Novel simulation of HIV evolution within and between hosts, and developing software tools, for inferring ancestry and transmission

---

Author:

Marianne Aspbury

Supervisor:

Dr. Yan Wong (BDI)

Big Data Institute

The University of Oxford

October 4, 2019

Word Count: 4395

Abstract

The aim of this project was to design a novel way of simulating and inferring evolution of the Human Immunodeficiency Virus (HIV) within and between-hosts, by exploiting newly developed tree-sequence methods. The efficiency and scalability of tree-sequences allow us to model the complex evolutionary history of HIV, and, in the future, to manipulate large amounts of HIV data. The focus of this project was on developing simulations of HIV transmission networks and its evolutionary processes including mutation and recombination, and use these to make and test inferences of evolutionary history from limited, and fragmented, samples. Simulations allow us to test the results of our inference methods against known properties of the simulated data, and to refine and add successive layers of complexity into the model, whilst testing at each step. The work so far shows promise that we will be able to build a model of HIV that can be applied to real data, once some additional software development has been completed, and further validation performed with simulated data. In this project, I have contributed to open-source tree-sequence software tools (*tskit*, *msprime* and *tsinfer*) in collaboration with other researchers (Yan Wong, Wilder Wohns & Jerome Kelleher) to handle necessary qualities of HIV evolution.

1 Introduction

The aim of this project is to design a novel way of modelling evolution of the Human Immunodeficiency Virus (HIV) within and between-hosts, exploiting newly developed tree-sequence methods [1]. The scalability and efficient data-storage capabilities allow us to cope with big data [2], and thus model the complex evolutionary history of HIV, brought about by factors including high recombination rates. The ultimate aim of this project was to simulate and infer realistic evolutionary history of HIV sequences, representing genetic relationships both within and between different HIV hosts, using the novel tree sequence representation. However, the initial aims were to build a reasonable small-scale, simplified, model which can be improved upon with future research. Before being able to simulate and infer near-complete HIV ancestry, we aimed to address the problem of inferring the direction of HIV transmission between hosts, given that HIV sequences sampled from hosts are fragmented; only spanning a limited section of the genome. Using simulated data allows us to test our methods and to refine and build-in successive layers of complexity into the model, including successive testing, in order to build a robust model that can be applied to real data once the model has been adequately validated on simulated data.

HIV has had a huge global impact on health, and this continues today particularly in Eastern Europe, Central Asia, the Middle East and Africa, due to challenges including legal, social, and economic restrictions to treatment and infection prevention [3]. The virus itself displays rapid evolution with high mutation and recombination rates, differing selection within hosts and during transmission, and changing growth rates [4–6]. These create research challenges for modelling the complexity of its behaviour and learning from these models in order to provide better treatment and infection control. Current models often fail to address important aspects of the problem at hand; often they discount the presence of recombination despite its high

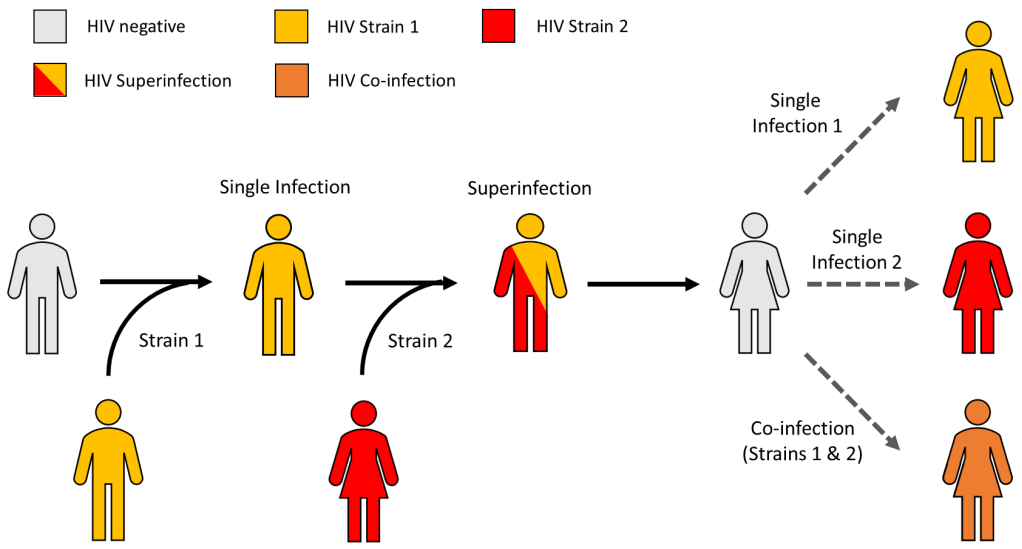


Figure 1: Example of the two types of dual infection: superinfection (two different infectors at different timepoints) and co-infection (two genetically distinct infections from one infector i.e. one infection event). Dual infection strains may be the same (intraclade) or different (interclade) sub-types of HIV. Figure adapted from [7].

rate for HIV, or consider only single infection events rather than dual infections. There are two kinds of dual infection, either distinct viral infections from one host, or separate single infections from two different hosts (see fig. 1). Dual infections are estimated to account for a significant proportion of HIV cases (possibly $10-20\%$, with as much as 50% incidence estimated in some cases [Christophe Fraser, personal communication]). Another limitation of current HIV models is that they generally do not consider the problem of missing data in HIV sampling. We aim to make incorporating such complexities feasible and scalable for HIV analysis by using novel simulation and inference methods.

1.1 Detailing The Human Immunodeficiency Virus (HIV)

At the time of its discovery, HIV-1 exhibited an unprecedented rate of genetic evolution, orders of magnitude higher than other known viruses at the time, and which continues to be one of the fastest known biological rates of evolution [4]. Genetic evolution is defined by the genetic change in the virus over generations of replication; predominantly described by mutation, selection and genetic drift, although many factors serve to complicate these processes [4, 8]. Accompanying the high rate of evolution, there are also record levels of genetic diversity recorded amongst HIV virions in hosts (surpassed only recently by hepatitis C [4]). Initially, this high genetic diversity was incorrectly thought to be due to an exceptionally high error rate during replication. However, the error rate for HIV replication is similar to other RNA viruses, and the underlying reasons for its genetic diversity are the very short replication cycles ($\sim 1-2$ days), long infection periods (currently, a human lifetime), and a large replicating population of virions, which together enable rapid response to selection pressures [4].

The average mutation rate of HIV is still debated, since it demonstrably varies between

in-vitro and in-vivo experiments, inside and outside of the human body, and due to the under-estimation of deleterious and lethal mutations in previous studies [9]. For models which do not attempt to include all the complexities of HIV, a mutation rate of $\sim 2 - 3 \times 10^{-5}$ mutations per genomic site per generation is often quoted [4, 5, 10], although the true HIV mutation rate may be on the order of 10^{-3} mutations per site per generation [9]. Therefore, for the purposes of our simplified simulations, we also assume a conservative mutation rate of 2.5×10^{-5} per site per generation.

Recombination rates for HIV are also an actively debated topic. In-vivo studies of HIV recombination are limited, but the few estimates published do not concur with in-vitro studies, and complexities in recombination, including template switching effects, complicate the task of estimating HIV recombination rates [5, 11]. We use a recombination rate of 1×10^{-4} per site per generation in our simulations, following overall estimates of recombination by various authors, and the typical recombination rates utilised in other HIV evolution models [5].

Other key aspects of HIV to consider include replication rates, population sizes of HIV particles in infected hosts, quantification of HIV transmission between individuals and viral population growth, as well as qualities of HIV sequences obtained from infected individuals. As mentioned previously, the replication cycle for HIV is very short, about 1-2 days [4], which enables the virus to evade the immune system through rapid evolution. For simplicity we assume a replication cycle, i.e. generation time, of one day. In our model, the population size we must consider is not the absolute number of virions in a host, which is $\sim 10^{11}$ in steady state [4], nor the number of infected cells in an individual, $\sim 10^{7-9}$ [4, 12], but the size required to model genetic properties of the population: the "effective population size" [12]. For HIV, coalescent estimations of the effective population size are $\sim 10^3$, although longitudinal estimates in patients vary from $\sim 10^{2-4}$ [12]. Hence, for our simulations, we either use populations of 1×10^2 or 1×10^3. These reflect other models in the literature and are appropriate sizes for simpler recombination models: the more complex processes introduced into the model, the higher the effective population size also required [Christophe Fraser, personal communication].

Another important consideration is the quantifiable transmission behaviour of HIV, both at the level of individual transmission events, and the possibility of dual infections (fig. 1). Despite their real-world importance, dual infection effects are currently under-researched in HIV models. One benefit of tree sequence methods is that it should be straightforward to implement dual infection scenarios, in order to study the effect of dual infections on HIV evolutionary history and inference capabilities (e.g. of transmission direction). However, in this preliminary work we only consider the common case of single infection events, which simplifies the development of the initial model. A key aim for the future is to extend this initial model to include the extra complexity of dual infections. In order to implement realistic transmission structure into our simulations, we use single-infection transmission networks supplied by collaborator William Probert (BDI, Oxford) [13, 14]. In single infection events, the virus is subject to extreme selection pressure in the transmission channel and only one virion is transmitted from the infecting to the infected person to seed the new infection [e.g. 4]. From this single virion

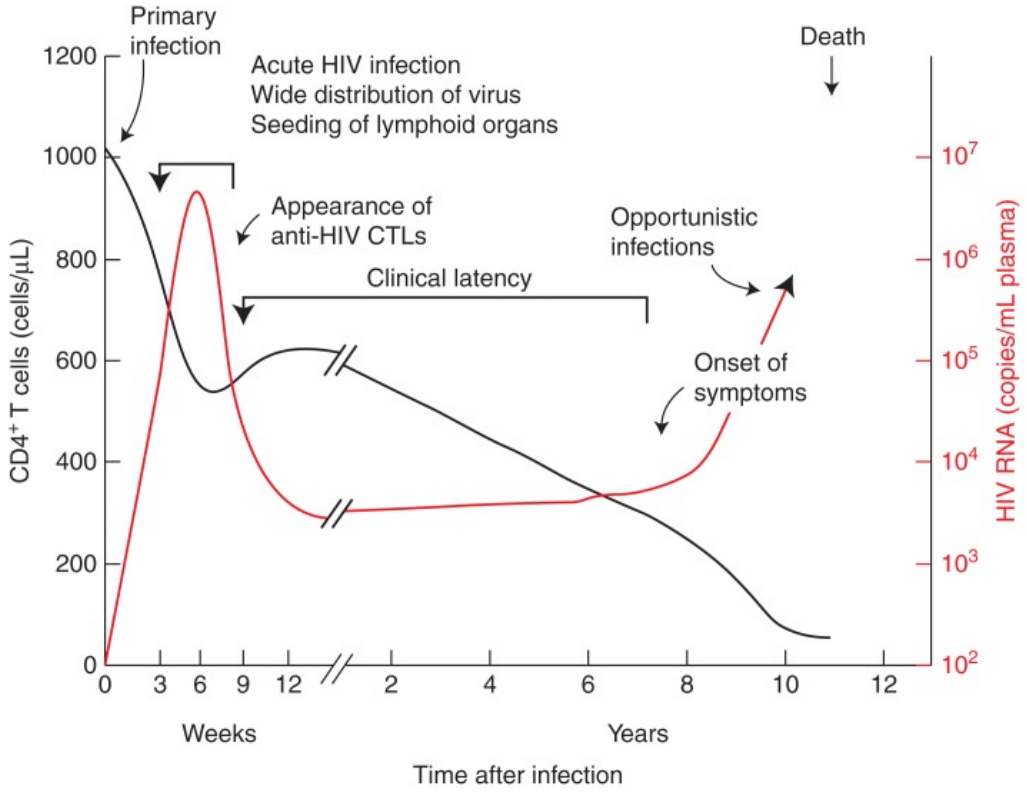


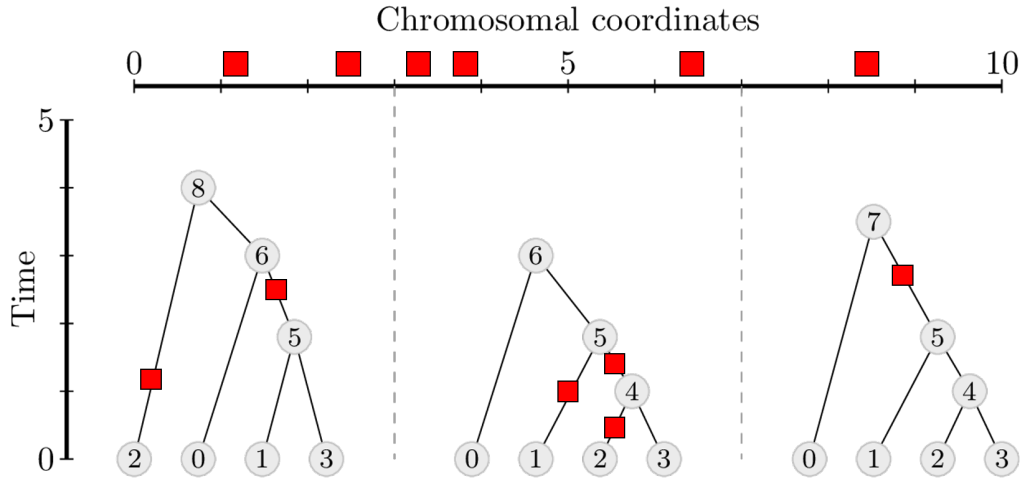
Figure 2: Time course of a typical HIV infection; patients' CD4+ T cells are shown in black, and on the left axis, whilst HIV RNA copies are shown in red, and on the right axis. Figure from [4].

there is a huge viral growth rate over several weeks, and a later decline following the host's immune response, before steady-state is achieved (see fig. 2).

Finally, we must consider the details of the HIV sequencing process when sampling HIV genomes from individuals, and the quality of data that is obtained. Modern high throughput sequencing methods, which reconstruct a single overall genome from many overlapping short reads, are inadequate for sequencing HIV. Each short-read sequence from a given patient is likely to come from a different HIV virion, and since there is such high genetic diversity amongst HIV virions in each infected individual, these short sequence samples cannot be aligned and overlapped with each other to reconstruct a whole genome. Hence, hundreds of thousands of short reads are taken per patient to maximise coverage and enable some phylogenetic analysis of the HIV under these difficult circumstances. Of the 9181 base pairs (bp) comprising the HIV genome [15], the current average sample read length is $200 - 300$ bp, with current maximum of ~ 500 bp (although with improving technology and experimental method, longer reads are likely to become available soon), whilst reads below 50 bp tend to be discarded [Matthew Hall and Christophe Fraser, personal communication]. These short sequences can be considered a missing data problem, in that we have a section of genome, e.g. 200 bp long, with known haplotypes, but for haplotypes to the left and right of these genomic positions we have no recorded information. Incorporating the reality of this missing data into any inferences we try to make from samples in our simulated data is important for building an inference framework which will actually hold up to real world data, and prove epidemiologically useful.

1.2 Tree Sequence Data Representation

Central to this project is an understanding of trees and tree sequences. A tree represents the genealogical history of a given set of genomes at some range of genomic co-ordinates (i.e. position along the genome or chromosomal co-ordinates, or an allele). A tree comprises nodes, representing a common ancestor between individual sequences at a certain point in time, and edges, which represent genealogical relationships between the sequences. However, one tree is not enough to represent the genealogical history of the entire DNA sequence of one sample (e.g. of one virion, or one person), due to the process of recombination. Recombination is the mixing of sections of DNA sequence during reproduction, and means that different sections of the genome can be inherited via different ancestral routes. For example, in humans a particular haplotype inherited from a person's father may consist of a mixture of sections of DNA from their father's father and sections from their father's mother. Thus, it is not possible to use one tree to represent the complete ancestry of an entire genome, since some sections of the genome will have different ancestors to other sections. The underlying inheritance process means that there exists a shared ancestry for our viral genomes, although there may not be enough information available to us in the fragmented sequence data to fully trace this ancestry. Graphically, we can represent such a tree sequence as in fig. 3.



	SNP 0	SNP 1	SNP 2	SNP 3	SNP 4	SNP 5
Sample 0	0	0	0	0	0	0
Sample 1	1	0	1	0	0	1
Sample 2	0	1	0	1	1	1
Sample 3	1	0	0	0	1	1

Figure 3: An example tree sequence comprising three trees, figure adapted from Georgia Tsambos (original available at https://github.com/gtsambos/SMBE-SGE-2019).

2 Simulation Method

In order to infer features of HIV evolution and test these inferences, it helps to be able to simulate this within a realistic HIV infection network. This enables us to determine whether our inferences are accurate and reliable, since with simulated data we know the history of the samples. We can also make any adjustments to the complexity of the system, and which effects we wish to can consider, in a controlled environment, so that we can interpret and improve our model in successive steps. Once we have evaluated the accuracy of our inferences based on simulated data, it is then possible to carry forward the inference scheme to real data and be able to assess the reliability and accuracy of these inferences.

The core tools of our simulations are a set of python software packages developed under the tree sequence toolkit (*'tskit'*) framework. This includes the underlying *tskit* library, as well as the specific tree sequence simulator software *msprime*, and the inference software *tsinfer*. More information on these tools can be found in publications detailing their performance [1, 2, 16] and all code for them is available at `https://github.com/tskit-dev`. Whilst these tools were initially developed for evaluating human ancestry, the underlying coalescent theory [1, 17] is applicable to general population genetics including pathogen genetics. With further development of the software packages' capabilities in this project, we are able to incorporate complexities required for studying HIV, including the missing-data problem of fragmented short sequences rather than whole-genome reads.

2.1 Transmission Networks

The first important step in our simulations is to base our transmission events off of realistic HIV transmission networks. For this purpose we collaborated with an HIV researcher based at the BDI; Dr William Probert. He was able to supply us with extensive realistically simulated transmission networks comprising single infection events totalling 49011 individuals, and two root infectors [13, 14]. Although the tree sequence simulation and inference tools are capable of handling large numbers of individuals, in this exploratory work it is more convenient to select smaller sub-networks for testing purposes. We create sub-networks by select a random sub-network from the possible sub-networks satisfying two conditions: it contains the number of overall infected people we desire to include, and the sub-network has only one root infector. A toy example selecting some sub-networks from a small overall network is shown in fig. 4. We filter the network to a set of available sub-networks, then use the python package *random* to select our sub-network from these options.

2.2 Simulation Software: *msprime*

Having chosen a subset of the transmission network with our desired number of individuals, we need a simulator for the genetic evolution of HIV within and between hosts in this network. For this we use the tree sequence simulation software package *msprime*. To use *msprime*, we first need to convert the transmission network data into the format required by this software package. The only information contained in the network that we require is the time and

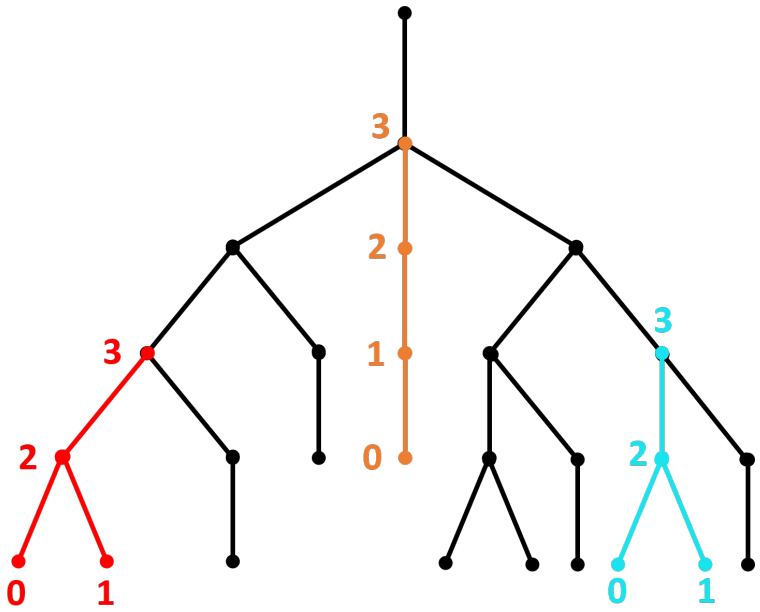


Figure 4: Illustration of a simple network and three ways you could select sub-networks of 4 individuals with one root infector (each individually labelled 0-3 and colour-coded). Circles are nodes or 'individuals', and lines represent infection events between two individuals (Top-down infection direction).

direction of infection events between individuals in our network. The simulation is based on the concept of generations, e.g. mutation and recombination rates are implemented in units 'per generation'; thus, we need to convert the time of infection, listed as a date in years by Probert et al., into the corresponding number of HIV generations elapsed in the simulation. Since one generation is one day, (see section 1.1), this is equivalent to days elapsed, and is a straightforward transformation from years to days (1 year is taken as 365 generations).

The *msprime* package was developed in the context of human evolution which means there are some terms to adapt for our viral evolution model. The standard set-up in *msprime* is to consider 'populations' of diploid 'individuals' (e.g. a geographically defined population of humans), where we want to track the ancestry of both genomes in each individual. In HIV, each 'individual' in the simulation is one virus particle with one genome (haploid), and 'populations' can be thought of as human hosts which each contain a large number of viral particles. Transmission events between hosts are directly comparable to human migration events where a member of one population (an HIV particle in a person) migrates to form another population elsewhere (an HIV infection in another person). Hence, we use 'migration events' defined in *msprime* to model transmission events, which seed new populations (infected hosts) containing viral particles. It is key that we encode each human host as a 'population' in *msprime* and consider transmission events, since we want to study the evolutionary history of the population of HIV particles in each person, as well as the relationships between these populations of HIV. Representing individuals as populations of HIV also allows us to manipulate population sizes (effective number of HIV particles) in each individual in *msprime*; to encode the single transmitted particle from another population followed by growth of the viral population in each newly infected individual.

A further technicality in our simulations is that we define a source population, such that the first individual, functioning as the root infector in our chosen sub-network, is also modelled by a standard single-infection transmission event. The time of infection seeding our first individual, infected from the source population, is given in Probert et al.'s data (see fig. 5). The source population is a convenient tool which serves to ensure that all populations coalesce, and we do not need to take samples to inspect any HIV haplotypes within this source population. Furthermore, the source population allows us to include the timescale of infection seeding, growth, and transmission, for the first true individual in our transmission network.

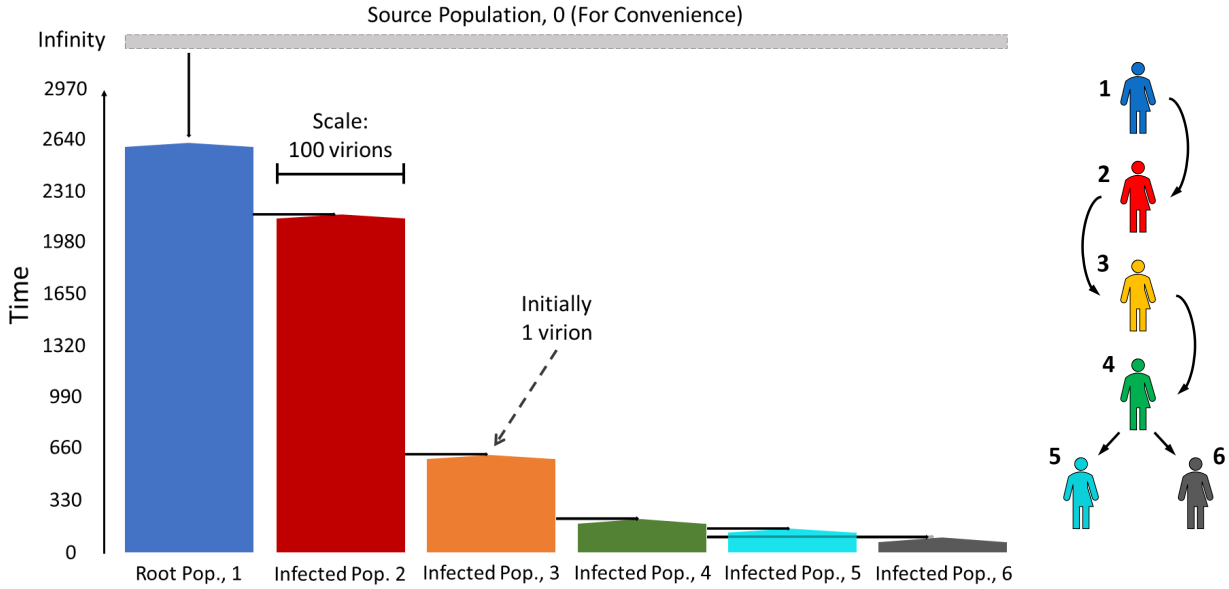


Figure 5: Illustration of transmission model with a simple network comprising 6 individuals/populations (Pop.) including one root infector. Each bar represents viral population in an individual host, arrows denote direction of transmission between hosts, the height is time in days (where taller is older) and width is the population size of virions in the host. The source population is arbitrarily large and of constant size. The schematic on the right re-illustrates the transmission network displayed within the bar graph.

2.3 Sampling

Core to the research and result analysis is the data we create; for this project we need to take sample genome sequences from our simulated HIV populations. In order to achieve an observable signal of the transmission pathways, preserving the possibility of genealogical inference of who infected whom, given the high rates of mutation, recombination, and fast replication cycle in HIV, we need to take samples from individuals close to the time of infection. From discussions with HIV collaborators (Matthew Hall et al.), taking many samples a month (30 days) after infection was deemed suitable and representative of typical sampling times in real-world infection scenarios. Since it helps us to build, visualise, and analyse genealogy in the trees, we will also take an equal number of contemporaneous samples, from all individuals at the end of the simulation, to the samples taken near-infection.

If we run such a simulation, sampling both contemporaneously and near-infection, we pro-

duce trees like the one shown in fig. 6. The tree in fig. 6 follows the simple model in fig. 5 with HIV-like evolution parameters (mutation rate, recombination rate, replication cycle), where we have taken 5 samples from each population 30 days after its infection event, and 5 samples from each population at the end of the simulation. The vertical axis corresponds to time, the samples close to infection time appear as tips deep in the tree, such as those on the far right side (e.g. samples 10-12). In contrast, the contemporaneous samples appear at the bottom of the tree, e.g. samples 5-9. We can clearly see that the transmission pathway reflects that which we are intending to model (as shown in fig. 5), since infector populations seed their infected population(s) in the order expected.

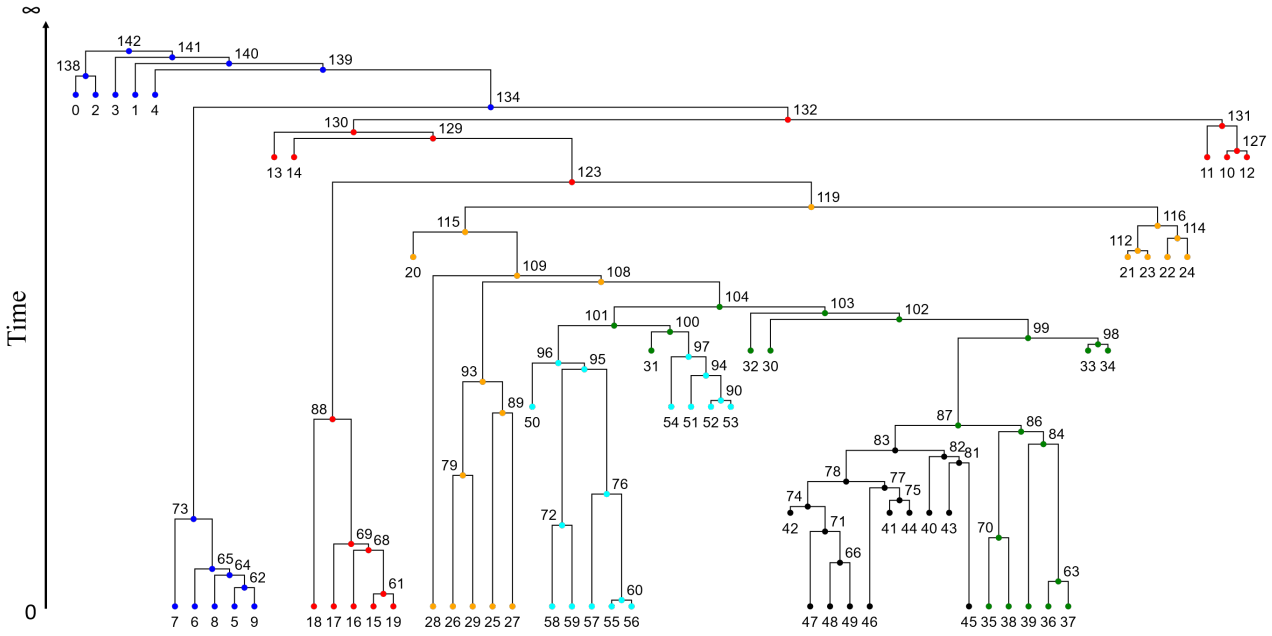


Figure 6: Simulated tree from arbitrary location along the tree sequence for the simple transmission network illustrated in fig. 5. Colour coding of populations, and their order, are the same as in fig. 5. 5 samples are taken from each population 30 days after the infection event for that individual, and 5 samples are taken at the end of the simulation (most recent time). Height is given by ranked age, where taller is older in time.

2.4 Missing Data: Truncating Sequences

Another important property that we should model is the 'missing data problem' that HIV sampling yields only short sequence fragments, as explained in section 1.1. We account for this after simulating genealogies based on full sequence data (see fig. 6). Once we have our complete simulated tree sequence samples, we then delete some of the information associated with each sample and truncate the sequences to just a fraction of their full length. Each sample no longer spans the whole genome, but only a certain range of chromosomal positions along the genome. We specify the average genomic span of each sample (e.g. 200 bp) and use a Poisson distribution to select the actual genomic span of each sample. This is exemplified in fig. 7, which illustrates that each sample sequence only covers a portion of the entire genomic span, and so each sample only appears in some of the trees along the tree sequence. Contrast this

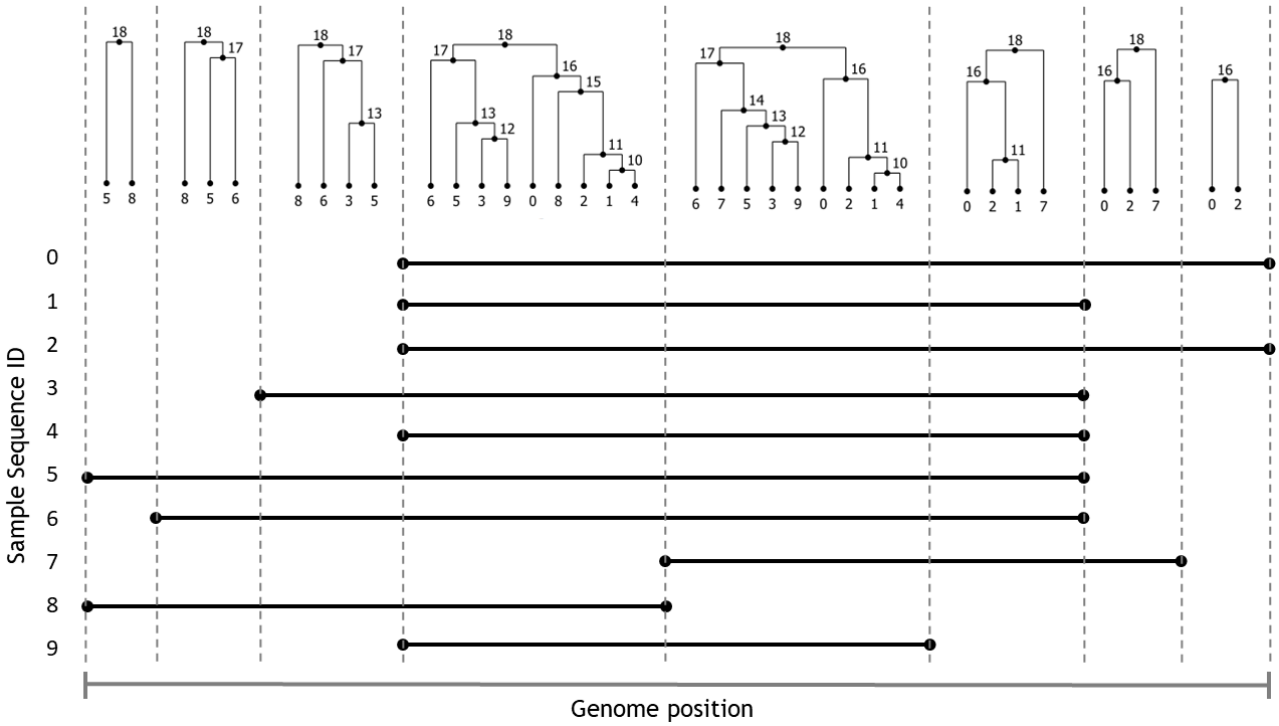


Figure 7: Illustrating effect of fragmented sequences on trees along the genome.

to fig. 3, which displays typical non-truncated data. Here, the samples comprise whole genome sequences and thus each tree in the tree sequence contains every sample.

Implementing sequence truncation required several amendments to the tree sequence software packages, including handling the missing data (no information at some genomic positions for samples) and handling the variation in number of samples for trees along the sequence. Both the data format which stores the tree data and visualisation code for trees needed to be updated to be compatible with these cases (within the *tskit* software package). Furthermore, when it comes to inferring the genealogy based only upon the sample sequences, the tree-based inference software *tsinfer* also needs to handle this missing data, and so several modifications to the *tsinfer* code were also required.

3 Preliminary Inference Results

One way to test the performance of our inference is to compute the Kendall-Colijn (KC) distance between our inferred and simulated trees, which is a metric to quantify the similarity of different trees [18]. The KC distance can be calculated considering just the topology of the trees (connections between nodes), or just the branch lengths (time between events), or by considering a weighted combination of the two. Since our inference protocol does not place non-contemporaneous samples at the right time on inferred trees (due to algorithmic complications in implementation), we do not want consider branch lengths, and thus use topology-only KC as our metric.

As a first pass, we computed KC for inferred trees using whole-genome (non-truncated) samples (fig. 8). We have calculated KC for each tree along the simulated tree sequence, and at the time of writing, we are only able to do this for non-truncated sample data. In this case,

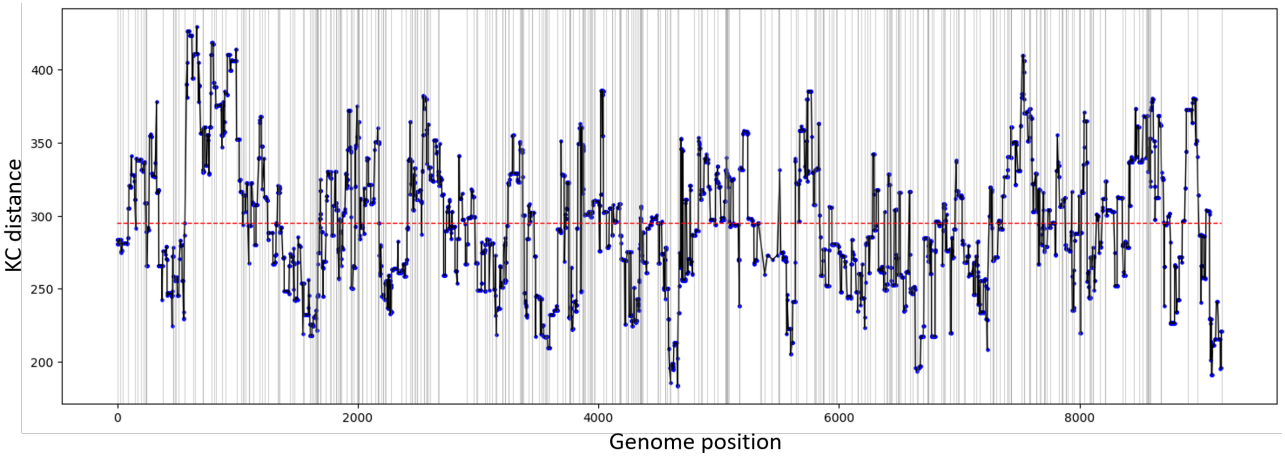


Figure 8: Illustrating comparison of simulated and inferred trees using topology only KC distance. These simulated and inferred trees contain contemporaneous and non-contemporaneous samples, but use whole genome sequences. Blue dots are calculated KC for each tree, black connects the KC points to emphasise variation; red line is the mean KC (294.9), and grey vertical lines show the genomic breakpoints of the inferred trees.

each tree along the tree sequence contains all the samples, which simplifies the calculations, and the variation in KC for different inferred trees along the sequence is immediately meaningful. In contrast, if some trees have a different number of samples to other trees in the sequence, then it hard to compare the KC between different trees along the sequence since the metric itself does not include any normalisation to account for the number of samples in a tree. Thus, whilst you can compare two trees with x samples and get a KC measure of their similarity, k_x, if we compare two trees with a different number of samples, y, and with a KC, k_y, then the KCs are not directly comparable, i.e. $k_y > k_x$ does not automatically imply that the two trees with x samples are more similar to each other than the two trees with y samples. To be able to compare simulated and inferred trees using fragmented sequences is not trivial, and will require further consideration. We must determine a method to quantitatively compare the two systems (whole and fragmented sequences), and first we must be able to infer trees which use both fragmented data and non-contemporaneous samples, to allow direct comparison with this whole-genome benchmark case. Nevertheless, the similarity between inferred trees using whole-genome sequences can act as a baseline for the accuracy of trees inferred with fragmented sequences.

Figure 8 shows the KC between simulated and inferred trees, which use non-contemporaneous and contemporaneous whole genome sample sequences, along the HIV genome. Clearly there is a lot of variation in KC between for trees along the genome, with the maximum KC (429.6) more than twice that of the minimum (183.6). An important point to understand is that there is a different number of simulated and inferred trees, 18783 and 352 respectively, but that both tree sequences span the entire genome. The inference clearly underestimates the impact of recombination and fails to replicate the frequency of tree changes across the genome, and hence estimates the presence of fewer trees in the sequence. This is a huge difference in number of trees, and must contribute to the sheer level of variation seen in KC in fig. 8. The grey lines

included in the figure give the boundaries (breakpoints) in genomic position between inferred trees. There are more KC measurements than inferred trees since we compare to the tree spanning the lesser range of genome, i.e. multiple simulated trees will be compared to the same inferred tree which contains the span of each simulated tree.

Although we cannot draw many conclusions from this metric alone, especially in this small example system, metrics such as these will become useful as this research continues. Once we have more simulations and inferred trees which simultaneously incorporate fragmented and non-contemporaneous data, we will be able to use baseline metrics from whole-genome sequence data, and compare results in different scenarios (e.g. purely single vs single and dual infection events). Further tests will be developed to better assess the performance of the model, such as the accuracy of predicted transmission direction from inferred trees using our tree-sequence method, compared to alternative inference methods.

4 Discussion

In this project, I have developed open-source tree-sequence software tools in collaboration with other researchers (Yan Wong, Wilder Wohns & Jerome Kelleher) to handle qualities of HIV evolution. These included finite site mutations and fragmented sequences ('missing data'), incorporating non-contemporaneous samples in inferred tree sequences, and fasta output of simulated sequences in order to integrate with other software tools. Whilst the analysis completed so far is limited, developing the methods and code to simulate HIV evolution, and debugging compatibility issues between different software tools (e.g. *tskit* handling missing data but *tsinfer* failing) was a time-consuming process. By the time it was possible to infer trees including non-contemporaneous samples, and compare these inferred trees to the simulated ones, there was not time to adequately assess the accuracy of the inferences. It was particularly difficult to use a metric in the case of fragmented sequences, due to varying sample numbers between different trees, and there was not enough time to solve this issue in this project. Nor did we manage to infer trees using both non-contemporaneous samples and fragmented sequences, we could only consider one complexity at a time for our inferences: *tsinfer* could handle fragmented sequences, but the method of including non-contemporaneous samples did not work with fragmented sequences. Although there is a reasonable amount of complexity in our simulated data, our pipeline remains incomplete because we cannot make inferences and test them with the full set of parameters we had included in our simulations.

We have completed a first step for testing inferred ancestry by measuring KC distance against simulated ancestry, based only on complete genome samples. This can act as a baseline to test how well the tree-based inference performs with complete data, compared to existing methods, and as a baseline for evaluating fragment-based inferences in the future. This will require some additional software development in order to enable both non-contemporaneous and fragmented samples to be included in the inference process, as well as careful consideration of how to evaluate the accuracy of inferences under these conditions. After exploratory small population models have been evaluated, and refined, we then aim to scale up to larger popu-

lations of 1000s or more individuals, and carry forward our approach to infer ancestry for real (fragmented sequence) HIV datasets.

References

[1] J. Kelleher, A. Etheridge, and G. McVean. Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLoS Computational Biology*, 12(5):e1004842, 2016.

[2] J. Kelleher, Y. Wong, A. Wohns, C. Fadil, P. Albers, and G. McVean. Inferring whole-genome histories in large population datasets. *Nature Genetics*, 51(9):1330–1338, 2019.

[3] World Health Organisation. HIV/AIDS Factsheet [Online]. https://www.who.int/news-room/fact-sheets/detail/hiv-aids, 2019. Last accessed 18-09-2019.

[4] J. Coffin and R. Swanstrom. HIV pathogenesis: dynamics and genetics of viral populations and infected cells. *Cold Spring Harbour Perspectives in Medicine*, 13:6a012526, 2013.

[5] D. Shriner, A. G. Rodrigo, D. C. Nickle, and J. I. Mullins. Pervasive Genomic Recombination of HIV-1 in Vivo. *Genetics*, 167(4):1573–1583, 2004.

[6] K. A. Lythgoe and C. Fraser. New insights into the evolutionary rate of HIV-1 at the within-host and epidemiological levels. *Proceedings of the Royal Society B: Biological Sciences*, 279(1741):3367–3375, 2012.

[7] A. D. Redd, T. C. Quinn, and A. A. R. Tobian. Frequency and implications of hiv superinfection. *The Lancet. Infectious Diseases*, 13:622–8, 2013.

[8] S. Andrews and S. Rowland-Jones. Recent advances in understanding HIV evolution. *F1000Research*, 6(597), 2017.

[9] J. M. Cuevas, R. Geller, R. Garijo, J. López-Aldeguer, and R. Sanjuán. Extremely High Mutation Rate of HIV-1 In Vivo. *PLoS Biology*, 13:e1002251, 2015.

[10] M. Kearney, F. Maldarelli, W. Shao, J. B. Margolick, E. S. Daar, J. W. Mellors, V. Rao, J. M. Coffin, and S. Palmer. Human Immunodeficiency Virus Type 1 Population Genetics and Adaptation in Newly Infected Individuals. *Journal of Virology*, 83(6):2715–2727, 2009.

[11] D. Cromer, A. J. Grimm, T. E. Schlub, J. Mak, and M. P. Davenport. Estimating the in-vivo HIV template switching and recombination rate. *AIDS*, 30(2):185–192, 2016.

[12] P. Lemey, A. Rambaut, and O. G. Pybus. HIV evolutionary dynamics within and among hosts. *AIDs Review*, 8(3):125–140, 2006.

[13] W. Probert, R. Sauter, A. Cori, M. Pickles, and C. Fraser. PopART-IBM Report. https://github.com/BDI-pathogens/POPART-IBM/tree/master/doc/Model_V1_4.pdf, 2018. Last accessed 04-10-2019.

[14] W. Probert, R. Sauter, A. Cori, M. Pickles, S. Floyd, D. MacLeod, E. A. Wilson, D. Donnell, H. Ayles, NuldaBeyers, P. Bock, S. Fidler, R. Hayes, and C. Fraser. HPTN071 PopART Analysis Plan IBM projections [Online]. https://www.hptn.org/sites/default/files/2018-12/HPTN071_PopART_AnalysisPlan_IBMprojections.pdf, 2018. Last accessed 04-10-2019.

[15] National Center for Biotechnology Information (NCBI). HIV-1, complete genome [Online], 1999. Accession No. AF033819.3. Available from: https://www.ncbi.nlm.nih.gov/nuccore/AF033819.3?report=fasta, Last accessed 20-09-2019.

[16] J. Kelleher, K. Thornton, J. Ashander, and P. Ralph. Efficient pedigree recording for fast population genetics simulation. *PLoS Computational Biology*, 14(11):e1006581, 2018.

[17] R. R. Hudson. Properties of a neutral allele model with intragenic recombination. *Theoretical Population Biology*, 23(2):183–201, 1983.

[18] M. Kendall and C. Colijn. Mapping Phylogenetic Trees to Reveal Distinct Patterns of Evolution. *Molecular Biology and Evolution*, 33(10):2735–2743, 2016.