The influence of viral within-host evolution on transmission chain reconstruction: a simulation study.

The influence of viral within-host evolution on transmission chain reconstruction: a simulation study.

Author:
Francesco Bosia
Supervisors:
Dr. Carsten Magnus, Prof. Dr. Tanja Stadler

August 2017

The Human Immunodeficiency Virus (HIV) is a measurably evolving pathogen responsible for the Acquired Immune Deficiency Syndrome (AIDS). Since the within-host evolution timescale is faster than the between-host transmission one, phylogenetical information can be used to infer epidemiological parameters. One possible piece of information that was proposed to be embedded in the sequences sampled from the hosts of an epidemics is the transmission tree. The phylogeny reconstructed from the sequences is used as an approximation of the transmission tree. This assumes that the within-host evolution has a negligible effect on the transmission tree reconstruction. In order to explore this assumption, this work presents the implementation of a mechanistic model of within-host viral evolutionary dynamics. A preliminary analysis identified the tree triplets distance as a suitable metric for reconstructed transmission tree comparison and reconstruction performance inference. Subsequent analysis explore the role of sampling time and mutation rate on transmission tree reconstruction performance. Different sampling strategies along the epidemics were proposed and confronted. The results suggested heterochronous sampling have a better reconstruction performance in comparison to homochronous sampling at the end of the simulation. Nevertheless, the heterochronous sampling strategies seem to present analogous performance. The role of the mutation rate was explored by simulating epidemics with a higher and a lower mutation rate, respectively. A higher mutation rate seems to provide more evolutionary signal for the reconstruction, improving its performance. This work represents a step towards the comprehension of the role of various factors on transmission tree reconstruction, and provides a software for hypothesis testing in a within-host evolution framework.

Acknowledgements

I would like to thank my supervisor Carsten Magnus for the exciting topic, for the always useful feedback and for being a straight to the point person. Not lastly, thanks for the very mathematical, as well as moral, support. Thanks for having believed in me, and trusted me with this ambitious and stimulating project. With Carsten I developed my passion for mathematical modelling and for their implementation. Many thanks to the whole Computational Evolution group for creating an incredibly friendly and pleasant working environment. Particularly, I would like to mention the very useful helping words spent by Hesam Montazeri, Venelin Mitov and Matteo Togninalli on the specificities of HIV modelling, computational implementation and general scientific chats. Thanks to Prof. Tanja Stadler for always offering an open door for questions. Further, I am grateful to Prof. Martin Pilhofer for hosting me as a laboratory assistant in his group. Lastly, my thoughts go to all of the inhabitants of my flat for the support and for making home *home*.

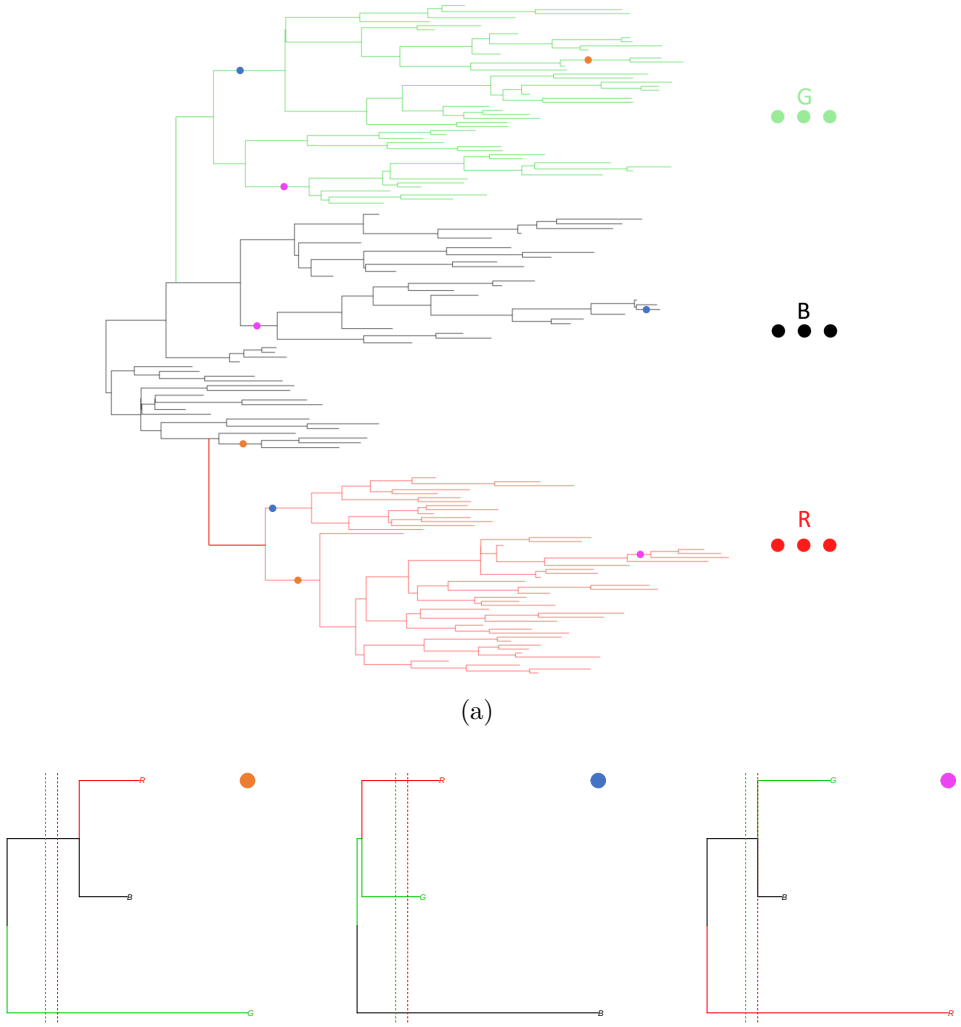
1 Introduction

After more than 30 years from its first isolation in 1983 [1], the Human Immunodeficiency Virus (HIV) still causes estimated 2.1 million new infections yearly worldwide and 1.1 million Acquired Immune Deficiency Syndrome (AIDS) related deaths [2]. The development of an effective treatment was an important milestone in the fight against this disease for two reasons: firstly, the decreased viral load in treated patient increases their life expectancy, restoring it to almost uninfected levels [3]. Secondly, treatment decreases HIV presence in body fluids, making patients under correct treatment regimen consistently less infectious [4]. Nevertheless, HIV still eludes an effective cure or a vaccine, partly due to the fast mutation rate characterizing retroviruses, partly due to its presence in latently infected cells [5]. Treatment is not available everywhere, for both economic and logistic reasons, and not everyone infected by HIV is aware of being infected. Furthermore, treated patients might not adhere to the treatment, and new infections happen also in countries where treatment is widespread [6]. For this reason, prevention policies share a central role with drug based intervention strategies to limit the spread of the disease.
The inference of transmission trees, clustering together donor-recipient pairs, gives deeper insights on the disease spreading patterns. Data and methods to construct these trees are of major interest for the understanding of, for example, the efficacy of the treatment, the disease dynamics in risk groups and the identification of the contributors to the spread of the disease. One possible method is to perform contact tracing on an epidemics subgroup, but it is complicated and often inaccurate [7]. The most informative and readily available data on disease dynamics are viral sequences from hosts infected by a measurably evolving pathogen. Tools to correctly infer transmission chains from viral sequences must thus be developed and tested. Inferring the unknown transmission history within an epidemics has paved the way to a better understanding of the infection spread dynamics and to potential control strategies.
This work reports the developing of a multiscale (within and between-host) HIV viral and evolutionary dynamics software for the investigation of the effects of within-host evolution on the performance of transmission tree reconstruction. More specifically, I will look into the effects different viral mutation rates and different host blood sampling times have on the quality of the transmission tree reconstruction of an epidemics. While host sampling times are not a within-host parameter, it results in sequences evolving for a different amount of time before being sampled, and thus actively influences the effects of within-host evolution. In the subsequent paragraphs, I will describe some of the extreme complexity of the HIV virion and its infection, in order to subsequently highlight the features deemed most important for inclusion in the model. The practice of transmission tree reconstruction is presented, together with an overview of some of the modelling techniques covering different aspects of the HIV infection, including its within-host dynamics and why a new model is deemed necessary.

The Virion HIV is a retrovirus with two 10 kb long RNA strands. Its genome consists of ten genes: *pol*, *gag*, *env*, *tat*, *rev*, vif, *vpr*, *vpu*, nef, nef and *tev*. *Pol*, coding for the protease, the reverse transcriptase, the RNase H and the integrase, *gag*, coding for capsid proteins, and *env*, coding for the $gp120$ and $gp41$ viral glycoproteins, are arguably the most understood and sequenced HIV genes. The genome is enclosed in a capsid composed of viral proteins and a lipidic envelope hosting viral glycoproteins.
The partial elucidation of HIV cell entry and infection mechanism was possible by many experimental efforts and theoretical works in the field of mathematical modelling [8] [9] [10] [11] [12]. Among host HIV transmission happens through mucosal membranes most

often by just one or few viral variants, and mainly targets activated $CD4^+$ cells [13]. Other cells types that can also be targeted by HIV infection include macrophages [14] and dendritic cells [15] [16]. After successful infection, HIV rapidly disseminate in gut, genitalia, gastric mucosa, liver, central nervous system, lymph nodes and breast tissue [17] [18] [19] [20], while a compartment of latent cells is maintained [21]. Cell entry is mediated by interaction with the CD4 coreceptor and at least one of the coreceptors CCR5 or CXCR4 [22]. Once entered, the viral reverse transcriptase acts on the viral RNA content and to produce a double stranded DNA, which is then integrated in the host cell genome. This step is particularly error prone, as the reverse transcriptase lacks exonucleolytic proofreading [23], resulting in a mutation rate of around $2{\cdot}10^{-5}$ mutations nt^{-1} cycle^{-1} [24]. Due to the fast mutation rate, HIV exists in a host as a quasispecies, the ensemble of all the mutants generated by the mutation-selection process, with selection exerted, in absence of treatment, by the host's immune system [25]. The innate immune system, particularly the natural killer compartment, responds rapidly and not specifically to the infection. Soon after infection, the adaptive immune system potently responds in a strain-specific way, selecting for escape variants [26], less susceptible to immune clearance [27] [28].

Transmission trees In order to unravel the transmission history of an HIV epidemics, one could perform contact tracing interviews with all the subjects studied, reconstruct the contact network and, thus, the transmission history [29] [30]. This approach easily becomes unfeasible with epidemics of bigger sizes, because of the need for complete and accurate epidemiological parameters like the infection times or the geographical location [7]. HIV within-host sequence evolution happens at a faster pace than between-host transmission events [31], thus a more recent approach proposes the exploitation of information on sequence similarity and sampling time of the infected hosts to extract new insights on the transmission pairs: the relatedness of the viral strains sampled, the phylogeny, is used to infer the transmission tree with the assumption that similar viruses are close in the transmission history [32] [33] [34] [35]. Blood samples from the hosts are sampled, viral strains are sequenced by direct Sangers bulk sequencing [36] [37] [38] or next-generation sequencing (NGS) [39] [40]. The sequence composed by the most frequent nucleotide at each loci is named consensus sequence and need not be in the blood sample. Consensus sequences are generated for each host and a phylogeny thereof is inferred. Leitner *et al.* [41] were the first to analyze the performance of using phylogeny for inferring a small transmission history known by contact tracing. They were able to reconstruct the transmission events in 12 out of 13 taxa. Phylogenetic relatedness was subsequently used, for example, to investigate transmissions in criminal cases [42], to investigate a possible surgeon transmitting HIV to a patient [43], to investigate on a male accused of rape [44], to elucidate foot-and-mouth virus transmission pathway [45], to infer social network structure in HIV epidemics [46] or infer sexual and parenteral transmission [47]. The reconstruction of the transmission tree from the phylogeny is all but trivial, as their topology can differ as a consequence of the earlier coalescence time of the leaves in the former as compared to the latter, incomplete patient sampling and within-host evolution. HIV measurably evolves within a host at a faster rate compared to the between-host transmission rate, possibly confounding transmission tree reconstruction. Internal nodes in a phylogeny represent the most recent common ancestor between two lineages, whereas in a transmission tree they represent the time of transmission. It follows that the transmitted strain coalescence in the phylogenetic tree must predate the infection time. Though phylogenetical reconstruction of transmission trees has become widespread, the data it yields ever more precious and the methods more advanced [31], caution about the basic assumptions of this method is



(a)

(b) (c) (d)

Figure 1: Between-host transmission with within-host evolution. Different colors represent different hosts. (a) The B host infects first the G host, then the R host. The within-host evolution occurs independently in every host, and transmissions are represented by immigration of a single viral strain from the donor to the recipient host. Sampling of a single viral strain per host was conceptually exemplified at three different times, represented by the orange, the blue and the violet points. The ellipses represent the continuation of the within-host evolution of the respective phylogeny. (b) resulting phylogram for the orange sampling. (c) resulting phylogram for the blue sampling. (d) resulting phylogram for the violet sampling. The dashed green and red lines represent the transmission times of the B host to the G and to the R hosts, respectively. The most recent common ancestor of each phylogram does not have to coincide with the transmission time.

necessary [48] [38] [49]. Here I propose a framework for the simulation of an HIV epidemics within- and among hosts and use it to test the influence of viral within-host evolution on transmission chain reconstruction. The interplay of within-host evolution and between-host transmission is idealized in figure 1, with three hosts, called B, R and G. HIV evolves within each host independently, and transmission events consist of the immigration of a virion of a specific strain from the donor to the recipient host. The effect of the within-host evolution is highlighted by the conceptual sampling of a single viral strain for each host at three different time points. This shows how different samples of a single strain from each host can give rise to completely different phylogenies. Most notably, transmission times do not forcibly coincide with the respective most recent common ancestors in the phylogenies.

Modelling HIV infection A quality measure of the reconstruction of transmission trees through phylogenetical inference is necessary. The validation of the method against real-world data through contact tracing is troublesome because of the high amount of high quality epidemiological data needed. Another possible route is the simulation of a transmission history according to an epidemiological model with explicit sequence mutation. The goal is taking into account not only the between-host transmission history, but also the within-host dynamics and possible bottlenecks at the time of transmission. One can subsequently reconstruct the phylogenetic tree from the simulated samples of sequences through the epidemics and compare it to the true transmission tree. In these 30 years of HIV epidemics, multiple attempts at simulating characteristic traits as the dynamics of infection [50] [51] [52] or the diversity and evolution [53] were performed. For a more complete review on the various HIV modelling approach, I referred to [54]. Typical viral dynamics models are an expansion of a target cell limitation model

$$\frac{dT}{dt} = \lambda - \beta VT - \delta_T T \tag{1a}$$

$$\frac{dI}{dt} = \beta VT - \delta_I I \tag{1b}$$

$$\frac{dV}{dt} = \kappa I - \beta VT - \delta_V V \tag{1c}$$

where T, I, V are the target cell, infected cell and virion compartments, respectively. λ represents the constant supply of new target cells, β the within-host infection rate constant, δ_i the removal rate constant of compartment i, and κ the burst rate constant of new virions from infected cells. The various expansions incorporate latency [55], the AIDS progression phase [56] cell-cell infection [57] [58], antiretroviral therapy [59] [60] [61], multiple spatial compartments [62] and simulation scales [63]. Genotype-aware simulations aimed at addressing questions about the evolution or the diversity of HIV usually integrate a coalescent model [64] of within-host evolutionary dynamics for phylogeny reconstruction and then sequence simulation along the branches of the phylogenetic tree [32] according to one among the most frequent nucleotide evolutionary models [65]. The development of a comprehensive mechanistic model of dynamics and evolution encompassing all of the aspects discussed so far is difficult, partly due to the overwhelming complexity of the system, partly due to the lack of knowledge thereof. A more specific mechanistic model focused on a specific aspect of the disease could potentially yield relevant and intuitive insights on HIV dynamics bridging within and between-host, like the identification of the confounding factors of transmission trees reconstruction from viral phylogenies.

Exploring the influence of evolutionary dynamics on transmission tree reconstruction A model should present a minimal set of characteristics to provide consensus sequences

from infected hosts coherent with the viral evolutionary history. This set encompasses the presence of transmissions among hosts, the presence of a within-host viral dynamics and that this viral dynamics be linked directly to the evolution of the sequence. The sequence evolution is driven by some external selective pressure. Typical coalescent models as described above, while returning a phylogenetical history consistent with the viral population change and usually modelling transmission events in the host population, do not link the within-host viral dynamics to the evolution of the sequence explicitly [49]. In a first step, cherries typically coalesce to an ancestral lineage in an exponentially distributed time with rate dependant on the population. In a second step, sequences may be simulated along the branches of the generated coalescence tree according to a substitution model. So, the coalescence events are conditioned on the population dynamics, but have no direct effect on it, being simulated only afterwards [66] [67]. This is correct only in case of the mutation of evolutionary neutral sites. A more direct approach is the modelling of a fitness landscape within which HIV evolves escaping a selective pressure, for example antiretroviral treatment or immune pressure. This way, both the viral dynamics and the selective pressure leave an intrinsic, explicit mark in the consensus sequence sampled from the host as result of the implementation of the model. The model presented in this work aims at doing exactly this.

2 Methods

The model In order to explore the various possible confounding factors of transmission tree recovery from its correspondent phylogenetic tree, in this work I develop a mechanistic model of within and between-host dynamics and evolution of HIV-1 in the blood compartment in absence of treatment. This implementation links within-host viral dynamics to sequence evolution through a fitness landscape modelling. Three different scenarios can arise as a result of a random mutation along the sequence: (i) if it occurs in a genomic region corresponding to an epitope recognized by the immune system of the host, that particular strain will escape immune pressure, thus gaining an evolutionary advantage and increasing its fitness. If the genome is mutated anywhere else, then a random draw decides whether the fitness (ii) remains the same, modelling a neutral mutation, or (iii) decreases, modelling a disadvantageous mutation leading to a less efficient protein. In such a way, the HIV quasispecies will evolve to escape the immune system response, while maintaining a slow changing signal of neutral mutations and a faster changing signal of the heavily selected genomic regions. In the current implementation of the model, the fitness may have an influence on the burst size of new virions from infected cells, on the removal rates of both virions and infected cells, and on the infection rate of new target cells by the virions, or any combination thereof. In this work, fitness is considered to only influence the burst size of new virions, i.e. it is meant as *replicative* fitness, f_i^b. The presence of multiple strains in the model changes equations (1) in its multistrain analogue

$$\frac{dT}{dt} = \lambda - \beta \sum_{i=0}^{n} \left(f_i^{inf} V_i \right) T - \delta_T T \tag{2a}$$

$$\frac{dI_i}{dt} = \beta \sum_{j=0}^{n} \left(f_j^{inf} V_j q_{ji} \right) T - \frac{\delta_I}{f_i^{Id}} I_i \tag{2b}$$

$$\frac{dV_i}{dt} = \kappa f_i^b I_i - \beta V_i f_i^{inf} T - \frac{\delta_V}{f_i^{Vd}} V_i \tag{2c}$$

where f_i^k is different from 1 in this work only if $k = b$, as in the term $\kappa f_i^b I_i$ in equation (2c)

Event	Rate expression	Expected value after a day
Infection of a single target cell	$\beta \sum_{i=0}^{n} \left(V_i\right)$	$1 - \exp\left(-\beta \sum_{i=0}^{n} \left(V_i\right)\right)$
Removal of a single element of compartment k	δ_k	$1 - \exp\left(-\delta_k\right)$
Burst of new virions from a single infected cell of strain i	κf_i	$\begin{cases} 0, & \text{if } K \sim \mathcal{N}(\kappa f_i^b, \left(\kappa f_i^b/3\right)^2) \leq 0 \\ K, & \text{if } K \sim \mathcal{N}(\kappa f_i^b, \left(\kappa f_i^b/3\right)^2) > 0 \end{cases}$

Table 1: Selection of transformations from rates to probabilities used in the implementation of the model. The time unit is a time step for both rates and probabilities.

and subscript i refers to the i-th strain of the compartment. f_j^{inf} represents the *infection* fitness, the relative ability of each strain to infect target cells, which could potentially be the case of the insurgency of more aggressive and virulent strains of HIV. f_i^{Id} and f_i^{Vd} are the virions and infected cells *removal* fitness. The rationale behind the possible integration of the removal fitness is the selective pressure acting over cells and virions on not showing easily recognizable antigens. In order to simplify the model and allow an easier interpretation of the effects of the changing strain-specific fitness, only the replicative fitness was set different from 1 in the model. The q_{ji} factor is the probability of strain j to mutate into strain i upon infection. n represents the number of viral strains in the host. The other parameters are equivalent to equation (1). Even though the model formulation consists in a set of ordinary differential equations (ODE), which are by nature deterministic, in the implementation the rates are converted to probabilities through appropriate modelling of the events distributions (see table 1), making the within-host viral dynamics model by all means stochastic.

Another deviation from the set of ODE is that the time in the implemented model is not continuous, but discrete. In this case, how many events of each kind happen in a time step is recorded as time jumps forward, in contrast to the exact (and slow) stochastic implementation of the Gillespie algorithm [68], where exponential waiting times are randomly drawn according to the rates of the considered reaction channel [69]. As already mentioned, a comprehensive HIV-1 model is hardly doable: as any other, also this model comes with a set of limitations and assumptions. The target cell limitation model is only able to recover the acute and asymptotic stages of disease progression, but not the AIDS stage. Since the simulation time is much shorter than the time necessary for AIDS progression, this does not pose a problem. For the same reason, hosts are not removed over time. The short time-frame of the simulation also poses good ground for approximating set point viral load and $CD4^+$ cell count in the asymptotic phase at an approximately constant level. The usual decrease in $CD4^+$ cell level and the increase in viral load might not be enough for changes to be detected in the short time-frame considered. Simple mass-action kinetics assumptions impose that the solution be a well-mixed system, which is a both common and straightforward claim for the blood compartment. The lack of a latent cell compartment is not as critical in a not treated host as in a treated one. Nevertheless, it decreases the chance of transmitting ancient viral variants to new hosts [21]. HIV is

only assumed to infect $CD4^+$ cells, even though other cells, as macrophages and dendritic cells could potentially also be targeted.
Assumptions on the evolutionary behaviour of HIV are the strongest, even though limited evidence is available on the fitness landscape of the virus [70] [71]. As an approximation, the only selective pressure acting on the virus is the immune system of the host, which is reasonable, but does not differentiate between innate and adaptive immune responses, and is for each viral variants the same over time. The immune response was chosen such that in absence of mutation the infection would last around 20 days, and the fitness bonus and maluses, discussed above, were set to guarantee no host recovery. The fitness itself is replicative, and thus only has an effect on the viral burst size of the infected cells of the corresponding strain.
Between-host transmission happens according to a SI model with constant susceptible population: the simplicity of the model has the advantage of preventing overly complicated network-structure interpretations in favour of a leaner variant with little disadvantage in term of identification of confounding factors of transmission tree reconstruction.

The parametrization I chose to calibrate the model with experimental parameters inferred for the blood compartment. An overview of the chosen parameters is depicted in Table 2.

The implementation I implemented the model in the C++14 language standard for efficiency and performance. At execution time, it takes a file specifying the parameters set as argument, avoiding useless and time-consuming hard coding of the parameters and multiple compilations. I will refrain from commenting every line of the code, and I will instead focus on some of the most important points of the program. The overarching goal was to create a simulation tool that could be executed on a high performance computing (HPC) platform, and get in a couple of hours the results back. For this reason, the simulation volume had to be limited to 10^4 mm^3, some implementation details had to be changed and parallel computing had to be included.

Three algorithms are mainly determining the course of the simulation. Algorithm 1 is responsible for the time propagation of the within-host dynamics. The time spent calculating the probability and performing the Bernoulli experiment (lines 4-5) is negligible in comparison to the time spent performing and updating the weighted sampling without replacement procedure (line 8). In order to speed up this step, the efficient weighted sampler with replacement proposed by Vose [72] complemented by a rejection technique was implemented to sample from an approximate target distribution at a higher rate.
Algorithm 2 is responsible for mutating a sequence and assigning it to the correct strain, if present, or generating a new strain. Checking each new sequence for equality with the currently present ones is an expensive task. I approached this problem by hashing the first 50 characters of each sequence upon generation and storing the hash value with the sequence. At the time of comparison, only sequences sharing the same hash values are compared at every locus. If the hash value is different, no string comparison is needed. This is particularly useful as the sequence is rearranged to have the genomic region under stronger selection first, hence the most variant fragment is hashed, decreasing the collision probability.
Algorithm 3 simulates the between-host infection. The computation spends a negligible amount of time doing so, as the computation is short and the condition in line 2 is expected to be true once every 20 simulation days. Code parallelization offers another potential decrease in wall-clock time. Implementing shared memory parallelism fully takes advantage

Parameter	Reference	Value
within-host parameters		
Blood volume	-	$1 \cdot 10^4$ mm^3
T constant supply λ	[50]	20 mm^{-3} day^{-1}
Cell infection rate constant β	[50]	$2.4 \cdot 10^{-3}$ mm^3 day^{-1}
Initial T compartment size	[50]	1000 mm^{-3}
Initial V compartment size	-	100 mm^{-3}
Initial I compartment size	Healthy individual	0 mm^{-3}
T removal rate constant δ_T	[50]	0.02 day^{-1}
I removal rate constant δ_I	[50]	0.26 day^{-1}
V removal rate constant δ_V	[50]	2.4 day^{-1}
Burst size κ	-	10 mm^3 mm^{-3} day^{-1}
between-host parameters		
between-host infection rate $k_B H$	-	$5 \cdot 10^{-5}$ day $^{-1}$
Constant susceptible population S	-	$1 \cdot 10^3$
Evolutionary parameters		
HIV-1 mutation rate	[24]	$2.2 \cdot 10^{-5}$ nt^{-1} cycle^{-1}
Fitness bonus for advantageous mutation	-	0.4
Possible fitness maluses for disadvantageous mutation	adapted from [71]	-1.0 -0.6 0.0
Relative probability weight for disadvantageous mutations	[71]	$\frac{1}{10}$ $\frac{11}{15}$ $\frac{1}{6}$
Change of fitness due to immune response	-	-0.035 day^{-1}

Table 2: Overview of the adopted parameters and their reference. "-" as a reference means the parameter has been calculated or set in this work. For runs with a lower mutation rate, a mutation rate of $6.0 \cdot 10^{-6}$ nt^{-1} cycle^{-1}, fitness maluses of -1, -0.5 and 0.0 and immune response fitness decrease of -0.025 day^{-1} were used instead.

Algorithm 1: within-host dynamics

Data: within-host parameters, Random Number Generator (RNG), current state of T, $I[\,]$, $V[\,]$, mutation transition matrix $TMat$.

Result: Propagation of T, $I[\,]$, $V[\,]$ at next time step.

begin
  Initialize temporary infected cell array $C[\,]$ of the same size as $I[\,]$ containing newly infected cells that will undergo evolution;
  for $i \in T$ **do**
    Calculate probability that the i^{th} cell gets infected;
    Perform a Bernoulli experiment with this probability, assign its outcome to *infected*;
    if *infected* **then**
      Sample infecting strain according to relative virion abundance;
      Strain = sample(...);
      $V[Strain] \leftarrow V[Strain] - 1$;
      $C[Strain] \leftarrow C[Strain] + 1$;
      $T \leftarrow T - 1$;
    end
    if *No more virions* **then**
      break;
    end
  end
  evolve($TMat$, $C[\,]$,...);
  Burst new virions and remove V, I, T according to removal rate;
end

Algorithm 2: Evolve

Data: Mutation Transition matrix $TMat$, Random Number Generator, mutation probability p_mut, temporary infected cells array $C[\,]$.

Result: Infected cells correctly assigned to a strain after evolution.

begin
  for $s \in Strains$ **do**
    for $c \in C[s]$ **do**
      $C[s] \leftarrow C[s] - 1$;
      $N_{mut} \sim \mathcal{B}(sequence\ size, p_{mut})$;
      if $N_{mut} == 0$ **then**
        $I[s] \leftarrow I[s] + 1$;
      end
      else
        Sample N_{mut} positions on sequence;
        Mutate the nucleotide according to $TMat$ and replace it;
        if *Mutated sequence already present* **then**
          Increment i^{th} strain's $I[i]$ by 1;
        end
        else
          Generate new strain i with $I[i] = 1$, assigning newly calculated fitness;
        end
      end
    end
  end
end

Algorithm 3: between-host infection

Data: Current host state $H[\,]$ with respective viral array $V[\,]$, time to next infection t_{next}, between-host parameters.

Result: New infected host and strain infecting, new time to next infection t_{next}.

begin
  while $t_{next} \leq 0$ **do**
    Sample host index $K \sim \mathcal{U}(1, size\ of\ H[\,])$;
    Sample a viral strain from $H[K]$ according to within-host relative abundance;
    Initialize a new infected host with the sampled strain;
    Sample new infection time according to infection rate $S \cdot k_{BH}$
    $t_{next} \sim Exp(S \cdot k_{BH})$;
  end
  if $t_{next} > 0$ **then**
    $t_{next} \leftarrow t_{next} - 1$;
  end
end

of the architecture of HPC cluster nodes, and the program exploits directive based shared memory parallelism through OpenMP at a host level to efficaciously speed up computation in simulated epidemics with multiple infected hosts.

The output A successful simulation run yields as result three pieces of information for each host: a general overview of the within-host dynamics of virions, target cells and total number of cells in the simulated volume, a table with the amount and an identifying number of every viral strain at each time step and a file for each time step containing the identifying number, the sequence and the fitness of the viral strain at that time step.

Simulations In this work I will name "simulation" a realization of the model described by equation (2) using the software presented above. I performed seven simulations, referred to as "lower mutation rate simulation runs" with parameters as in table 2, mutation rate of $6.0 \cdot 10^{-6}$ nt^{-1} cycle^{-1} and immune response fitness decrease of -0.025 day^{-1}. The seven simulations differ only in the seed of the random number generator used. Afterwards, I performed 7 simulations with the parameters described in table 2, referred to as "higher mutation rate simulation runs", differing only in the random number generator seed. The transfer of the files of two of the latter simulations from the HPC cluster failed, resulting in a total number of higher mutation rate simulations of 5. The within-host figures (figures 10, 11) were constructed with a custom R statistics [73] script (supplmentary material 1, CEvo_Calls.R, CEvo_Analysis.R).

Sequence sampling A custom R statistics script (supplmentary material 1, CEvo_Calls.R, CEvo_Sequence.R) computationally sampled sequences from the output discussed above at different times through the simulation. Sequences were not sampled as in figure 1, i.e. just one viral strain, as it does not mirror the real-world situation. Instead, consensus sequences were constructed by sampling 100 sequences with probability proportional to relative amounts from each host after and before 30 and 60 days relative to the time of first transmission (labelled *Af30* and *Af60*, *Be30* and *Be60*, figures 2 and 3), at the exact time of transmission (*In*, figure 4), at random times (*Ra*, figure 5) and at the end of the simulation (*Ti*, figure 6) and setting at each locus the most frequent nucleotide of the sequence collection at that position. The transmitted sequence alone was also sampled (*Tr*). In case of no transmission event prior to the simulation end, or if the sampling time would be after the simulation end, the program sampled the last time step. The program writes the consensus sequences in a single FASTA format alignment and the sampling times in a separate table. The described sequence sampling are referred as "sampling schemes" in this work.

Phylogeny reconstruction For each sampling scheme the phylogenetical analysis was set up with the BEAUty utility software and run on the BEAST2 Bayesian inference software package [74]. The preparation requires the specification of whether to add information on the sequence partitioning, the sampling times, a site model, a clock model, the prior knowledge on the data and the Markov chain Monte Carlo (MCMC) preferences. I specified the alignment described above as not partitioned sequence collection, as in my model each site has the same mutation rate, and I did not want to give information on the sites under higher selective pressure, mimicking an analysis without prior genomic knowledge of the system. The recorded sampling times were used as tips times. A gamma distributed rates site model with 4 categories, with 0 as invariant proportion was used. I used a general time reversible (GTR) model with estimated nucleotide frequencies and substitution rates (except C$\leftrightarrow$T substitution, the default option) estimated according to an unbounded

gamma distributed prior with alpha parameter set to 0.05 and beta parameter set to 10 except for A↔G substitutions where the beta parameter was set to 20, as suggested by the default options. I set up a strict clock model with clock rate estimated according to a Log-Normal distribution prior with mean equal to the known mutation rate in real space and S parameter equal to 1.25 bounded at 0 and 0.001, a good range of possible substitution rates. Since there is no recovery from the infection in the proposed model, I assumed a Yule tree prior with an exponentially distributed birth rate with mean 20 days, the real transmission rate, bounded at 0 and 100 days. The MCMC chain length was $1 \cdot 10^7$, the default option, and trees were logged every $1 \cdot 10^4$ steps to save storage space. The TreeAnnotator software was then used to calculate the maximum clade credibility (MCC) tree from the BEAST2 tree distribution discarding the first 1% of the trees as burn-in and using mean node height.

Tree comparisons The MCC trees of the respective sampling scheme were collected in a single NEXUS file. The transmission tree reconstructed from the known simulated transmission chain was also added as a reference. Phylograms were plotted with the R package *ape*. The distance between the reconstructed phylogenies and the true transmission tree were used as indicator of the reconstruction performance. In order to identify the most suitable tree metric for subsequent analysis, three of them were compared on both the lower and the higher mutation rate simulation runs. The triplets distance metric [75] implemented in the R package *treeman* [76], the Kendall Colijn [77] and the Robinson Foulds tree metrics [78] implemented in the R package *treespace* were used to compare the trees deriving from each sampling scheme with the reference tree. The Kendall Colijn metric calculates the euclidean distance between two tree-specific vectors. These vectors are a linear combination dependent on a factor λ of a vector accounting exclusively for the topology and one that takes into account the branch lengths. I used $\lambda = 0$, i.e. only the topology vector, as I was interested in the correct placement of donor-recipient pairs more than in the correct timing of infection. The length of this vector is the number of possible combinations of two tips in the tree. The value of the elements of the vectors is the distance between the most recent common ancestor of the two tips and the root. The Robinson Foulds metric counts the number of splits induced exclusively by one of the trees. It is very conservative, as only clades that are perfectly shared do not increase the distance. The triplets distance compares the topology of every possible tip triplet in the tree. It gives as result the fraction of tip triplets shared by the two trees.
An intrinsic error is unavoidable and is due to the different tips sampling times (figures 2-6). This deviation can change the topology of the whole tree (figure 7). An additional reference was thus added to assess the influence of the sampling time alone on the quality of the transmission tree reconstruction. I scripted a custom piece of R code to construct the sample scheme-specific references and add them to the NEXUS tree comparison file (supplmentary material 1, CEvo_Calls.R, CEvo_Sequence.R). These are constructed by modifying the known simulated transmission tree to mirror the topology changes caused by the sampling (figure 7).
I summarized all replicas of the same mutation rate simulation run (same parameters, different seeds) with the three tree metrics, and constructed a boxplot for each metric. The respective metric was applied to compare each tree with the transmission tree reconstructed from the simulated transmission chain. The coloring of the scheme is color-blind friendly, and classifies the sampled phylogeny and the respective tree reference together, sharing the same color. The sampling scheme reference has the same label as the reconstructed sampling scheme phylogeny, but is complemented by the suffix *_Ref*. As an example, *Af30_Ref* is the specific reference of the epidemics sampled 30 day before the

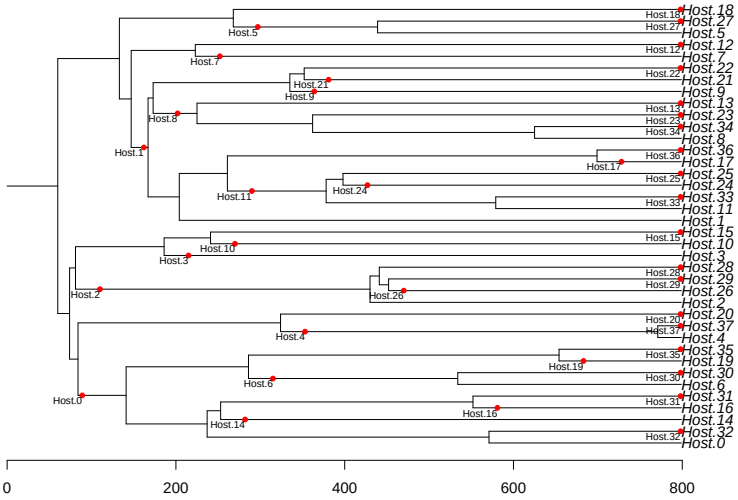


Figure 2: Sampling scheme labelled *Af30*. The hosts are sampled 30 days after their first transmission. If a host does not transmit to anyone, then it is sampled at the tip. The *Af60* sampling scheme is analogous, with sampling happening 60 days after transmission.

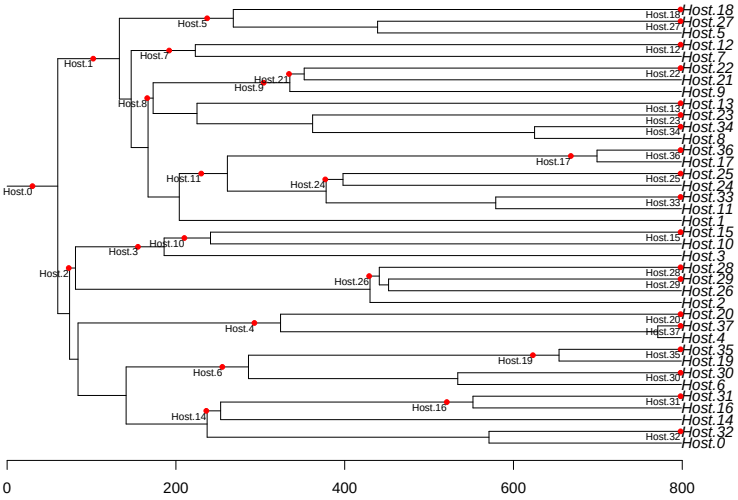


Figure 3: Sampling scheme labelled *Be30*. The hosts are sampled 30 days before their first transmission. If a host does not transmit to anyone, then it is sampled at the tip. The *Be60* sampling scheme is analogous, with sampling happening 60 days before transmission.

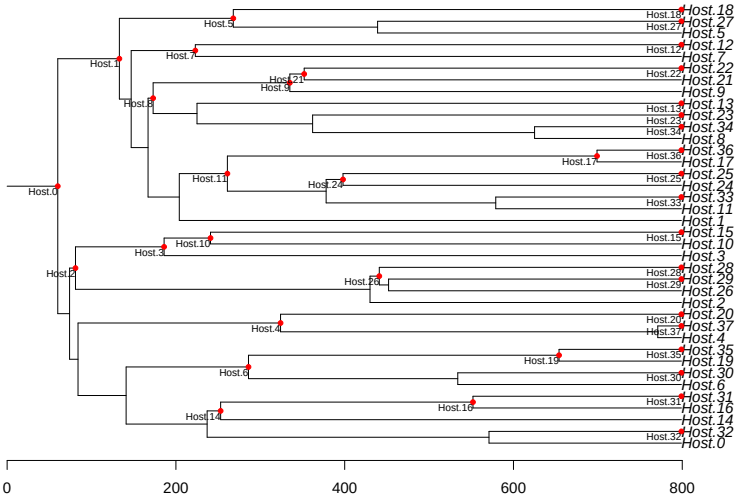


Figure 4: Sampling scheme labelled *In*. The hosts are sampled at the time of their first transmission. If a host does not transmit to anyone, then it is sampled at the tip.

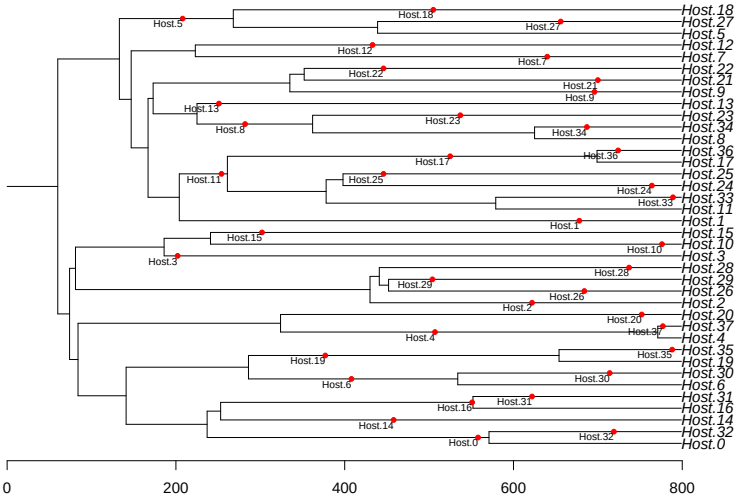


Figure 5: Sampling scheme labelled *Ra*. The hosts are sampled at random times.

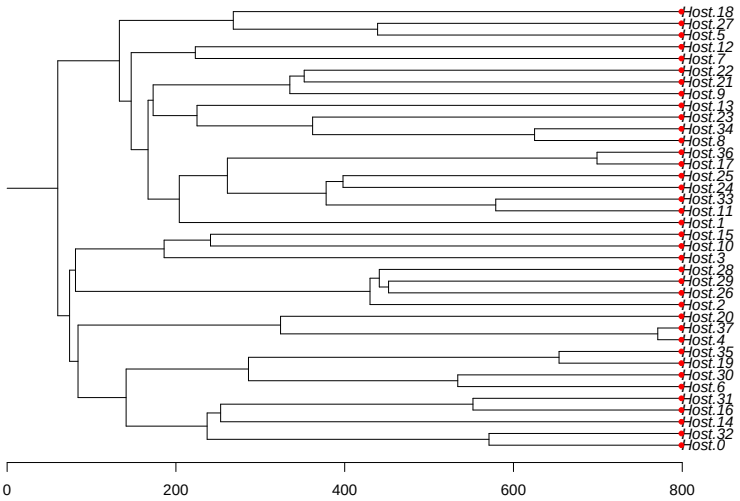
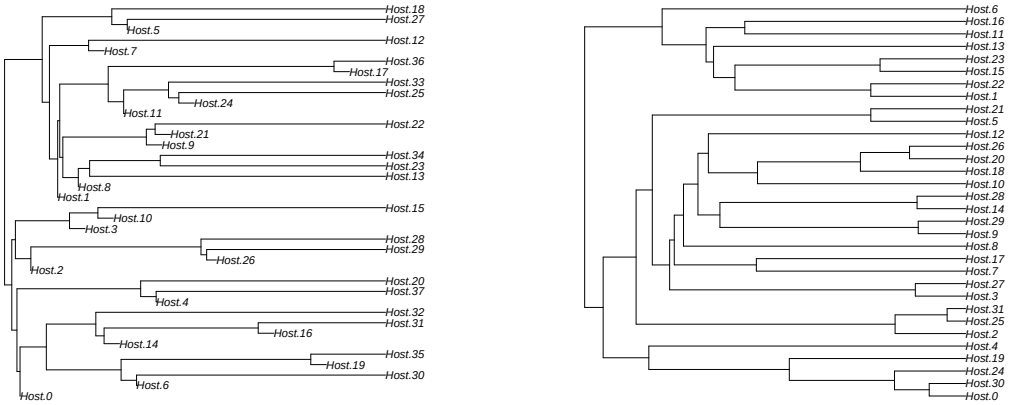


Figure 6: Sampling scheme labelled *Ti*. The hosts are sampled at the tips.



(a) Transmission tree with hosts sampled 30 days after the first infection of the host, or at the end if no infection occurs.

(b) Transmission tree constructed from known transmission history.

Figure 7: Comparison between the true transmission tree and the correspondent tree with hosts sampled at random times. Even without phylogenetic reconstruction error, the topology might change due to the different sampling schemes. Tip labelled "Host.0" on the lower part of both panels is an example thereof.

time of first infection.
The identification of a suitable tree metric for the comparison of the trees reconstructed from the simulations output is necessary to perform subsequent analysis. The tree distance from the known transmission tree would represent the indicator of transmission tree reconstruction performance: the shorter the distance, the better the performance.

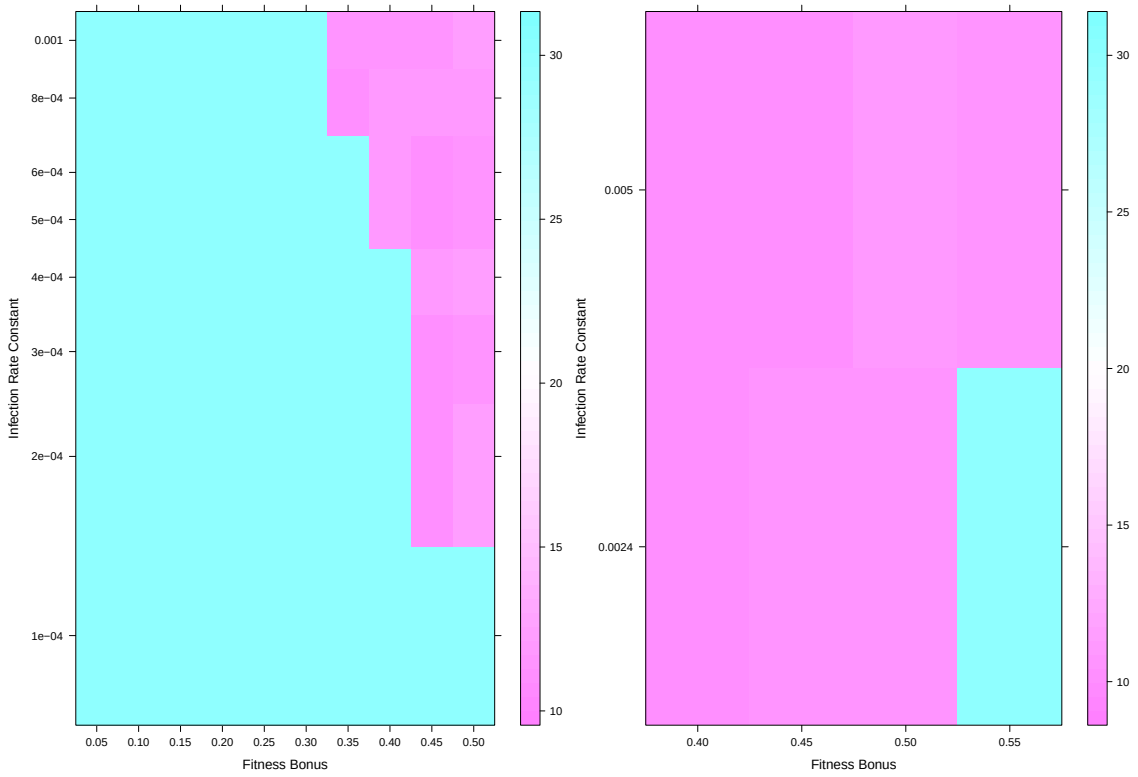
3 Results

3.1 Simulation parameters

As already mentioned above, I parametrized my model with a set of parameters consistent with the blood compartment. While there is a consensus about many parameters of HIV infection and evolutionary dynamics needed in this simulation, the literature is lacking confidence on some of the essentials ones. As a preliminary analysis, viable combinations of the missing parameters were identified. A combination was considered viable if the resulting simulation does not abruptly terminate before reaching the requested number of time steps, if no host recovers from the HIV infection and if the within-host dynamics mirrors approximately the expected one, i.e. viral loads of around 150 mm^{-3} and $CD4^+$ cell count not lower than 200 mm^{-3}. Small deviations from these target values were tolerated, as long as the overall dynamics was coherent.
Figure 8 summarizes the search over a parameter space composed by the two parameters: k_{inf} and the fitness bonus. A score was designed, that would allow an easy screening of wide search ranges. This score calculates the host-averaged natural logarithm of the integral of the viral load exceeding 150 mm^{-3}. The integral range is depicted in figure 9. This allows to discard the acute stage viral load peak, and to minimize the viral load to a more realistic level. I chose a viable combination of the two parameters minimizing the score measure and that yielded robust results (table 2), i.e. not on the edge of the viable range (pink region in figure 8), with the fitness bonus greater or equal than 0.40, and the infection rate constant greater than $8.0 \cdot 10^{-4}$ mm^3 day^{-1}.

3.2 Tree metric

In order to assess the most suitable tree metric for my analysis, in a preliminary analysis three of them were compared (figure 12) on the lower mutation rate simulation runs. It is important to find a tree metric that is adequate for the specific problem of comparing a possibly diverse set of trees. The Kendall Colijn tree metric (figure 12a) and the Robinson Foulds tree metric (figure 12b) share some critical points on this analysis for two reasons. (i) the sampling-scheme reference trees are not consistently close to the true reference transmission tree, i.e. their distances are not consistently close to 0 in comparison with the reconstructed phylogenies. This makes the imputation of the distance measured between a reconstructed phylogeny and the true transmission tree difficult, as it could be due to both a poor reconstruction and an inherent, unavoidable, error resulting from topology changes after sampling (figure 7). The amount of the error imputable to the change in topology due to the sampling scheme adopted is unknown without knowing the true transmission tree. (ii) the distance measure does not simplify the comparison between trees with different number of tips. The triplets distance measure (figure 12c) has more consistent sampling-schemes reference trees, and gives as distance a number from 0 to 1, which makes it suitable for comparing trees with different number of tips.
I gathered reference tree and reconstructed tree distances in two groups for each metric. A two-sided t-test on the two groups with the Welch correction for the degrees of freedom



(a) (b)

Figure 8: Heatmap of the score of simulations with different infection rate constants k_{inf} and fitness bonuses. The score is the host-averaged natural logarithm of the within-host numeric integral of the number of virions exceeding 150 mm^{-3} over the course of the simulation except the acute phase. A low score is index of a successful run. Failed runs were assigned a score of 30. (a) wide, early search. (b) subsequent more focused search around plausible combinations.

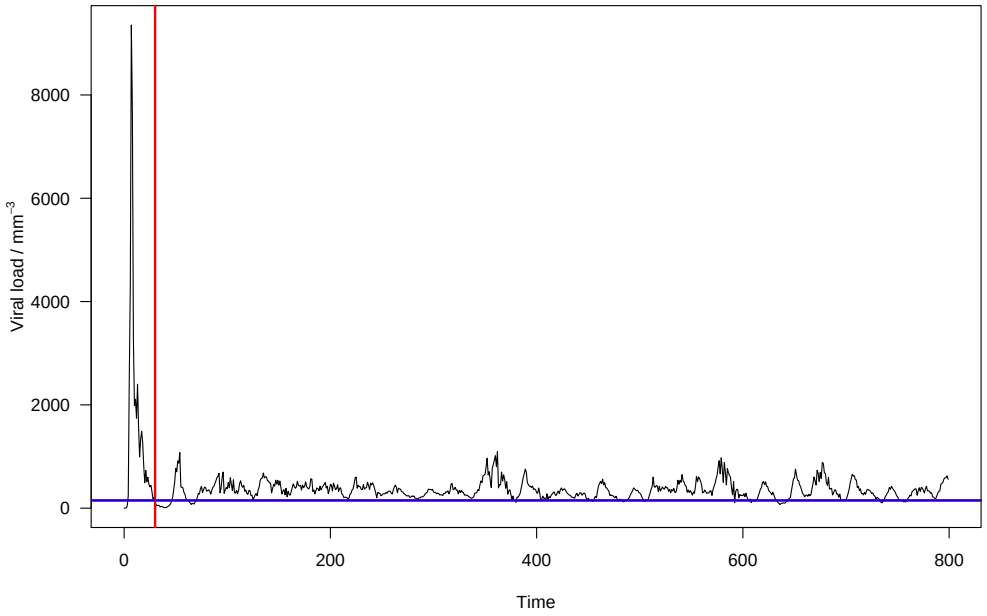
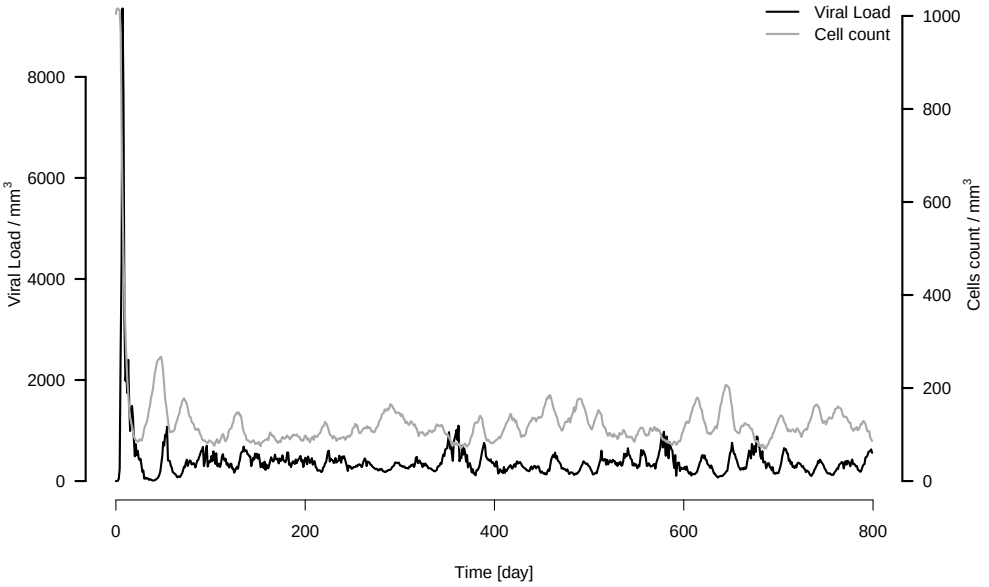


Figure 9: Bounds of the integral for the score calculation. The integral is computed in the range bounded by the red line and the end of the simulation, and takes into account just the values above the blue line.

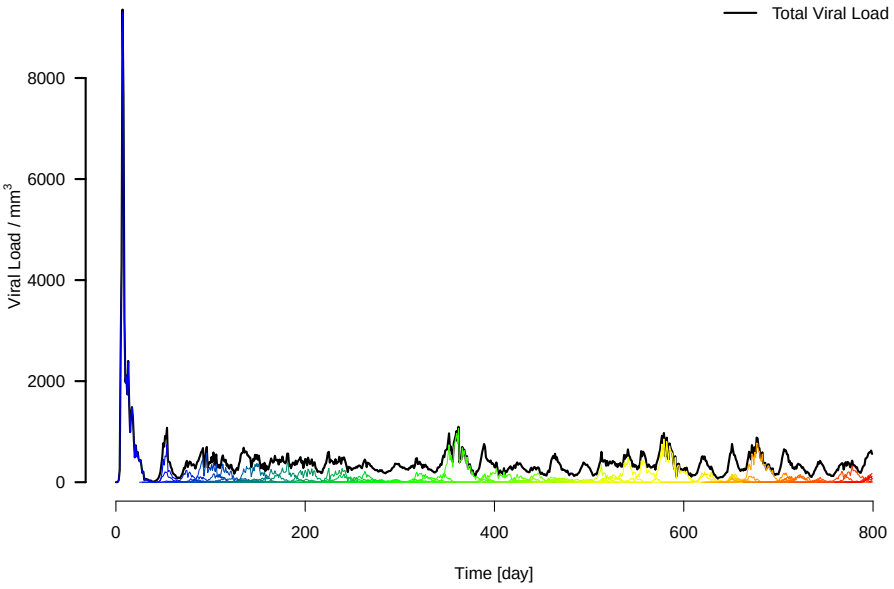
showed a significant difference in mean for each metric. The difference in mean of the two groups was 13.9 ± 11.2 ($p = 1.6 \cdot 10^{-2}$) for Kendall Colijn metric, 18.5 ± 7.0 ($p = 1.1 \cdot 10^{-6}$ for the Robinson Foulds metric and 0.32 ± 0.034 ($p = 3.7 \cdot 10^{-31}$) for the triplets distance. Results are reported as *Mean difference* $\pm$ 1/2 *Confidence Interval*. Figure 13 shows the result of an ancestry-based tree comparison: notably, correctly reconstructed subtrees might get a low score because of errors in the ancient branching history of the tree. As an example, the subtree composed of tip labelled *Host.10*, *Host.22*, *Host.36* and *Host.44* were closely reproduced in the reconstructed tree, but still were assigned low scores. Since the triplets metric does not present the inconveniences of the Kendall Colijn and the Robinson Foulds metrics in the ease of interpretation and comparison of trees having different amounts of tips presented in this section, I chose the triplets tree metric for further analysis in this work.

3.3 Within-host dynamics

Each host in a simulation corresponds to an independent realization of the model described in equation (2). Figure 10 shows the characteristic within-host viral dynamics of a simulation with a lower mutation rate, whereas figure 11 shows the within-host dynamics of a simulation with a higher mutation rate. In both cases, after an acute phase, lasting approximately 30 days, the viral load and the cell count wiggle around a constant value, reaching the set point. The total amount of cells stabilizes around 150 cells mm^{-3}, and the viral load around 400 virions mm^{-3}. Over time, only few strains dominate the quasispecies fitness landscape, and succeed in waves. The next generation overtakes the previous and possibly dominates the quasispecies fitness landscape.

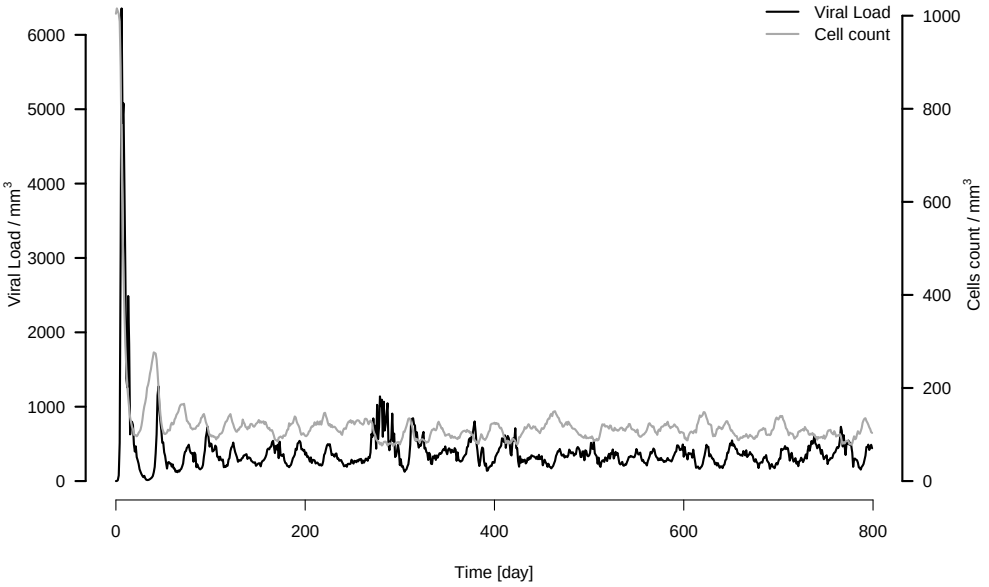


(a)

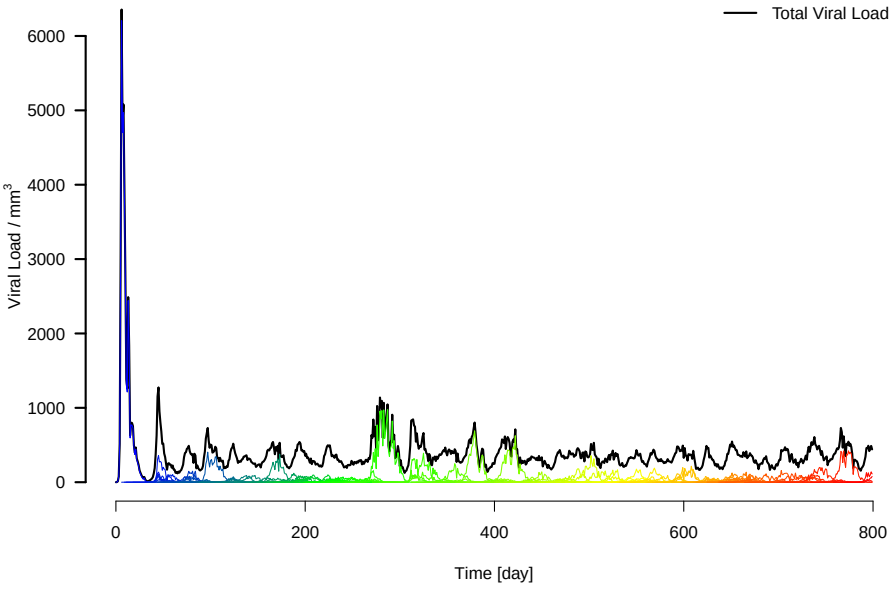


(b)

Figure 10: Example of within-host viral dynamics of a lower mutation rate run. 10a: trajectory over time of the viral load (black line) and the total cell count (gray line). The axis are different. 10b: the most frequent viral strains are plotted together with the total viral load. Colors show the appearance time of a strain, blue being earlier, red being later.

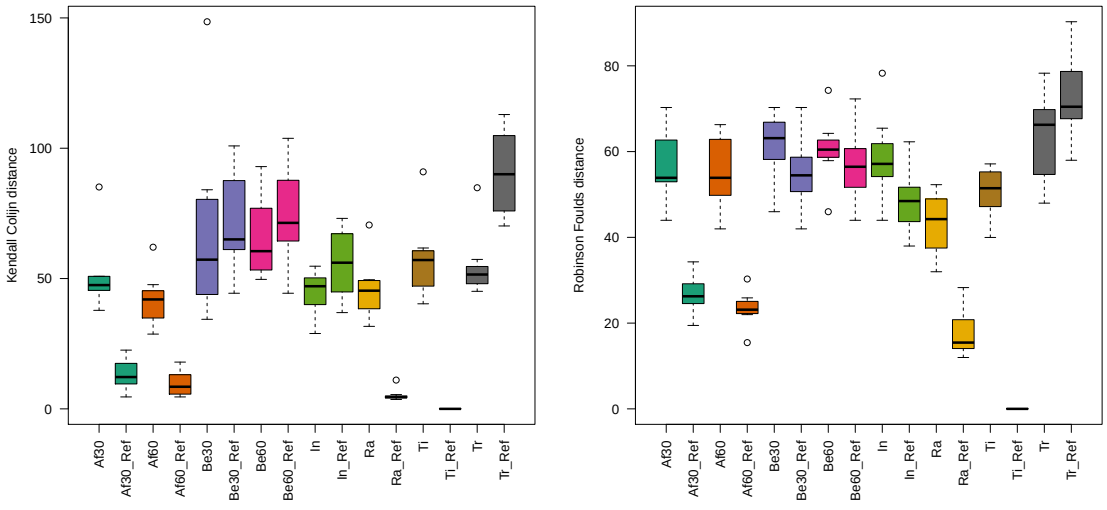


(a)



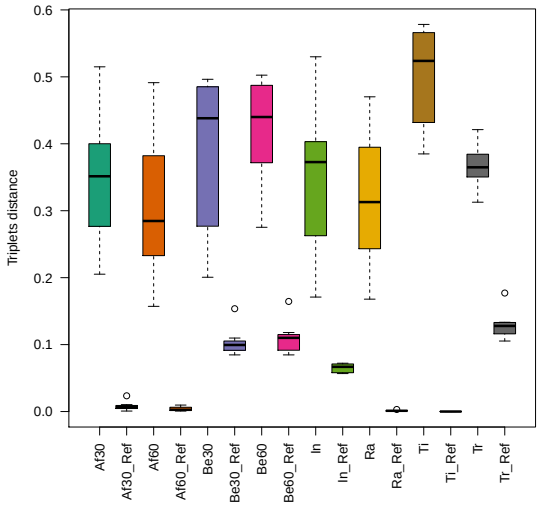
(b)

Figure 11: Examples of within-host viral dynamics of a higher mutation rate run. 11a: trajectory over time of the viral load (black line) and the total cell count (gray line). The axis are different. 11b: the most frequent viral strains are plotted together with the total viral load. Colors show the appearance time of a strain, blue being earlier, red being later.



(a)

(b)



(c)

Figure 12: Comparison of different tree metrics on a simulated data set. The labels are the abbreviation of the sampling schemes followed by the number of days and the *_Ref* tag for reference trees. *Af*, *Be*, *In*, *Ra*, *Ti* and *Tr* stand respectively for *After*, *Before*, *At infection time*, *At random times*, *At the tips* and *Just the transmitted strain*. Depicted are the distances between the labelled tree and the known transmission tree.

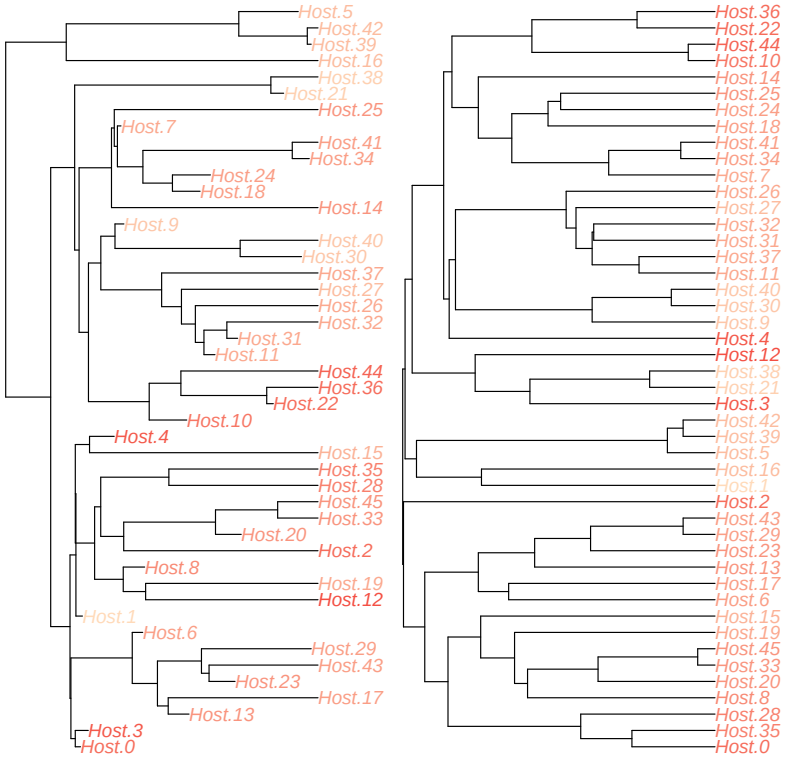
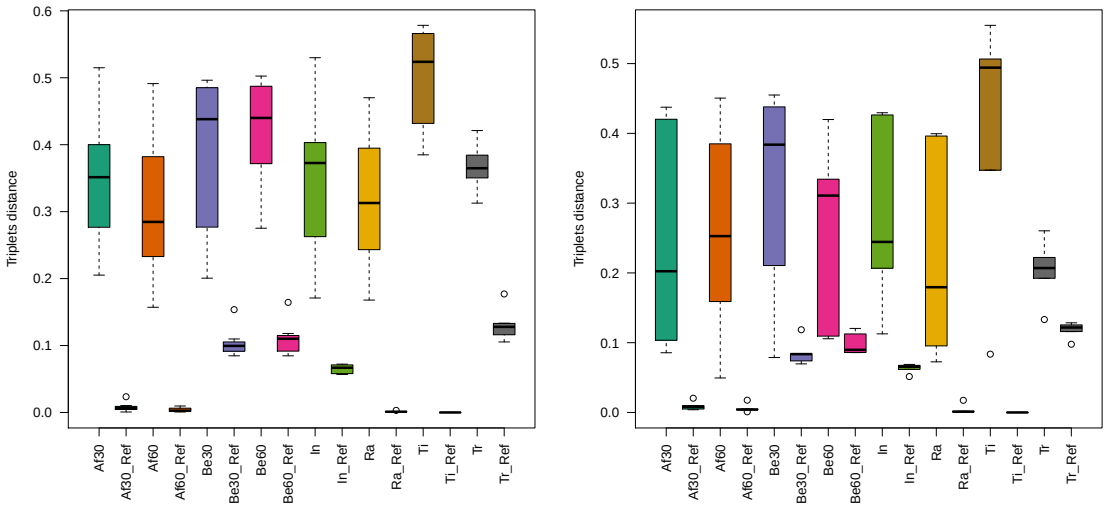


Figure 13: Tree comparison with the number of ancestral difference for each tree tip between the simulated transmission tree (right panel) and the tree sampled after 60 days from the first infection (left panel). Lighter colors of the tip label mean less difference between ancestry in the two trees. Notable is the influence of deep coalescence events on the asserted quality of correctly reconstructed subtrees.



(a) (b)

Figure 14: Comparison of transmission tree reconstruction with different sampling schemes. 14a: lower mutation rate simulation runs data. 14b: higher mutation rate simulation runs. The tree triplets metric was employed analogously to figure 12 to compare the distances of the various sampling schemes to their, and the overall, reference tree.

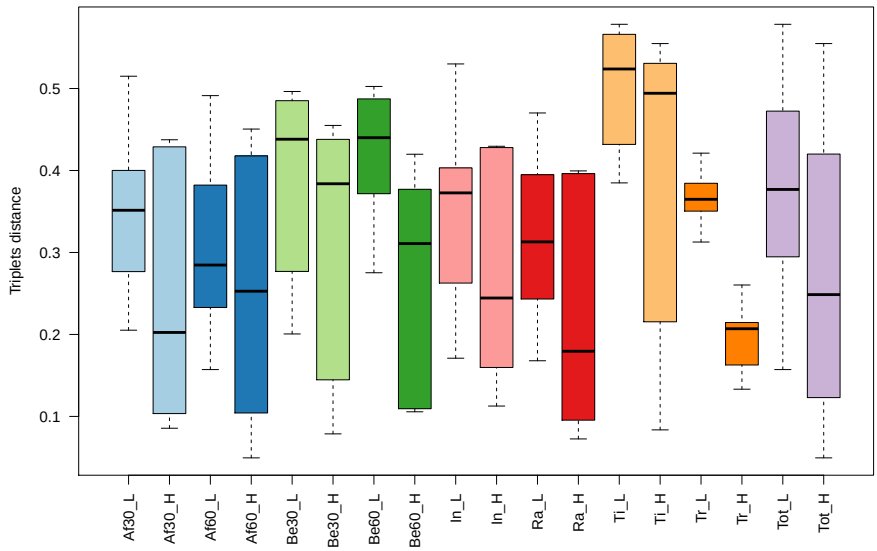
3.4 Sampling time

I sampled sequences from both mutation rate simulation runs, reconstructed the respective phylogenies and compared as described in the methods section, and summarized the distances of the resulting phylogenies with the true transmission tree in figure 14. For both mutation rate simulation runs, respectively, each sampling scheme was pairwise compared against every other through a double sided t-test with variance estimated with the Welch approximation of the degrees of freedom. Reference phylogenies were excluded from this analysis, as its goal was to detect changes in the phylogenetic reconstruction performance. After Bonferroni correction for multiple testing (28 t-tests), no significant difference in mean between the distance of any pair of sampling schemes in neither of the two simulation runs was detected (see supplementary materials 1, Compare_Sampling_Times.R for analysis and data).

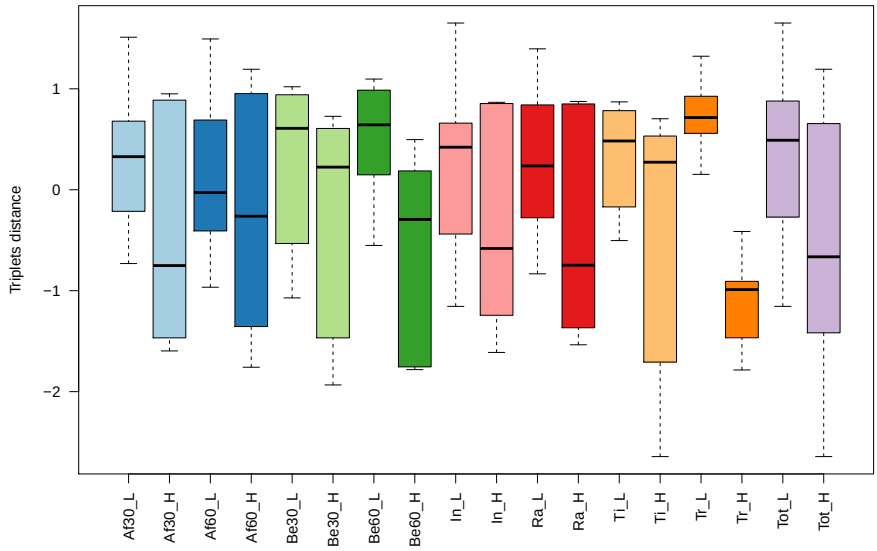
3.5 Different mutation rates

I performed two runs as described in the methods section, with differing mutation rate (see table 2 for reference), and compared the results with the triplets metric.

In figure 15 the median of the reconstructed phylogeny distances is higher in the lower mutation rate simulation runs than in the higher mutation rate ones, and the interquartile range in narrower. Additionally, the pooled sampling scheme specific data for both mutation rates runs were centered and the respective standard deviation was standardized (as clarification example, the data from *Af30_L* and *Af30_H* were pooled, the mean thereof was subtracted, and then the data were divided by the pooled standard deviation). I performed a double sided t-test with variance estimated with the Welch approximation of the degrees of freedom to test the difference between the triplets distance of the lower mutation rate simulation runs (n = 7) and the higher mutation rate simulation runs (n =



(a)



(b)

Figure 15: Comparison of transmission tree reconstruction with different sampling schemes. Each pair sharing the same colour represents the same sampling scheme from the low mutation rate runs (suffix *_L*) and the high mutation rate runs (suffix *_H)*). The *Tot* label refers to the pooled data from every sampling scheme. 15a: unstandardized data. 15b: the data were centered according to the pooled mean of the low and high mutation rate runs sampling scheme, and then divided by the pooled standard deviation. The tree triplets metric was employed analogously to figure 12 to compare the distances of the various sampling schemes to their, and the overall, reference tree.

5) for each sampling strategy for the unstandardized data and for the standardized data. This resulted in the absence of any significant difference in both datasets, except for the *Tr* sampling scheme ($p_{adj,unst} = 0.0024$ and $p_{ads,std} = 5.5 \cdot 10^{-5}$, adjusted with the Bonferroni correction for 8 tests, one for every sampling scheme). While the differences are not significant, the trend towards having a better reconstruction performance in the higher mutation rate simulation runs is consistent with every sampling scheme. Two additional tests were thus performed on the totality of the data from the high and low mutation rate simulation runs. There is no significant difference on the 95% confidence level after Bonferroni correction for testing both the totality of unstandardized data ($p_{adj,unstd} = 0.152$, Bonferroni correction for 4 tests, 2 overall sampling scheme specific tests and 2 tests on the pooled data). There is a significant difference in the means of the pooled data of the standardized dataset ($p_{adj,std} = 4.8 \cdot 10^{-4}$, Bonferroni correction for 4 tests). Uncorrected test results are summarized in table 3.

Label	Low mutation rate Mean ± SD	High mutation rate Mean ± SD	p-value
Unstandardized data			
Af30	0.346 ± 0.106	0.250 ± 0.169	0.302
Af60	0.309 ± 0.122	0.259 ± 0.163	0.583
Be30	0.380 ± 0.130	0.313 ± 0.163	0.471
Be60	0.419 ± 0.092	0.256 ± 0.141	0.062
In	0.344 ± 0.123	0.284 ± 0.140	0.466
Ra	0.318 ± 0.113	0.229 ± 0.156	0.319
Ti	0.498 ± 0.083	0.397 ± 0.192	0.322
Tr	0.367 ± 0.035	0.203 ± 0.047	$2.7 \cdot 10^{-4}$
Pooled	0.213 ± 0.113	0.161 ± 0.149	0.038
Standardized data			
Af30	0.291 ± 0.770	-0.365 ± 1.223	0.257
Af60	0.152 ± 0.898	-0.233 ± 1.300	0.532
Be30	0.196 ± 0.917	-0.386 ± 1.208	0.331
Be60	0.493 ± 0.700	-0.674 ± 1.061	0.033
In	0.194 ± 0.964	-0.301 ± 1.145	0.399
Ra	0.275 ± 0.836	-0.349 ± 1.156	0.272
Ti	0.297 ± 0.592	-0.574 ± 1.498	0.191
Tr	0.737 ± 0.379	-1.134 ± 0.501	$6.8 \cdot 10^{-6}$
Pooled	0.329 ± 0.747	-0.462 ± 1.046	$1.2 \cdot 10^{-4}$

Table 3: Summary of the t-tests comparing the means of the distance of the labelled phylogenies between the simulation runs with low mutation rate ($n_{sim} = 7$) and high mutation rate ($n_{sim} = 5$). "Pooled" represents the t-test comparing the mean of two groups composed by pulling together the data of the low ($n_{low} = 56$) and high ($n_{high} = 40$) mutation rate runs, respectively. The p-values are not adjusted for multiple testing. The same procedure is applied to the standardized and unstandardized datasets.

4 Discussion

The overarching goal of this work was to investigate the influence of viral within-host evolution on transmission chain reconstruction. With this aim, I developed a mechanistic

simulation tool for within-host evolution and between-host transmission. The analysis consisted in the comparison of the phylogenies reconstructed from the simulation with the known transmission history of two sets of simulations, one with a lower mutation rate, one with a higher mutation rate. The phylogenies were reconstructed from sequences sampled according to 8 different sampling schemes to assess the influence of sampling times on the reconstruction.

As a preliminary exploration, three tree metrics were compared, in order to find which one would suit the above-mentioned analysis best. The Kendall Colijn is a very useful metric for comparing and classifying trees resulting from Bayesian inference, like a MCMC run on the BEAST2 software package. It is able to incorporate the importance of the branch length at an extent of choice, and can also just take into account the topology of the trees. In the latter case, the distance measure is the euclidean distance between two tree-specific vectors. These vectors have a length equal to the number of MRCA of the tips, and an element value equal to the number of splits separating the MRCA to the root. The Robinson Foulds is one of the most widespread tree topology metrics. It relies in identifying the tree splits that are present in one of the trees, but not in the other one. While intuitive and popular, it is very conservative, not increasing the distance only if the considered clades are exactly identical. These two metrics are not suitable for comparing distances originating from different simulations. They result in a not normalized distance measure, the interpretation of which is not obvious for trees with a different amount of tips. Moreover, the Kendall Colijn tree metric seems to be less adequate for phylogenies whose deep branching events are inaccurately reconstructed (figure 13). This was the case in the reconstruction of phylogenies with little information on the early stages of the epidemics. The triplets distance metric measures the fraction of triplets that are equal in the two trees, so the distance relative to the true transmission tree in each simulation is a number between 0 and 1 for any number of tips. The tree metrics exploration led to the choice of using the triplets tree metric as tree comparison measure. The identification of a suitable tree metric is central for the quality of subsequent analysis. Designing a tree metric is a complicated matter, as the shrinking of a multidimensional object as a tree to a scalar number causes a drastic dimensionality reduction. The result of this preliminary analysis does not thus imply that the Kendall Colijn metric, or the Robinson Foulds metric perform worse in general: it just states that the triplets metric is more suitable for this specific kind of analysis.

The simulation tool mechanistically realizes the model described in equation (2) at a within-host level independently for every host. Two infection stages were included: the acute phase and the asymptotic phase. These stages are qualitatively well resolved in the simulation (figures 10 and 11). The acute phase lasting around 30 days is followed by the stabilization of the set point viral load. Quantitatively, some deviations with respect to the *in vivo* values are expected, as the target cell limitation model hardly can cover all the details the extremely complex system presented in the introduction. The subsequent simplifications are then mirrored in simulated values deviating to a certain extent from the expected ones. In particular, the cell count at the asymptotic stage is around 150 cells mm^{-3}, which would already be a level closely preceding the AIDS phase *in vivo*. The viral set point, conversely, is higher than expected. Both these value are in the correct order of magnitude of typical *in vivo* HIV dynamics values, making the deviation acceptable as part of the unavoidable model assumptions. A remarkable fact is that the overall within-host dynamics results from a complex strain specific turnover: there is not a single strain mutating over time, but a constant rise-and-fall mechanism. Old strains are outgrown by newer, fitter strains. The quasispecies thus migrates through the fitness landscape determined by the immune response throughout the simulation.

Sequence length and simulation volume were chosen according to the trade-off between computation time and simulation accuracy. A sequence length of 1000 bp is hardly representative of the whole HIV sequence, which is an order of magnitude longer. The fitness landscape implied in the sequence is also much narrower, influencing the viral dynamics through the fitness according to equation (2). The typical human blood volume of some liters is several orders of magnitude more than the simulated 10 mL. In a higher volume, more mutations can happen at a given time step, and more viral variants are created. With the quasispecies unfolding wider over the fitness landscape, the chance of finding a maximum in the latter is higher. The viral dynamics is thus more than directly dependent on the volume. Increasing the mutation rate proportional to the ratio of the volumes does not solve the issue, either. The total number of virions, in fact, is also affecting the evolutionary dynamics of the system: more virions mutating cause a more efficient and wider search over the fitness landscape. Other than the immune system, an additional selection force shaping the fitness landscape could be represented by the addition of antiretroviral treatment to a fraction of the hosts.
In subsequent versions of the simulation tools, more effort must be diverted into increasing the efficiency of the simulation in order to overcome, at least in part, the so far demanding computing time and memory hurdles. Another point that will be included soon is the addition of a latent compartment. This has been indicated as one of the reasons HIV treatment has not yet succeeded in defeating the infection definitely [5]. It could add a layer of realism to the viral dynamics, influencing the range of within-host parameters available and possibly making the calibration of within-host dynamics to *in vivo* values easier. Moreover, it could allow to explore the hypothesis of ancient strain transmission [79][80] over quick adaptation to the recipient host by recent viral variants [81] in the donor host. Another future possibility could be the inclusion of the AIDS disease stage. In order to do so, three routes, possibly combined, should be explored: (i) the inclusion of a time dependent random variable parameter, in such a way that after a certain time, the fitness of a new viral strain overcomes the defense by the immune system, (ii) the expansion of the viral infection to other kinds of cells, as described in the introduction, expanding the target repertoire, and (iii) linking the HIV fitness explosion and consequent immune system failure to a certain phenotype on the virus, which would be reached after some time. While easy to implement, point (i) has the disadvantage of not having a direct mechanistic interpretation. Point (ii) would still need a mechanism to trigger the target repertoire shift, which could be represented by point (i) and (iii). Point (iii) needs special attention, in order to avoid the possibility of transmission of virions close to or already presenting the AIDS phenotype, as that would immediately cause AIDS in the recipient host. Improving the evolutionary dynamics model and expanding it to different stages of the infection could potentially lead to a tool for exploring the potential of next generation sequencing for between-host directional transmission discrimination and other techniques exploiting the within-host viral diversity.
The fact that within-host evolution affects the quasispecies over time lead to the conception of different sampling schemes to test the effects of time on the reconstruction of the transmission tree. The generation of the viral variant infecting a new host might predate the infection time by some time. Comparing different sampling schemes could shed a light on the sampling time dependency of the reconstruction performance. This reasoning is trivial if one focuses on a specific transmission event, less so if the sampling strategy must comprise the transmission history of a whole epidemics. Figure 14 shows the triplets distance from the known simulated transmission tree as a function of the sampling scheme. There is no significant difference between the phylogenies reconstructed (i.e. excluding the reference sampled trees) according to the sampling schemes *Af30*, *Af60*, *Be30*, *Be60*,

In and *Ra* at the same mutation rate. The reason behind it might be intrinsic in the conception of the sampling scheme: by choosing the sampling relative to the time of first infection, all subsequent infections lack any temporal information given by the sampling. As depicted in figures 2, 3 and 4, the difference between these sampling schemes is minimal. Frequently, hosts are sampled at the end of the simulation, and the difference in time of those who are not is little. The presence or absence of the actually transmitted strain plays little role, as the sequences used for the reconstruction are the consensus sequences of the 100 most frequent strains. It comes thus as no surprise that the reconstruction of the aforementioned sampling schemes result in similar performances. The simulations results suggest the inadequacy of homochronous sampling. *Ti* seems to present the worst performance in both mutation rate regimens is true. More data is needed to corroborate this suggestion. *Tr* represents the phylogeny reconstructed from the single transmitted strain of each host first infection, i.e. no consensus sequence is built. It contains two sources of error. Firstly the error due to the fact that only the first infection of each host is taken into account, and no further temporal information for subsequent infections is given. This characteristic is shared by all the heterochronous sampling schemes devised so far. Secondly, and more interesting, *Tr* can suffer from the loss of generality of the consensus sequence. If the transmitted sequence is a minor variant, or if the quasispecies migrates into different regions of the fitness landscape, the sequence transmitted in the first infection might not give much information on subsequent infections by the same host. Building a consensus sequence out of the most frequent viral variants might improve the reconstruction performance, as it is more robust towards quasispecies migration on the fitness landscape, and is inherently resistant to minor variants.

In order to discuss the effect of mutation rate on the reconstruction of the transmission tree, I performed two simulation runs with different mutation rates and analyzed different sampling schemes for each simulation. The results are summarized in figure 15. The reconstruction performance improves consistently from the lower mutation rate simulation runs to the higher mutation rate simulation runs for every sampling scheme (suffix *_L* and *_H* in figure 15, respectively). Even though the single change of the corresponding sampling scheme might not be significant in the unstandardized data (figure 15a), the overall trend towards improving the reconstruction performance with a higher mutation rate suggests that there is a real effect involved, albeit not significant at the 95% confidence level. Standardizing the data (centering and dividing by the standard deviation of the pooled sampling scheme specific data of both high and low mutation rate simulation runs) reveals the pattern that was hidden before, removing the effects due to the different sampling schemes. In the standardized data set, the total means of the pooled data from the low and high mutation rate simulation runs, respectively, are significantly different, with the latter showing a better phylogenetical reconstruction performance than the former. As expected, the sampling-scheme specific references (*_Ref* suffix in figure 14), which are dependent only from the sampling scheme and the known transmission tree topology, but not on the reconstruction performance, are consistent between the two simulation setups. This is evidence that the detected change is not due to artifacts of sampling, which could derive by applying a certain sampling scheme to a particular topology. If that would be the case, both sampling-scheme specific references and reconstructed phylogenies would be affected. Since the sampling-scheme specific references are at comparable distances between the two simulation runs, it is possible to rule out the presence of important sampling artifacts.

Even though this result consists in a two points estimate (lower and higher mutation rates), it seems to suggest that the faster mutation rate provides a better reconstruction performance than the lower. Two different, possibly competing effects involve the muta-

tion rate. One is the loss of signal over time, by means of multiple mutation in the same loci, which possibly hides the evolutionary history of a sequence. The second effect is the generation of evolutionary signal in the sequences over time. In order to get a better phylogenetical reconstruction performance, the mutation rate must thus be high enough to generate a sufficient amount of characteristic mutations to relate the sequences, but not as high as to mutate the same loci multiple times in the timescale of sampling. The intuition behind this mechanisms may be explained by a small thought experiment: imagine having a fictitious pathogen evolving at a nearly infinite pace, in such a way that every information contained in the original infecting pathogen is lost before another transmission event could occur. Any attempt at reconstructing the transmission tree from the phylogeny of such a pathogen would fail, as the sequences of the pathogens sampled from different infected hosts would not contain enough information to relate them. On the contrary, if we would have another fictitious pathogen, this time not evolving at all, any attempt at reconstructing the transmission tree from the phylogeny would still fail, as there would be no information to tell the different hosts apart. They would be infected by the same pathogen variant. The conclusion of this thought experiment is that there must exist an intermediate mutation rate for which the reconstruction performance is maximal. Further works could concentrate on exploring the effects of a wider range of mutation rates on the reconstruction performance, to test the limits of these two mechanisms.

I created a software for the simulation of within and between-host HIV dynamics and evolution. This links mechanistically the evolution of a viral sequence to its within-host dynamics through the modelling of a fitness landscape. The goal was to develop a tool for studying the effects of within-host evolution on the reconstruction of transmission trees from sequences sampled from hosts in an HIV epidemics. This software could easily be expanded to answer other research questions linked to the within-host evolution and viral diversity, as the testing of methods to ascertain transmission directionality [82][83]. Minimal changes could be implemented to account for more diverse between-host transmission models, including contact networks and structured populations.

I discussed two effects on the reconstruction performance: the sampling scheme adopted and the role of the mutation rate. While in the future different sampling schemes can be engineered, the data collected so far in this work seem to suggest a better performance in transmission tree reconstruction for heterochronously sampled epidemics with respect to the homochronously sampled epidemics. Would this be confirmed, it would be comforting news, since new infected hosts are naturally discovered - and sampled - at different times. Additionally, different reconstruction performances by different sampling schemes represent a confounding factor of transmission tree reconstruction, as no information about the infection time is usually known prior to sampling hosts in a real epidemics. By this logic, the fact that there seems to be no difference between the heterochronous sampling schemes reduces the amount of potential confounding factors in the transmission tree reconstruction. Moreover, the construction of consensus sequences might help improving the reconstruction performance of epidemics presenting infected hosts infecting multiple people on a long time frame. The construction of a consensus sequence comprising the most frequent viral variants is robust towards quasispecies fitness landscape migration and the error of possible transmission of a minor viral variant on subsequent infections. On the short sequence analyzed (1000 bp), a higher mutation rate provides a higher evolutionary signal, which improves the reconstruction performance. No transmission tree was perfectly reconstructed, and indicates that the performance of transmission tree reconstruction might be poor in absence of much evolutionary signal.

Future efforts might be directed towards improving the computational efficiency, increasing the simulated volume and sequence length to values closer to the observed system.

The evolutionary dynamics model could benefit from a overhaul and improvement of the fitness model, taking into account both synonymous/non-synonymous mutation, determining genomic regions under neutral evolution, avoiding the possibility of escape to already defeated viral phenotypes, and the implementation of antiretroviral treatment as a selection force shaping the fitness landscape. The addition of a latent cell compartment could help understanding the role of ancient viral variants in the between-host transmission, and add a layer of realism to the within-host dynamics. An AIDS disease stage could be included through different mechanisms. Better performing sampling schemes must be devised to get a more precise mechanistic insight on the role of the sampling time on the reconstruction of the transmission trees. In the future, this program could be of help testing hypothesis that need a within-host evolution framework, as the transmission directionality discrimination, thus helping progress in the research fields connected to the epidemiology of fast evolving pathogens.

References

[1] L Montagnier et al. "A new human T-lymphotropic retrovirus: characterization and possible role in lymphadenopathy and acquired immune deficiency syndromes". In: *Human T-cell leukemia/lymphoma virus* 363 (1984).

[2] Joint United Nations Programme on HIV/AIDS et al. "Global AIDS update 2016". In: *Geneva: UNAIDS* (2016).

[3] Gilles Wandeler, Leigh F Johnson, and Matthias Egger. "Trends in life expectancy of HIV-positive adults on antiretroviral therapy across the globe: comparisons with general population". In: *Current Opinion in HIV and AIDS* 11.5 (2016), pp. 492–500.

[4] Myron S Cohen et al. "Antiretroviral therapy for the prevention of HIV-1 transmission". In: *New England Journal of Medicine* 375.9 (2016), pp. 830–839.

[5] Anthony S Fauci. "25 years of HIV". In: *Nature* 453.7193 (2008), pp. 289–290.

[6] Maithe Enriquez and David S McKinsey. "Strategies to improve HIV treatment adherence in developed countries: clinical management at the individual level." In: *HIV/AIDS (Auckland, N.Z.)* 3 (2011), pp. 45–51. ISSN: 1179-1373. DOI: 10.2147/HIV.S8993.

[7] Sara Fernandez-Lopez et al. "Transmission intensity and impact of control policies on the foot and mouth epidemic in Great Britain". In: *Biol* 2 (2000), E153–E157.

[8] Martin A Nowak, Robert M May, and Roy M Anderson. "The evolutionary dynamics of HIV-1 quasispecies and the development of immunodeficiency disease". In: *Aids* 4.11 (1990), pp. 1095–1103.

[9] Martin A Nowak and Robert M May. "Mathematical biology of HIV infections: antigenic variation and diversity threshold". In: *Mathematical Biosciences* 106.1 (1991), pp. 1–21.

[10] Martin Nowak and Robert M May. "AIDS pathogenesis: mathematical models of HIV and SIV infections." In: *Aids* 7 (1993), S3–S18.

[11] Martin A Nowak. "AIDS pathogenesis: from models to viral dynamics in patients." In: *Journal of acquired immune deficiency syndromes and human retrovirology: official publication of the International Retrovirology Association* 10 (1994), S1–5.

[12] Alison L Hill et al. “Predicting the outcomes of treatment to eradicate the latent reservoir for HIV-1”. In: *Proceedings of the National Academy of Sciences* 111.37 (2014), pp. 13475–13480.

[13] Sarah B Joseph et al. “Bottlenecks in HIV-1 transmission: insights from the study of founder viruses”. In: *Nature reviews. Microbiology* 13.7 (2015), p. 414.

[14] Vikki Gill et al. “Macrophages are the major target cell for HIV infection in long-term bone marrow culture and demonstrate dual susceptibility to lymphocytotropic and monocytotropic strains of HIV-1”. In: *British journal of haematology* 93.1 (1996), pp. 30–37.

[15] MJ McElrath et al. “Comprehensive assessment of HIV target cells in the distal human gut suggests increasing HIV susceptibility toward the anus”. In: *Journal of acquired immune deficiency syndromes (1999)* 63.3 (2013), p. 263.

[16] Christopher J Miller and Robin J Shattock. “Target cells in vaginal HIV transmission”. In: *Microbes and infection* 5.1 (2003), pp. 59–67.

[17] Valentina Svicher et al. “Understanding HIV compartments and reservoirs”. In: *Current HIV/AIDS Reports* 11.2 (2014), pp. 186–194.

[18] Stephanie M Dorosko and Ruth I Connor. “Primary human mammary epithelial cells endocytose HIV-1 and facilitate viral infection of CD4+ T lymphocytes”. In: *Journal of virology* 84.20 (2010), pp. 10533–10542.

[19] Ronald Swanstrom and John Coffin. “HIV-1 pathogenesis: the virus”. In: *Cold Spring Harbor perspectives in medicine* 2.12 (2012), a007443.

[20] Tae-Wook Chun et al. “Early establishment of a pool of latently infected, resting CD4+ T cells during primary HIV-1 infection”. In: *Proceedings of the National Academy of Sciences* 95.15 (1998), pp. 8869–8873.

[21] Renée M van der Sluis, Rienk E Jeeninga, and Ben Berkhout. “Establishment and molecular mechanisms of HIV-1 latency in T cells”. In: *Current opinion in virology* 3.6 (2013), pp. 700–706.

[22] Oliver F Brandenberg et al. “The HIV-1 entry process: a stoichiometric view”. In: *Trends in microbiology* 23.12 (2015), pp. 763–774.

[23] John D Roberts, Katarzyna Bebenek, and Thomas A Kunkel. “The accuracy of reverse transcriptase from HIV-1”. In: *Science* 242.4882 (1988), p. 1171.

[24] Kelly Jo Huang and Dawn P Wooley. “A new cell-based assay for measuring the forward mutation rate of HIV-1”. In: *Journal of virological methods* 124.1 (2005), pp. 95–104.

[25] Christof K Biebricher and Manfred Eigen. “What is a quasispecies?” In: *Quasispecies: Concept and Implications for Virology*. Springer, 2006, pp. 1–31.

[26] Gary Maartens, Connie Celum, and Sharon R Lewin. “HIV infection: epidemiology, pathogenesis, treatment, and prevention”. In: *The Lancet* 384.9939 (2014), pp. 258–271.

[27] Martin A Nowak and Andrew J McMichael. “How HIV defeats the immune system”. In: *Scientific American* 273.2 (1995), pp. 58–65.

[28] Haynes W Sheppard et al. “The Initial Immune Response to HIV and Immune System”. In: *Journal of acquired immune deficiency syndromes* 4.7 (1991).

[29] Michael W Adler and Anne M Johnson. “Contact tracing for HIV infection”. In: *British medical journal (Clinical research ed.)* 296.6634 (1988), p. 1420.

[30] George W Rutherford and Jean M Woo. “Contact tracing and the control of human immunodeficiency virus infection”. In: *JAMA* 259.24 (1988), pp. 3609–3610.

[31] Bram Vrancken et al. “The genealogical population dynamics of HIV-1 in a large transmission chain: bridging within and among host evolutionary rates”. In: *PLoS computational biology* 10.4 (2014), e1003505.

[32] Matthew Hall, Mark Woolhouse, and Andrew Rambaut. “Epidemic reconstruction in a phylogenetics framework: transmission trees as partitions of the node set”. In: *PLoS computational biology* 11.12 (2015), e1004613.

[33] M Aldrin et al. “Modelling the spread of infectious salmon anaemia among salmon farms based on seaway distances between farms and genetic relationships between infectious salmon anaemia virus isolates”. In: *Journal of The Royal Society Interface* 8.62 (2011), pp. 1346–1356.

[34] Rolf JF Ypma et al. “Unravelling transmission trees of infectious diseases by combining genetic and epidemiological data”. In: *Proceedings of the Royal Society of London B: Biological Sciences* 279.1728 (2012), pp. 444–450.

[35] Marco J Morelli et al. “A Bayesian inference framework to reconstruct transmission trees using epidemiological and genetic data”. In: *PLoS computational biology* 8.11 (2012), e1002768.

[36] F Sanger, S Nicklen, and A R Coulson. “DNA sequencing with chain-terminating inhibitors.” In: *Proceedings of the National Academy of Sciences of the United States of America* 74 (12 Dec. 1977), pp. 5463–5467. ISSN: 0027-8424.

[37] Anne Derache et al. “HIV drug resistance mutations in proviral DNA from a community treatment program.” In: *PloS one* 10 (1 2015), e0117430. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0117430.

[38] Mary K Grabowski and Andrew D Redd. “Molecular tools for studying HIV transmission in sexual networks.” In: *Current opinion in HIV and AIDS* 9 (2 Mar. 2014), pp. 126–133. ISSN: 1746-6318. DOI: 10.1097/COH.0000000000000040.

[39] Matthew R Henn et al. “Whole genome deep sequencing of HIV-1 reveals the impact of early minor variants upon immune recognition during acute infection.” In: *PLoS pathogens* 8 (3 2012), e1002529. ISSN: 1553-7374. DOI: 10.1371/journal.ppat.1002529.

[40] Hirotaka Ode et al. “Quasispecies Analyses of the HIV-1 Near-full-length Genome With Illumina MiSeq.” In: *Frontiers in microbiology* 6 (2015), p. 1258. ISSN: 1664-302X. DOI: 10.3389/fmicb.2015.01258.

[41] T Leitner et al. “Accurate reconstruction of a known HIV-1 transmission history by phylogenetic tree analysis.” In: *Proceedings of the National Academy of Sciences of the United States of America* 93 (20 Oct. 1996), pp. 10864–10869. ISSN: 0027-8424.

[42] Michael L Metzker et al. “Molecular evidence of HIV-1 transmission in a criminal case”. In: *Proceedings of the National Academy of Sciences* 99.22 (2002), pp. 14292–14297.

[43] Edward C Holmes et al. “Molecular investigation of human immunodeficiency virus (HIV) infection in a patient of an HIV-infected surgeon”. In: *Journal of Infectious Diseases* 167.6 (1993), pp. 1411–1414.

[44] Jan Albert et al. “Analysis of a rape case by direct sequencing of the human immunodeficiency virus type 1 pol and gag genes.” In: *Journal of virology* 68.9 (1994), pp. 5918–5924.

[45] Eleanor M Cottam et al. “Integrating genetic and epidemiological data to determine transmission pathways of foot-and-mouth disease virus”. In: *Proceedings of the Royal Society of London B: Biological Sciences* 275.1637 (2008), pp. 887–895.

[46] Federica Giardina et al. “Inference of transmission network structure from HIV phylogenetic trees”. In: *PLoS computational biology* 13.1 (2017), e1005316.

[47] Tom FW Wolfs et al. “HIV-1 genomic RNA diversification following sexual and parenteral virus transmission”. In: *Virology* 189.1 (1992), pp. 103–110.

[48] Rolf JF Ypma, W Marijn van Ballegooijen, and Jacco Wallinga. “Relating phylogenetic trees to transmission trees of infectious disease outbreaks”. In: *Genetics* 195.3 (2013), pp. 1055–1062.

[49] Ethan Romero-Severson et al. “Timing and order of transmission events is not directly reflected in a pathogen phylogeny”. In: *Molecular biology and evolution* 31.9 (2014), pp. 2472–2482.

[50] Rebecca V Culshaw and Shigui Ruan. “A delay-differential equation model of HIV infection of CD4+ T-cells”. In: *Mathematical biosciences* 165.1 (2000), pp. 27–39.

[51] Alan S Perelson, Denise E Kirschner, and Rob De Boer. “Dynamics of HIV infection of CD4+ T cells”. In: *Mathematical biosciences* 114.1 (1993), pp. 81–125.

[52] John M Coffin et al. “HIV population dynamics in vivo: implications for genetic variation, pathogenesis, and therapy”. In: *Science-AAAS-Weekly Paper Edition* 267.5197 (1995), pp. 483–489.

[53] Sebastian Bonhoeffer and Martin A Nowak. “Mutation and the evolution of virulence”. In: *Proceedings of the Royal Society of London B: Biological Sciences* 258.1352 (1994), pp. 133–140.

[54] Samuel Alizon and Carsten Magnus. “Modelling the course of an HIV infection: insights from ecology and evolution”. In: *Viruses* 4.10 (2012), pp. 1984–2013.

[55] Sunpeng Wang and Libin Rong. “Stochastic population switch may explain the latent reservoir stability and intermittent viral blips in HIV patients on suppressive therapy”. In: *Journal of theoretical biology* 360 (2014), pp. 137–148.

[56] Martin A Nowak et al. “Antigenic diversity thresholds and the development of AIDS”. In: *Science* 254.5034 (1991), pp. 963–969.

[57] Xia Wang et al. “Mathematical analysis of an HIV latent infection model including both virus-to-cell infection and cell-to-cell transmission”. In: *Journal of biological dynamics* 11.sup2 (2017), pp. 455–483.

[58] Jinliang Wang, Jie Yang, and Toshikazu Kuniya. “Dynamics of a PDE viral infection model incorporating cell-to-cell transmission”. In: *Journal of Mathematical Analysis and Applications* 444.2 (2016), pp. 1542–1564.

[59] Jinhu Xu, Yan Geng, and Yicang Zhou. “Global dynamics for an age-structured HIV virus infection model with cellular infection and antiretroviral therapy”. In: *Applied Mathematics and Computation* 305 (2017), pp. 62–83.

[60] Christophe Fraser. “HIV recombination: what is the impact on antiretroviral therapy?” In: *Journal of The Royal Society Interface* 2.5 (2005), pp. 489–503.

[61] Sebastian Bonhoeffer, John M Coffin, and Martin A Nowak. “Human immunodeficiency virus drug therapy and virus load.” In: *Journal of Virology* 71.4 (1997), pp. 3275–3278.

[62] Sompop Moonchai, Yongwimon Lenbury, and Wannapong Triampo. “Cellular automata simulation modeling of HIV infection in lymph node and peripheral blood compartments”. In: *International Journal of Mathematics and Computers in Simulation* 4 (2010), p. 124.

[63] Anass Bouchnita et al. “Towards a Multiscale Model of Acute HIV Infection”. In: *Computation* 5.1 (2017), p. 6.

[64] John Frank Charles Kingman. “The coalescent”. In: *Stochastic processes and their applications* 13.3 (1982), pp. 235–248.

[65] Filip Bielejec et al. “π BUSS: a parallel BEAST/BEAGLE utility for sequence simulation under complex evolutionary scenarios”. In: *BMC bioinformatics* 15.1 (2014), p. 133.

[66] Liming Liang, Sebastian Zöllner, and Gonçalo R Abecasis. “GENOME: a rapid coalescent-based whole genome simulator”. In: *Bioinformatics* 23.12 (2007), pp. 1565–1567.

[67] David Posada. “Evaluation of methods for detecting recombination from DNA sequences: empirical data”. In: *Molecular biology and evolution* 19.5 (2002), pp. 708–717.

[68] Daniel T Gillespie. “A general method for numerically simulating the stochastic time evolution of coupled chemical reactions”. In: *Journal of computational physics* 22.4 (1976), pp. 403–434.

[69] Daniel T Gillespie. “Exact stochastic simulation of coupled chemical reactions”. In: *The journal of physical chemistry* 81.25 (1977), pp. 2340–2361.

[70] EJJH Domingo and JJ Holland. “RNA virus mutations and fitness for survival”. In: *Annual Reviews in Microbiology* 51.1 (1997), pp. 151–178.

[71] Fabio Zanini et al. “In vivo mutation rates and the landscape of fitness costs of HIV-1”. In: *Virus Evolution* 3.1 (2017).

[72] Michael D. Vose. “A linear algorithm for generating random numbers with a given distribution”. In: *IEEE Transactions on software engineering* 17.9 (1991), pp. 972–975.

[73] R Core Team. *R: A language and environment for statistical computing [Internet]. Vienna, Austria; 2014*. 2017.

[74] Remco Bouckaert et al. “BEAST 2: a software platform for Bayesian evolutionary analysis”. In: *PLoS computational biology* 10.4 (2014), e1003537.

[75] Douglas E Critchlow, Dennis K Pearl, and Chunlin Qian. “The triples distance for rooted bifurcating phylogenetic trees”. In: *Systematic Biology* 45.3 (1996), pp. 323–334.

[76] Dominic J Bennett, Mark D Sutton, and Samuel T Turvey. “treeman: an R package for efficient and intuitive manipulation of phylogenetic trees”. In: *BMC research notes* 10.1 (2017), p. 30.

[77] Michelle Kendall and Caroline Colijn. “A tree metric using structure and length to capture distinct phylogenetic signals”. In: *arXiv preprint arXiv:1507.05211* (2015).

[78] David F Robinson and Leslie R Foulds. “Comparison of phylogenetic trees”. In: *Mathematical biosciences* 53.1-2 (1981), pp. 131–147.

[79] Andrew D Redd et al. “Previously transmitted HIV-1 strains are preferentially selected during subsequent sexual transmissions.” In: *The Journal of infectious diseases* 206 (9 Nov. 2012), pp. 1433–1442. ISSN: 1537-6613. DOI: 10.1093/infdis/jis503.

[80] Katrina A Lythgoe and Christophe Fraser. “New insights into the evolutionary rate of HIV-1 at the within-host and epidemiological levels.” In: *Proceedings. Biological sciences* 279 (1741 Aug. 2012), pp. 3367–3375. ISSN: 1471-2954. DOI: 10.1098/rspb.2012.0595.

[81] Irina Maljkovic Berry et al. “Unequal evolutionary rates in the human immunodeficiency virus type 1 (HIV-1) pandemic: the evolutionary rate of HIV-1 slows down when the epidemic rate increases.” In: *Journal of virology* 81 (19 Oct. 2007), pp. 10625–10635. ISSN: 0022-538X. DOI: 10.1128/JVI.00985-07.

[82] Diane I Scaduto et al. “Source identification in two criminal cases using phylogenetic analysis of HIV-1 DNA sequences.” In: *Proceedings of the National Academy of Sciences of the United States of America* 107 (50 Dec. 2010), pp. 21242–21247. ISSN: 1091-6490. DOI: 10.1073/pnas.1015673107.

[83] Jing Yang, Meng Ge, and Xian-Ming Pan. “A time lag insensitive approach for estimating HIV-1 transmission direction.” In: *AIDS (London, England)* 26 (8 May 2012), pp. 921–928. ISSN: 1473-5571. DOI: 10.1097/QAD.0b013e3283536b89.