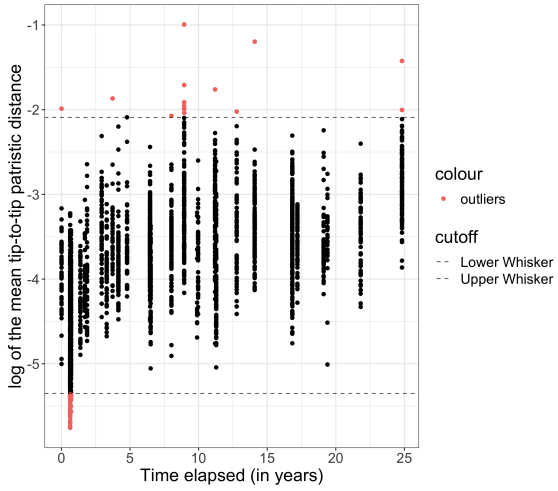
Appendices

Appendix A: Data transformation

The use of the logarithmic scale on the response is motivated heteroscedasticity in the raw scale and a reduction of the outliers' influence by applying the logarithmic scale. Figure A1 shows the distribution of the mean tip-to-tip patristic distance, in raw and logarithmic scale, against time elapsed. In Figure A1a, we observe an increase in the response's variance when the elapsed time increases. This response behavior is not observed with the log scale. Moreover, using the log scale reduces the number of outliers from 127 to 56.



(a) Mean patristic distance against time elapsed



(b) Log mean patristic distance against time elapsed

Figure A1: Mean tip-to-tip patristic distance in raw and log scale against time elapsed. We consider the 27 related couples from Subsection 5.2. We compare the mean tip-to-tip patristic distance distribution in the original (A1a) and logarithmic(A1b) scale to the time elapsed. Red dots highlight outliers. Dashed lines indicates the lower whisker (Q1 - 1.5× IQR) and the upper whisker (Q3 + 1.5× IQR). After applying the logarithmic scale, we observe a reduction in the number of outliers and an homoskedastic response.

Appendix B: Additional models for the Mean tip-to-tip patristic distance likelihood

We want to test the robustness of the time elapsed effect on mean tip-to-tip patristic distance found in the main text. This parameter is particular important in biology as it quantifies the development of patristic distance per unit of time. It describes the change in the distance for an additional year elapsed. Considering two additional models, one with only the fixed effect and one that adds independent random effects at the window level, we observe the median interval and 95% credibility interval of the fixed effect.

The first additional model model is:

$$logD_{ijw}|\mu_{ijw}, \sigma \sim t_\nu(\mu_{ijw}, \sigma), \tag{B.1a}$$
$$\mu_{ijw} = \alpha + \beta \times t^E_{ij}, \tag{B.1b}$$
$$\alpha \sim \mathcal{N}(0, 100), \tag{B.1c}$$
$$\beta \sim \mathcal{N}(0, 10), \tag{B.1d}$$
$$\sigma \sim C^+(0, 1), \tag{B.1e}$$
$$\nu \sim Gamma(2, 0.1). \tag{B.1f}$$

The second additional model is:

$$logD_{ijw}|\mu_{ijw}, \sigma \sim t_\nu(\mu_{ijw}, \sigma), \tag{B.2a}$$
$$\mu_{ijw} = \alpha + \alpha_w + \beta \times t^E_{ij}, \tag{B.2b}$$
$$\alpha \sim \mathcal{N}(0, 100), \tag{B.2c}$$
$$\alpha_w \sim \mathcal{N}(0, 10), \tag{B.2d}$$
$$\beta \sim \mathcal{N}(0, 10), \tag{B.2e}$$
$$\sigma \sim C^+(0, 1), \tag{B.2f}$$
$$\nu \sim Gamma(2, 0.1). \tag{B.2g}$$

The two additional models were fitted with Stan version 2.18.1, using 3 chains of 15,000 iterations each, the first 2,000 iterations being considered as burn-in. Chains converged and mixed correctly. The minimum and maximum effective sample sizes were respectively 9,586 and 17,948 for the first additional model; 2,108 and 94,201 for the second additional model. Table A1 indicates the empirical median and 95% credible intervals of the fixed effect on the models B.1 and B.2. We observe that the fixed effect identified in the main text is robust. Indeed, its credibility interval lies in the credibility intervals of the two additional models.

	Model B.1		Model B.2	
Parameter	Median	Credible interval	Median	Credible interval
β	0.0511	[0.0480, 0.0542]	0.0466	[0.0438, 0.0495]

Table A1: Median and credible interval of the fixed effect of model B.1 and model B.2. We fit models B.1 and B.2 with Stan version 2.18.1, using 3 chains with 15,000 iterations each, of which the first 2,000 iterations were considered as burn-in. Each iteration (excluding burn-in) is considered as observation of the posterior distribution for the parameter. Taking into account all these observations, a marginal empirical posterior distribution can be formed. We report the empirical median and 95% confidence interval of the time elapsed parameter.

Appendix C: Additional models for the Phylogenetic relationship likelihood

As above, we want to fit additional models on the phylogenetic relationships to investigate the behavior of the fixed effect. One of these models with only the fixed effect and another that adds independent random effects at the window level.

The first additional model model is:

$$R_{ijw}|\boldsymbol{\pi}_{ijw} \sim Categorical(\boldsymbol{\pi}_{ijw}), \tag{C.1a}$$

$$\boldsymbol{\pi}_{ijw} = (\pi_{ijw}^{R^{1\to2}}, ..., \pi_{ijw}^{R^{1\not\to2}}), \quad \sum_l \pi_{ijw}^l = 1, \tag{C.1b}$$

$$log\frac{\pi_{ijw}^r}{\pi_{ijw}^{R^{1\to2}}} = \tilde{\alpha}^r + \tilde{\beta}^r \times t_{ij}^E, \tag{C.1c}$$

$$\tilde{\alpha}^r \sim \mathcal{N}(0, 100), \tag{C.1d}$$

$$\tilde{\beta}^r \sim \mathcal{N}(0, 10),\ r \in \{R^{2\to1}, R^{1\leftrightarrow2}, R^{1\sqcup2}, R^{1\not\to2}\}. \tag{C.1e}$$

The second additional model is:

$$R_{ijw}|\boldsymbol{\pi}_{ijw} \sim Categorical(\boldsymbol{\pi}_{ijw}), \tag{C.2a}$$

$$\boldsymbol{\pi}_{ijw} = (\pi_{ijw}^{R^{1\to2}}, ..., \pi_{ijw}^{R^{1\not\to2}}), \quad \sum_l \pi_{ijw}^l = 1, \tag{C.2b}$$

$$log\frac{\pi_{ijw}^r}{\pi_{ijw}^{R^{1\to2}}} = \tilde{\alpha}^r + \tilde{\alpha}_w + \tilde{\beta}^r \times t_{ij}^E, \tag{C.2c}$$

$$\tilde{\alpha}^r \sim \mathcal{N}(0, 100), \tag{C.2d}$$

$$\tilde{\alpha}_w \sim \mathcal{N}(0, 10), \tag{C.2e}$$

$$\tilde{\beta}^r \sim \mathcal{N}(0, 10),\ r \in \{R^{2\to1}, R^{1\leftrightarrow2}, R^{1\sqcup2}, R^{1\not\to2}\}. \tag{C.2f}$$

The two additional models were fitted with Stan version 2.18.1, using 3 chains of 30,000 iterations each, the first 2,000 iterations being considered as a burn-in. Chains converged and mixed correctly. The minimum and maximum effective sample sizes were respectively 44,929 and 48,399 for the first additional model; 4,875 and 29,303 for the second additional model. Table A2 indicates the empirical median and 95% credible intervals of the fixed effect on models C.1 and C.2. The fixed effects identified in the main text are robust for every category. Indeed, their credibility intervals lie in the credibility intervals of the two additional models.

	Model C.1		Model C.2	
Parameter	Median	Credible interval	Median	Credible interval
$\tilde{\beta}^{R^{2\to 1}}$	-0.0371	[-0.0731, -0.0019]	-0.0519	[-0.0937, -0.0107]
$\tilde{\beta}^{R^{1\leftrightarrow 2}}$	-0.0241	[-0.0643, 0.0148]	-0.0389	[-0.0846, 0.0061]
$\tilde{\beta}^{R^{1\sqcup 2}}$	0.0067	[-0.0467, 0.0585]	-0.0070	[-0.0647, 0.0499]
$\tilde{\beta}^{R^{1\not\to 2}}$	0.1194	[0.0016, 0.2444]	0.1084	[-0.0148, 0.2379]

Table A2: Median and credible interval of the fixed effect of model C.1 and model C.2. We fit models C.1 and C.2 with Stan version 2.18.1, using 3 chains with 30,000 iterations each, of which the first 2,000 iterations were considered as burn-in. Each iteration (excluding burn-in) is considered as observation of the posterior distribution for the parameter. Taking into account all these observations, a marginal empirical posterior distribution can be formed. We report the empirical median and 95% confidence interval of the time elapsed parameters for each category.