Inferring Multi-Type Birth-Death Parameters for a Structured Host Population with Application to HIV Epidemic in Africa

Hassan W. Kayondo¹, Samuel Mwalili², John M. Mango³

¹Pan African University, Institute of Basic Sciences, Technology and Innovation, Nairobi, Kenya

²Department of Statistics and Actuarial Sciences, Jomo Kenyatta University of Agriculture and Technology, Nairobi, Kenya

³Department of Mathematics, Makerere University, Kampala, Uganda

Email: whkayondo@gmail.com, samuel.mwalili@gmail.com, mango@cns.mak.ac.ug

How to cite this paper: Kayondo, H.W., Mwalili, S. and Mango, J.M. (2019) Inferring Multi-Type Birth-Death Parameters for a Structured Host Population with Application to HIV Epidemic in Africa. *Computational Molecular Bioscience*, 9, 108-131.

<https://doi.org/10.4236/cmb.2019.94009>

Received: May 2, 2019

Accepted: September 10, 2019

Published: December 13, 2019

Copyright © 2019 by author(s) and Scientific Research Publishing Inc. This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>

Open Access

Abstract

Human Immunodeficiency Virus (HIV) dynamics in Africa are purely characterised by sparse sampling of DNA sequences for individuals who are infected. There are some sub-groups that are more at risk than the general population. These sub-groups have higher infectivity rates. We came up with a likelihood inference model of multi-type birth-death process that can be used to make inference for HIV epidemic in an African setting. We employ a likelihood inference that incorporates a probability of removal from infectious pool in the model. We have simulated trees and made parameter inference on the simulated trees as well as investigating whether the model distinguishes between heterogeneous and homogeneous dynamics. The model makes fairly good parameter inference. It distinguishes between heterogeneous and homogeneous dynamics well. Parameter estimation was also performed under sparse sampling scenario. We investigated whether trees obtained from a structured population are more balanced than those from a non-structured host population using tree statistics that measure tree balance and imbalance. Trees from non-structured population were more balanced basing on Colless and Sackin indices.

Keywords

HIV, Likelihood Inference, Multi-Type Birth-Death Process, Probability of Removal, Structured Population

1. Introduction

Human Immunodeficiency Virus (HIV) is among the most infectious diseases

globally. According to [1], an estimated 36.7 million people were living with HIV globally in 2016. Sub-Saharan Africa is the most affected by the epidemic as of 2016, 19.4 million people were living with HIV in Eastern and Southern Africa. Of the total HIV infections in 2016, 6.1 million infections were registered in Western and Central Africa [1]. This implies that about 70% of all people living with HIV resided in Africa in 2016. The world population nearly hit 7.6 billion as of mid-2017 [2]. African continent had approximately 1.3 billion, translating into only 17% of the global population [2]. African continent therefore has less than twenty percent of the global population, yet having almost three quarters of the total number of people living with HIV Worldwide. With Africa having the most of HIV infections, the epidemic poses a threat to both health and socio-economic development of people in Sub-Saharan Africa region. African HIV epidemic is generalised in most parts of the continent. According to [3], a generalised HIV epidemic has a prevalence greater than 1% in the general population and the epidemic self-sustains through heterosexual transmission. For developed countries, HIV epidemic is concentrated. A concentrated HIV epidemic is defined in [3] as where HIV has spread in one or more sub-populations while the general population has a prevalence of less than 1%. The generalised HIV epidemic in Africa requires slightly different methods compared to a concentrated epidemic.

Many African countries are battling with HIV epidemic. For example, the total burden of HIV in Uganda is increasing according to spectrum estimates from the Ministry of Health as documented in [4]. The number of individuals living with HIV in Uganda rose from 1.4 million in 2013 to 1.5 million in 2015. This was due to continuous spread of the disease and increased life expectancy for people living with HIV [4]. From [4], the number of AIDS related deaths was 28,000, while new HIV infections were 83,000 in 2015. In Uganda, HIV infects individuals disproportionally according to both their geographical and risk behaviours [5]. Some groups are more at risk compared to the general population in Uganda. Such groups have high incidence for HIV. These groups include sexual workers, fish folk community and truck drivers [6] [7] [8] [9]. It was observed that sex workers, their clients and regular partners of clients, contributed between 7% and 11% of new infections in Uganda, Swaziland and Zambia [8]. Dis-proportionality is also visualised among females and males. Although the general prevalence of HIV among adults in Uganda was 6.2% in 2016, it was 7.6% among females and 4.7% among males [10]. HIV prevalence differs in various regions in Uganda as well. It was stated in [10] that the prevalence of HIV in Central, Eastern, West Nile, Northern and Western Uganda was 8.0%, 5.1%, 3.1%, 5.7% and 7.9%, respectively in 2016. Therefore, HIV epidemic in Uganda is structured as well as generalised.

In order to study such a host population so as to make inference about HIV epidemic, viewing it as a structured population gives more reliable parameter inference due to heterogeneity manifested by individuals in terms of infectivity. In this paper, we are interested in a model that depicts HIV dynamics in an

African setting. Dis-proportionality in terms of infectivity is one of the characteristics significant for HIV dynamics in Africa. In order to account for this, we employ models that make inference for structured host population. Incorporating such models for HIV dynamics in Africa is still a challenge as the models should entail characteristics like sparse sampling and heterogeneity in infectivity among infected individuals. We incorporate such characteristics in the model for a structured host population. This aids in making parameter inference for HIV epidemic in an African setting. Molecular sequence data uncovers underlying disease dynamics in structured host populations much better than prevalence and incidence data.

Molecular sequence data is increasingly becoming available even in Africa, for example Southern African Treatment and Resistance Network has a growing database of greater than 70,000 HIV sequences [11]. The availability of this genomic data has resulted in analysis of DNA HIV sequences, for example, The Phylogenetics and Networks for Generalized Epidemics in Africa consortium (PANGEA-HIV) utilises viral sequence analyses to establish transmission patterns of HIV in Africa [11]. Analysis of sequence data is explored by many researchers for making inference about epidemiological parameters of epidemics like in computation of basic reproductive number. The basic reproductive number of the pathogen for HIV-1 epidemic in Switzerland was estimated using DNA sequences [12]. Some researchers have integrated both sequence data and compartment models to make inference about disease dynamics. The use of sequence data with a compartmental susceptible-infected-removed (SIR) enabled the joint estimation of epidemiological parameters and establishing disease origin [13]. The method was applied to HIV-1 sampled data from the UK to estimate the basic reproductive ratios within clusters and on hepatitis C virus data set from Argentina to establish the origin of the disease [13]. Analysis of sequence data involves using phylogenetic trees.

We extend the multi-type birth death model of likelihood inference implemented by [14]. We incorporate an additional parameter which is the probability of removal from an infectious pool. In this paper, a sampled individual becomes non-infectious immediately after sampling with a probability r and remains infectious with probability $1-r$. The added parameter accounts for scenarios where some infected individuals remain infectious even upon diagnosis like in most of Africa's HIV epidemic. Most of individuals who are diagnosed with HIV are not necessarily removed from the infectious pool owing to risky behaviour practised by some individuals. We extend the likelihood inference due to its flexibility in varying other parameters like sampling proportions of different types or sub-groups when making parameter inference. We estimate parameters by varying the sampling proportions as well. The effect of sparse sampling on parameter inference is investigated as well. This is because, analysing parameter inference in such scenarios of sparse sampling is important since HIV epidemic in Africa is still characterized by limited DNA sequences for HIV positive individuals. The model employed in this paper was used to infer parameter

estimation in a structured population with two types. The model distinguished between heterogeneous and homogeneous dynamics very well. We also succeeded in making inference about tree balance for both structured and non-structured populations.

The model used in this paper differs from the Bayesian inference that was implemented by [15] as it extends the likelihood techniques of a birth-death multi-type model developed by [14]. In the Bayesian framework, both epidemiology and fossil calibration of parameters were estimated. We extended the likelihood inference of [14] by adding an extra parameter that accounts for individuals that remain infectious even after being diagnosed with HIV. The extended model which is used for parameter inference in this paper is helpful in studying HIV epidemic in an African setting. Section 2 gives a detailed explanation of the method we are employing as well as explaining extension of the model in [14]. The results of our study are presented in Section 3, while discussion of the results and conclusions are presented in Section 4.

2. Methods

2.1. Modelling Framework

Using ideas developed in [14] on multi-type birth-death process (MTBD), we develop a multi-type HIV model that captures the dynamics of the spread of HIV within the context of a generalised HIV epidemic in Africa. In this model, the host population comprises of different demes. A deme is a subdivision of a population that consists of individuals that possess closely related characteristics. The demes are represented by various types. The disease dynamics are modelled using birth-death processes, hence the name multi-type birth-death infection model. Within each deme, HIV dynamics are homogeneous. However, at the population level, the resultant dynamics are heterogeneous. **Figure 1** shows a structured host population with two demes and the parameters for the model.

Figure 1. An illustration of a multi-type model with 2 demes.

For example, deme 1 can be a general population and deme 2 a sub-population at risk.

We describe the parameters in the first deme. Those for the second deme are described analogously. The following are the parameters for the model in deme 1;

1) Parameter λ_{11} is the rate at which an individual in deme 1 transmits or infects an individual of deme 1. Similarly, λ_{12} is the rate at which an individual in deme 1 transmits to an individual of deme 2. These are the speciation (transmission) rate parameters in deme 1. There is both within (λ_{11}) and across (λ_{12}) deme transmission in this multi-type birth-death model.

2) Parameter μ_1 is the rate at which an individual in deme 1 dies. This is the extinction rate parameter.

3) Parameter ψ_1 is the rate at which an individual of deme 1 is sampled. Sampling in this context is the way HIV sequences are included into the study.

4) Parameter r is the probability that an individual in either of the demes becomes non-infectious upon sampling (being removed from infectious pool). An individual therefore remains infectious with probability $1-r$. Parameter r is the added parameter in our model which is not in the model presented in [14]. In our model framework, an individual either remains infectious or is removed from the infectious pool upon sampling. Therefore, $r\mu_1\psi_1$ is the rate at which an individual in deme 1 is removed from the infectious pool upon sampling. The rate at which an individual in deme 1 who remains infectious, transmits to an individual in deme 1 upon sampling is $(1-r)\lambda_{11}\psi_1$. The rate, $(1-r)\lambda_{12}\psi_1$ is analogously defined.

For the derivation of the likelihood inference for parameter estimation, we still build on ideas used in [14]. **Figure 2** illustrates the multi-type birth-death tree that we use in the derivation for likelihood of multi-type birth-death tree.

In **Figure 2**, black dots are sampled individuals, individuals either belong to deme 1 or 2. E is an edge which represents an individual. For parameters on the

Figure 2. Multi-type birth-death trees. (A) A complete tree with 2 demes; (B) A sampled tree obtained after deleting non-sampled individuals from a complete tree.

dotted vertical line, 0 is the present, τ is the time at which a certain individual that belongs to deme 2 was sampled, t is the time taken to give raise to the sub-tree (dashed oval) that is subtended by an individual represented by an edge E . T is the time taken for the process or epidemic. This is the time from the origin of the epidemic to present. The model illustrated in **Figure 1** results in a multi-type tree as visualized in **Figure 2**. In our model, the multi-type tree can have a sampled descendant, while a sampled individual in the model presented in [14] had no sampled descendants.

In this derivation, $D_{E1}(t)$ is the probability density for an individual represented by an edge E in deme 1 giving raise to the sub-tree of age t (dotted oval). For $D_{E1}(t)$, $t > \tau$ as shown in **Figure 2**, where τ is the time at sampling of an individual. This implies that time increases backwards from present to the origin. We define another term $Z_1(t)$, which is the probability that an individual in deme 1 is not sampled and does not have a descendant between t and 0. In order to obtain $D_{E1}(t + \Delta t)$, we assume that $D_{E1}(t)$ is known and computed where Δt is a very small time step. Along an edge E for time step Δt , either no event happens or birth events happen. This gives:

$$D_{E1}(t + \Delta t) = \left(1 - \left(\sum_{j=1}^2 \lambda_{1,j} + \mu_1 + \psi_1 \right) \Delta t \right) D_{E1}(t) + \sum_{j=1}^2 \lambda_{1,j} \Delta t Z_j(t) D_{E1}(t) + \sum_{j=1}^2 \lambda_{1,j} \Delta t Z_1(t) D_{Ej}(t) + O(\Delta t^2) \quad (1)$$

From Equation (1), the first term on the right is for no birth, death and sampling events, second term represents birth of an individual of deme j while lineage j produces no samples in time t , third term signifies birth of an individual of deme j while lineage 1 produces no samples in time t . The last term on right is the probability that more than one event occurs in the time interval Δt .

After re-arranging the terms in Equation (1) and letting $\Delta t \rightarrow 0$, we obtain the differential equation:

$$\frac{d}{dt} D_{E1}(t) = - \left(\sum_{j=1}^2 \lambda_{1,j} + \mu_1 + \psi_1 \right) D_{E1}(t) + \sum_{j=1}^2 \lambda_{1,j} Z_j(t) D_{E1}(t) + \sum_{j=1}^2 \lambda_{1,j} Z_1(t) D_{Ej}(t) \quad (2)$$

The initial condition at τ is given as:

$$D_{E1}(\tau) = r\mu_1\psi_1 + (1-r) \sum_{j=1}^2 \lambda_{1,j}\psi_1, \quad \text{for } j=1, \text{ and } D_{E1}(\tau) = 0, \text{ for } j \neq 1, \quad (3)$$

The first term on the right of Equation (3) depicts individuals who become non-infectious upon sampling while the last term signifies those who are not removed from the infectious pool, *i.e.*, they remain infectious even after being sampled. In typical African communities, many individuals infected with HIV remain infecting others either knowingly or unknowingly. When r is 1, the initial condition given in Equation (3) reduces to that used in [14].

Equation (2) is used to obtain the quantity $D_{E1}(T)$ which is the desired probability density for the multi-type birth-death tree. The derivation for $Z_1(t)$ is required to obtain the density of the tree. $Z_1(t)$ is derived using similar steps as those used in the derivation for Equation (2). The differential equation for $Z_1(t)$ is given as:

$$\frac{d}{dt}Z_1(t) = (1 - \psi_1)\mu_1 - \left(\sum_{j=1}^2 \lambda_{1,j} + \mu_1 + \psi_1 \right) Z_1(t) + \sum_{j=1}^2 \lambda_{1,j} Z_1(t) Z_j(t), \quad (4)$$

with the initial condition,

$$Z_1(0) = Z_2(0) = 1. \quad (5)$$

The first term on the right of Equation (4) represents death without sampling, the second signifies no birth, death and sampling as well as an individual in deme 1 producing no samples. The last term depicts a birth of an individual j but both individuals 1 and j producing no samples in time t . The initial condition given in Equation (5) implies that an individual at time $t = 0$ is not sampled. This is the case because the model used accounts for sampling through time only. Contemporaneous sampling, which is sampling at only one time point especially the present is not modelled in this inference.

The differential equations (2) and (4) with initial condition given by (3) are for deme 1. The corresponding differential equations for deme 2 were analogously derived. The system of differential equations with their initial conditions that is integrated along a given edge for 2 demes is given as:

$$\begin{aligned} \frac{d}{dt}Z_1(t) &= (1 - \psi_1)\mu_1 - (\lambda_{1,1} + \lambda_{1,2} + \mu_1 + \psi_1)Z_1(t) + \lambda_{1,1}Z_1^2(t) + \lambda_{1,2}Z_1(t)Z_2(t) \\ \frac{d}{dt}Z_2(t) &= (1 - \psi_2)\mu_2 - (\lambda_{2,1} + \lambda_{2,2} + \mu_2 + \psi_2)Z_2(t) + \lambda_{2,1}Z_2(t)Z_1(t) + \lambda_{2,2}Z_2^2(t) \\ \frac{d}{dt}D_{E1}(t) &= -(\lambda_{1,1} + \lambda_{1,2} + \mu_1 + \psi_1)D_{E1}(t) + (\lambda_{1,1}Z_1(t) + \lambda_{1,2}Z_2(t))D_{E1}(t) \\ &\quad + \lambda_{1,1}Z_1(t)D_{E1}(t) + \lambda_{1,2}Z_1(t)D_{E2}(t) \\ \frac{d}{dt}D_{E2}(t) &= -(\lambda_{2,1} + \lambda_{2,2} + \mu_2 + \psi_2)D_{E2}(t) + (\lambda_{2,1}Z_1(t) + \lambda_{2,2}Z_2(t))D_{E2}(t) \\ &\quad + \lambda_{2,1}Z_2(t)D_{E1}(t) + \lambda_{2,2}Z_2(t)D_{E2}(t) \\ Z_1(0) &= Z_2(0) = 1 \\ D_{E1}(\tau) &= r\mu_1\psi_1 + (1-r)(\lambda_{1,1} + \lambda_{1,2})\psi_1 \\ D_{E2}(\tau) &= r\mu_2\psi_2 + (1-r)(\lambda_{2,1} + \lambda_{2,2})\psi_2 \end{aligned} \quad (6)$$

Solving the system of equations in (6) numerically gives the calculations along an edge of a tree. This is then followed by pruning at the nodes and finally reaching the root of the tree. This results in obtaining either $D_{E1}(T)$ or $D_{E2}(T)$ depending on whether the root was in deme 1 or 2.

2.2. Probability Density of the Tree

The probability density of the tree with the first individual at time T conditioned on observing at least an individual who is sampled for a structured population

with 2 demes is given as:

$$p(T | \lambda, \mu, \psi, T) = f_1 \frac{D_{E1}(T)}{1 - Z_1(T)} + f_2 \frac{D_{E2}(T)}{1 - Z_2(T)} \quad (7)$$

The probability that an individual at time T is of deme 1 is given by f_1 and its distribution must be specified. Equation (7) is the likelihood of the parameters given the data. This is used for parameter inference. It should be noted that neither $Z_1(T)$ nor $Z_2(T)$ can be 1. This is because if any takes on 1, then the process will not give rise to the inferred tree. The conditioning is necessary since studies have shown that this conditioning gives more accurate parameter values [16]. The parameters that we estimated were:

$$\lambda = (\lambda_{1,1}, \lambda_{1,2}, \lambda_{2,1}, \lambda_{2,2}) \quad \mu = (\mu_1, \mu_2), \quad \psi = (\psi_1, \psi_2), \quad r \quad \text{and} \quad T.$$

2.3. Parameter Inference

The necessary changes suggested in Equation (3) to reflect HIV dynamics in an African setting were effected in R using an R package, Treesim of [17]. The parameters for simulated trees were then estimated using an R package, TreePar of [18] with changes to suit our model. In the first set of simulations, 50 trees with 100 tips were simulated under multi-type birth-death model with 2 demes (MTBD-2), with sampling probability of 0.2 for each of the two sub-populations ($\psi_1 = \psi_2 = 0.2$). We first set $r = 0.2$ and $r = 0.8$ to investigate the effect of the probability of removal on parameter inference. We then increased the number of simulated trees to 100 but kept $r = 0.8$. This was to investigate the effect of increasing the number of simulated trees on parameter estimates.

Another set of simulations was done for 100 simulated trees, with 200 sampled tips. We varied the probability of removal, taking on values, $r = 0.8$ and $r = 0.5$. The sampling probability remained at 0.2 for both sub-populations. In another set of simulations, the sampling probabilities were set to $\psi_1 = \psi_2 = 0.05$ for both the sub-populations. This depicted sparse sampling which is highly the case in several HIV epidemics for Africa. Again the probability of removal was varied as it was set at $r = 0.2$ and $r = 0.8$. To investigate the effect of varying the sampling proportions, we simulated trees when the sampling proportions were different. We set $\psi_1 = 0.2$ and $\psi_2 = 0.01$ for this kind of parameter inference.

2.4. Heterogeneous and Homogeneous Dynamics

For comparison between two models, we used likelihood ratios. From [19], suppose that H_0 is derived from a full alternative (H_1). Suppose further Θ_0 and Θ_1 are defined such that H_0 corresponds to a situation that a true parameter θ is in $\Theta_0 \subseteq \theta$, and H_1 corresponds to a case $\theta \in \Theta_1 - \Theta_0$. The log likelihood ratio statistic is given as:

$$LR = -2 \left[\max_{\theta \in \Theta_0} \log(L(\theta)) - \max_{\theta \in \Theta_1} \log(L(\theta)) \right] \quad (8)$$

In [20], they defined the likelihood ratio statistic as $2 \left(\frac{\ln(X|H_1)}{\ln(X|H_0)} \right)$, where $(X|H)$ is the likelihood of the data (X) under hypothesis H . This likelihood statistic simplifies to the one given in Equation (8). Likelihood ratio statistic is asymptotically chi-squared distributed under the null hypothesis according to [19].

For heterogeneous dynamics, we assume that the population is structured into demes, say 2 as in our model. Each of the deme has different birth and death parameters. For homogeneous case, the population is not structured and individuals have the same birth and death rates. We investigated how well our model distinguished between heterogeneous and homogeneous dynamics. For simplicity, we had two sets of simulations, named A and B.

In A, both trees simulated and parameter inference were made under MTBD-2. For B, parameter inference was made for multi-type birth-death model with 1 deme (MTBD-1), yet trees were obtained under MTBD-2. We then performed likelihood tests to find out the number of trees that were rejected in support of the MTBD-2 (heterogeneous) dynamics. For homogeneous dynamics, we simulated trees under MTBD-1 and made parameter inference under both MTBD-1 and MTBD-2. Likelihood ratio tests were used to find out how many trees were rejected whose inference was made under MTBD-2.

Since the use of likelihood ratios requires that H_0 is derived from a full alternative H_1 , in A, we tested H_0 : Tree simulation under MTBD-2 and parameter inference under MTBD-1 against H_1 : Tree simulation under MTBD-2 and parameter inference under MTBD-2. In B, we had H_0 : Tree simulation under MTBD-1 and parameter inference under MTBD-1 against H_1 : Tree simulation under MTBD-1 and parameter inference under MTBD-2. H_0 and H_1 are defined as null and alternative models, respectively.

2.5. Simulated DNA Sequences

In the absence of real DNA sequence data for a structured host population, DNA sequences were simulated using a stochastic agent based model called, the Discrete Spatial Phylo Simulator (DSPS). DSPS was also used in [21]. The simulated DNA sequences were for a structured population which was composed of homosexual men (MSM), heterosexuals and bisexual men. The homosexual men were further split into LowMSM, MedMSM and HighMSM that were of increasing risk behaviour. Heterosexuals consisted of Low Females, High Females, Low Males and High Males. Bisexual men contacted both heterosexuals and homosexual men. During the simulations, half of bisexual men were diagnosed as if they were homosexual men (this was more likely early in infection) and the other half like heterosexual men (this was more likely late in infection). It was assumed that all bisexual men contacted men and women at the same rate on average. This group was not sub-divided further. The sampling was at 100% for 10 years (25 - 35, equivalent to 1995-2005). The simulated DNA sequences were

from pol gene.

In order to make analysis faster, we down sampled the set of sequences which we used in the analysis. In this set, there were 21532 sequences and we sampled 200 sequences uniformly at random. We further reduced the sub-groups to homosexuals, heterosexuals and bisexuals. Out of 200 sequences, 157 were for homosexuals, 38 for bisexuals and only 5 for heterosexuals. Since we were interested in MTBD-2 model, we discarded 5 sequences for heterosexuals. This resulted in analysis of 195 sequences in our set of simulated HIV sequences.

2.6. BEAST Study of Simulated DNA Sequences

We made a Bayesian study for a sampled set of sequences using Bayesian Evolution Analysis by Sampling Trees (BEAST) developed by [22]. This was done using the multi-type birth-death model package version 0.2.0 of BEAST. The inferred parameters included reproductive number for each deme, sampling proportions, becoming non-infectious rate, among others. For this study, we used the HKY substitution model with 4 gamma categories and estimated base frequencies. We set 0.867 as a proportion of invariant sites. Since our alignment contained sequences sampled at different times and the times were measured in years, we used a strict clock model with starting value set to 0.005. We set sampling change time to 10.0 which was slightly greater than time for the oldest sample. The sampling proportions were set to 0 before the time of the first sample.

We used the mean estimates for the parameters obtained from this Bayesian study to simulate trees with 200 sampled tips. The death rate parameters used in the simulation were equivalent to becoming non-infectious parameters from the Bayesian study. The birth-rate parameters were not inferred directly from this study. The birth-rate parameters were however computed from the reproductive numbers for each deme. This was possible since the BEAST study inferred the reproductive numbers of each deme. Since the reproductive number of each was

obtained using $R_{0i} = \frac{\sum_{j=1}^2 \lambda_{i,j}}{\mu_i}$, and yet we obtained R_{0i} and μ_i from the

BEAST study, we therefore set $\sum_{j=1}^2 \lambda_{i,j} = \mu_i R_{0i}$ during the simulations with the

mean parameter estimates from the BEAST study of the simulated sequences. From the simulated trees, we estimated the parameters to evaluate the performance of our model with parameters that were obtained from the simulated DNA sequences.

2.7. Tree Balance versus Structured Host Population

We investigated whether there was a relationship between some tree statistics and the structuring of the host population. This was done using tree statistics that measure balance and imbalance of a phylogeny. In [23], they state that

measuring the degree of imbalance or asymmetry of a tree topology may provide support for the hypothesis that species have different potential for speciation. It also serves as an indication of the patterns of speciation events for organisms under study. We used three tree statistics that measure tree balance and imbalance. We employed gamma-statistic, Sackin and Colles indices.

Gamma (γ) statistic was defined in [24]. Let $g_2, g_3, \dots, g_n$ be the inter-node distances of the reconstructed phylogeny with n taxa, the γ -statistic is defined as:

$$\gamma = \frac{\left(\frac{1}{n-2} \sum_{i=2}^{n-1} \left(\sum_{k=2}^i kg_k \right) \right) - \left(\frac{T}{2} \right)}{T \sqrt{\frac{1}{12(n-2)}}} \quad (9)$$

where

$$T = \sum_{j=2}^n jg_j$$

We analysed whether the values for γ -statistic are affected by the structuring of a host population. We computed gamma-statistic for both structured and non-structured host population. In [25], it is stated that under pure birth process, γ -statistic values of complete reconstructed phylogenies follow a standard normal distribution. If gamma-statistic values are greater than zero, then the internal nodes are closer to the leaves than expected under a pure birth process. On the other hand, if gamma-statistic values are less than zero, then internal nodes are closer to the origin (root) than expected under a pure birth model.

Another tree statistic which we used is called Sackin index which adds the number of internal nodes between each leaf and the root. It is defined as:

$$I_s^n = \sum_{i=1}^n N_i \quad (10)$$

where n is the number of leaves of the tree and N_i is the number of internal nodes crossed in the path from leaf i to the root. It is well known in systematic biology that the expectation of I_s^n under Yule model is of order $2n \ln(n)$. The normalized Sackin index converges in distribution as the number of leaves, n grows to infinity. The normalized Sackin index is defined in [25] as $\hat{I}_s^n = \frac{I_s^n - E(I_s^n)}{n}$.

Colless index is another tree statistic that assesses tree imbalance. It is defined as:

$$I_c = \frac{2}{(n-1)(n-2)} \sum_{i=1}^{n-1} |T_R - T_L| \quad (11)$$

where i is an internal node, n is the total number of leaves. For each internal node, T_R is the number of terminal taxa subtended by the right hand branch and T_L is the number of terminal taxa subtended by left hand branch. For a normalised I_c , it ranges from 0 (perfect balance) to 1 (incomplete) balance.

Tree balance is an important consideration for phylogenies because the bal-

ance of the true phylogeny affects the accuracy of its estimates as stated in [26]. For Sackin and Colless indices, the higher the values for these two indices, the more unbalanced the trees are [26].

We simulated phylogenetic trees in R using TreeSim package of [17]. For a structured host population, we set our parameters as: $\lambda_{11}=17$, $\lambda_{12}=4$, $\lambda_{21}=6$, $\lambda_{22}=8$, $\mu_1=9$, $\mu_2=3$ and $\psi_1=\psi_2=0.6$. For a non-structured host population, the parameters were set to $\lambda_{11}=8$, $\lambda_{12}=8$, $\lambda_{21}=8$, $\lambda_{22}=8$, $\mu_1=3$, $\mu_2=3$ and $\psi_1=\psi_2=0.6$. Parameter λ_{ij} is the rate at which an individual in sub-population i infects an individual in sub-population j , μ_i is the rate at which an individual in sub-population i dies and ψ_i is the rate at which an individual in sub-population i is sampled, for $i, j = 1$ and 2 .

In the first set of simulations, we simulated 5000 trees with 200 sampled tips under both structured and non-structured host populations (A1 and A2). In the second set of simulations, we increased the number of sampled tips to 500 while other parameters remained the same as those for the first set of simulations under both structured and non-structured host populations (B1 and B2). In the third set of simulations, parameters were the same as those in first set of simulations but increased the number of sampled trees to 15,000 for both structured and non-structured populations (C1 and C2). Under each of the simulations, we computed gamma-statistic values for the simulated trees. Using an R package apTreeshape of [27], we computed Colless and Sackin indices for the simulated trees. We additionally computed their mean, median and 95% confidence intervals for the three tree statistics in each of the simulation sets.

3. Results

3.1. Parameter Inference

Figure 3 and **Figure 4** display the estimates for some of the selected parameters graphically. Parameter estimates for λ_{11} and μ_1 when $\psi_1=\psi_2=0.2$ and $r=0.2$ are depicted in **Figure 3**. **Figure 4** shows parameter estimates for λ_{22} and μ_2 when $\psi_1=\psi_2=0.05$ and $r=0.8$. Apart from μ_2 which is shown in lower part of **Figure 4**, other parameter estimates are uniformly scattered across their true values (solid red line).

For fixed sampling probabilities ($\psi_1=\psi_2=0.2$), parameter estimates for $r=0.8$, yielded slightly better estimates compared to estimates obtained for $r=0.2$ as shown in **Table 1**, upper and middle parts. There was an overall poor estimation for the second death parameter (μ_2). The lower part of **Table 1** shows parameter estimates when the number of simulated trees was increased from 50 to 100. Better estimates were realized when the number of simulated trees was increased. Parameter estimation for μ_2 greatly improved when the number of sampled trees was increased to 100. A close look at the middle and lower parts of **Table 1**, reveals that 95% confidence interval for μ_2 narrows by close to half of the interval for the same parameter when the number of simulated trees was increased from 50 to 100.

Figure 3. Selected parameter estimates. Red line is the true parameter value while the yellow and green lines represent the 95% upper and lower limits for the shown parameters, respectively. The two selected parameters were estimated when $\psi_1 = \psi_2 = 0.2$ and $r = 0.2$. The vertical axis shows parameter estimate values, while the horizontal axis shows the number of simulated trees.

Figure 4. Selected parameter estimates. Red line is the true parameter value while the yellow and green lines represent the 95% upper and lower limits for the shown parameters, respectively. The two selected parameters were estimated when $\psi_1 = \psi_2 = 0.05$ and $r = 0.8$. The vertical axis shows parameter estimate values, while the horizontal axis shows the number of simulated trees.

Having observed that increasing the number of simulated trees to 100 improved parameter estimates, we increased the number of sampled tips from 100 to 200 in the next set of simulations. This is shown in **Table 2**. On making comparisons between the lower part of **Table 1** and the upper part of **Table 2**, we note that increasing the number of sampled tips to 200 and keeping the number of simulated trees constant (100), registers a slight improvement in the parameter

Table 1. Maximum likelihood parameter estimates from 50 and 100 simulated trees with 100 sampled tips (sampling probability $\psi_1 = \psi_2 = 0.2$) under MTBD-2.

For $r = 0.2$ & 50 simulated trees	True value	Mean	Median	95% CI
$\lambda_{1,1}$	17.000	16.542	16.489	[11.007, 22.704]
$\lambda_{1,2}$	4.000	4.724	2.895	[0.000, 25.838]
$\lambda_{2,1}$	6.000	10.782	6.489	[0.000, 39.772]
$\lambda_{2,2}$	8.000	7.273	6.827	[0.000, 16.052]
μ_1	9.000	11.593	9.781	[0.000, 23.597]
μ_2	3.000	0.935	1.243	[-25.061, 20.470]
For $r = 0.8$ & 50 simulated trees				
$\lambda_{1,1}$	17.000	16.464	16.520	[11.207, 20.851]
$\lambda_{1,2}$	4.000	3.983	3.765	[0.000, 10.515]
$\lambda_{2,1}$	6.000	8.092	6.922	[0.000, 32.281]
$\lambda_{2,2}$	8.000	8.153	7.651	[0.000, 17.024]
μ_1	9.000	10.015	9.252	[0.000, 23.330]
μ_2	3.000	3.014	2.104	[-14.388, 33.727]
For $r = 0.8$ & 100 simulated trees				
$\lambda_{1,1}$	17.000	16.137	16.920	[5.484, 20.558]
$\lambda_{1,2}$	4.000	5.007	4.483	[0.000, 15.587]
$\lambda_{2,1}$	6.000	6.611	5.016	[0.000, 25.069]
$\lambda_{2,2}$	8.000	7.959	8.684	[0.000, 12.711]
μ_1	9.000	11.302	9.492	[3.754, 30.200]
μ_2	3.000	3.237	1.899	[-2.926, 16.476]

Table 2. Maximum likelihood parameter estimates from 100 simulated trees with 200 sampled tips (sampling probability $\psi_1 = \psi_2 = 0.2$) under MTBD-2.

Parameters for $r = 0.8$	True value	Mean	Median	95% CI
$\lambda_{1,1}$	17.000	16.438	16.898	[7.990, 19.241]
$\lambda_{1,2}$	4.000	4.407	3.598	[0.000, 15.162]
$\lambda_{2,1}$	6.000	6.508	6.108	[0.000, 18.591]
$\lambda_{2,2}$	8.000	7.693	8.484	[0.000, 11.829]
μ_1	9.000	10.983	9.038	[4.330, 26.605]
μ_2	3.000	2.985	2.058	[-2.162, 13.363]
For $r = 0.5$				
$\lambda_{1,1}$	17.000	16.869	16.905	[12.755, 20.473]
$\lambda_{1,2}$	4.000	3.766	3.510	[0.000, 10.938]
$\lambda_{2,1}$	6.000	7.258	6.429	[0.000, 21.580]
$\lambda_{2,2}$	8.000	7.959	8.107	[2.528, 11.630]
μ_1	9.000	10.094	9.433	[2.249, 23.875]
μ_2	3.000	2.335	2.395	[-11.925, 13.089]

estimation. The 95% confidence intervals for parameters in the upper part of **Table 2** are narrower than those in the lower part of **Table 1**. When we varied the probability of removal (r) from 0.8 to 0.5, there was no significant difference in the parameter estimates as seen in the lower part of **Table 2**.

To investigate whether sampling intensity had an effect on parameter estimates, we varied the sampling proportion from 0.2 to 0.05. Parameter estimates when the sampling intensity was fixed at $\psi_1 = \psi_2 = 0.05$ are shown in **Table 3**. Slightly wider intervals were obtained for the parameter estimates for this setting. Parameter μ_2 was poorly estimated.

When parameter inference was made with different sampling proportions, *i.e.*, for $\psi = 0.2$ and $\psi = 0.01$, there was no difference noted in parameter inference for the simulated trees, results not shown here. Parameter μ_2 was poorly estimated again.

For the error bars shown in **Figure 5** and **Figure 6**, the intervals for parameter estimates obtained when each simulated tree had 200 sampled tips were narrower than those when the simulated trees had 100 sampled tips. The error bars for two parameters ($\lambda_{1,1}$ and μ_2) are shown. The two selected parameters gave a representation for the rest of the parameters that were estimated.

3.2. Heterogeneous and Homogeneous Dynamics

To establish whether our model distinguished between heterogeneous population (MTBD-2) from homogeneous population (MTBD-1), we performed likelihood ratio tests. In the likelihood ratio inference, we used a decision rule of rejecting the null-hypothesis if the likelihood ratio (test statistic) is greater than

Table 3. Maximum likelihood parameter estimates from 50 simulated trees with 100 sampled tips (sampling probability $\psi_1 = \psi_2 = 0.05$) under MTBD-2.

Parameters for $r = 0.2$	True value	Mean	Median	95% CI
$\lambda_{1,1}$	17.000	16.347	15.770	[0.000, 29.305]
$\lambda_{1,2}$	4.000	4.351	2.190	[0.000, 14.730]
$\lambda_{2,1}$	6.000	12.800	6.120	[0.000, 41.502]
$\lambda_{2,2}$	8.000	7.482	5.842	[0.000, 25.795]
μ_1	9.000	11.718	9.402	[2.312, 34.556]
μ_2	3.000	3.865	0.886	[-18.917, 36.919]
For $r = 0.8$				
$\lambda_{1,1}$	17.000	16.788	16.200	[7.715, 22.555]
$\lambda_{1,2}$	4.000	4.499	3.623	[0.000, 16.305]
$\lambda_{2,1}$	6.000	12.477	8.373	[0.000, 36.119]
$\lambda_{2,2}$	8.000	5.452	5.410	[0.000, 12.472]
μ_1	9.000	13.710	11.637	[0.000, 43.433]
μ_2	3.000	-1.834	-1.238	[-18.265, 11.133]

Figure 5. Error bars for λ_{11} and μ_2 . In all the 4 sets of simulations, there were 100 sampled tips in each tree. Red line is the true parameter value.

Figure 6. Error bars for λ_{11} and μ_2 . In all the 3 sets of simulations, there were 200 sampled tips in each tree. Red line is the true parameter value.

χ^2_5 (9.236) for a significance level (α) of 0.1. We used χ^2_5 since in MTBD-2 inference, 7 parameters were estimated including the likelihood value of the tree, while 4 parameters were estimated for MTBD-1. This gave an average of 6 parameters during the inference and since we needed a theoretical value of χ^2_{n-1} , we used χ^2_5 for the inference. In the first scenario, 41 out of 50 trees rejected MTBD-1 model in favour of MTBD-2 using likelihood ratio tests at 0.9 level ($\alpha = 0.1$). All the 50 trees supported MTBD-1 while rejecting MTBD-2. Some of the results are shown in **Table 4** and **Table 5**.

Table 4. Maximum likelihood parameter estimates from 50 simulated trees with 100 sampled tips ($\psi_1 = \psi_2 = 0.2$ and $r = 0.2$). In A, both tree simulation and inference is performed under MTBD-2 while in B, tree simulation is under MTBD-2 and inference under MTBD-1.

(A)	True value	Mean	Median	95% CI
$\lambda_{1,1}$	17.000	16.586	17.067	[8.634, 23.338]
$\lambda_{1,2}$	4.000	4.285	3.076	[0.000, 12.518]
$\lambda_{2,1}$	6.000	10.645	7.853	[0.000, 33.991]
$\lambda_{2,2}$	8.000	6.939	6.774	[0.000, 15.088]
μ_1	9.000	12.285	9.990	[0.019, 28.269]
μ_2	3.000	0.519	-0.396	[-21.688, 18.662]
(B)				
$\lambda_{1,1} = \lambda_{2,1}$		35.774	33.637	[2.879, 61.085]
$\lambda_{2,2} = \lambda_{1,2}$		4.139	3.489	[1.741, 8.112]
$\mu_1 = \mu_2$		-0.969	-1.999	[-7.964, 9.120]

Table 5. Maximum likelihood parameter estimates from 50 simulated trees with 100 sampled tips ($\psi_1 = \psi_2 = 0.2$ and $r = 0.2$). In A, both tree simulation and inference is performed under MTBD-1 while in B, parameter estimation is made still under MTBD-1 but with some constraints.

(A)	True value	Mean	Median	95% CI
$\lambda_{1,1}$	15.000	14.707	14.514	[9.530, 20.433]
$\lambda_{1,2}$	7.500	5.625	5.871	[0.000, 15.468]
$\lambda_{2,1}$	15.000	24.068	18.018	[2.885, 81.330]
$\lambda_{2,2}$	7.500	6.973	6.460	[0.000, 20.342]
μ_1	6.000	6.515	3.488	[0.000, 30.117]
μ_2	6.000	9.578	4.414	[-35.129, 96.456]
(B)				
$\lambda_{1,1} = \lambda_{2,1}$	15.000	8.079	6.877	[0.873, 19.679]
$\lambda_{2,2} = \lambda_{1,2}$	7.500	17.461	16.330	[3.898, 35.268]
$\mu_1 = \mu_2$	6.000	5.303	2.467	[-12.899, 45.491]

However, according to [28], he suggests using a χ^2 with 1 degree of freedom. So at α of 0.1, we reject the null hypothesis if the likelihood ratio is greater than 2.706. With this rule, instead of 41 out of 50, it is 49 trees that rejected MTBD-1 model in favour of MTBD-2 in first scenario and results for the second scenario remain the same when we use the idea of [28].

3.3. Results from BEAST Study of Simulated DNA Sequences

We analysed 195 sequences from the sample. These sequences were structured into two groups, namely MSM and Bisexuals. For this sample, we run MCMC in

BEAST which was of 12,000,000 length. The mixing was not very good as some parameters had effective sample sizes (ESS) which were low although some had high ESS of above 400. Our interest was mainly on three parameters, namely R_0 , becoming non-infectious rate and sampling proportion.

Figure 7 shows one of the trees sampled in the MCMC runs from BEAST. We also simulated 100 trees with 200 tips using the mean values of the parameters obtained from the analysis of the simulated sequences. These results are shown in **Table 6**.

3.4. Tree Balance versus Structured Host Population

From results shown in **Table 7**, generally the gamma-statistic values for simulated trees between structured and non-structured populations were all negative values apart from upper limits of the confidence intervals. The mean values for gamma-statistic values were generally more negative under non-structured population when we considered simulation cases A1, A2, B1 and B2. This was not true for simulation sets C1 and C2. When the number of tips was increased from 200 to 500, the mean and median values for gamma-statistic became more

Figure 7. One of the sampled trees from BEAST for structured host population. Red and blue represent bisexual and homosexual group of individuals, respectively.

Table 6. Maximum likelihood parameter estimates from 100 simulated trees with 200 sampled tips using median estimates from for the simulated sequences (sampling probability $\psi_1 = \psi_2 = 0.05$).

Parameters for $r = 0.5$	True value	Mean	Median	95% CI
$\lambda_{1,1}$	0.100	0.374	0.134	[0.000, 1.813]
$\lambda_{1,2}$	0.140	0.178	0.147	[0.000, 0.759]
$\lambda_{2,1}$	1.280	1.306	1.315	[0.991, 1.613]
$\lambda_{2,2}$	0.800	0.746	0.627	[0.035, 2.274]
μ_1	0.215	0.325	0.264	[0.000, 1.009]
μ_2	0.215	0.175	0.110	[-0.950, 1.354]

Table 7. Tree statistics under structured and non-structured populations.

A1: 5000 simulated trees with 200 sampled tips in a structured population			
Tree statistic	Mean	Median	95% CI
Gamma-statistic	−8.0765	−35.0077	[−95.1967, 36.576]
Colless' index	1848.324	1763	[1111, 3090]
Sackin's index	2873.392	2785	[2191, 4062]
A2: 5000 simulated trees with 200 sampled tips in a non-structured population			
Tree statistic	Mean	Median	95% CI
Gamma-statistic	−25.4391	−34.6674	[−125.8019, 55.931]
Colless' index	1558.561	1490	[947, 2532]
Sackin's index	2609.295	2539	[2050, 3533]
B1: 5000 simulated trees with 500 sampled tips in a structured population			
Tree statistic	Mean	Median	95% CI
Gamma-statistic	−52.1761	−51.7531	[−109.566, 4.93]
Colless' index	5964.97	5714	[3821, 9482]
Sackin's index	8974.741	8734	[6955, 12375]
B2: 5000 simulated trees with 500 sampled tips in a non-structured population			
Tree statistic	Mean	Median	95% CI
Gamma-statistic	−103.5037	−52.006	[−154.277, 36.237]
Colless' index	4934.575	4747	[3247, 7631]
Sackin's index	8016.174	7830	[6441, 10616]
C1: 15,000 simulated trees with 200 sampled tips in a structured population			
Tree statistic	Mean	Median	95% CI
Gamma-statistic	−44.139	−35.071	[−105.172, 40.067]
Colless' index	1854	1764	[1092, 3121]
Sackin's index	2878.715	2788	[2173, 4090]
C2: 15,000 simulated trees with 200 sampled tips in a non-structured population			
Tree statistic	Mean	Median	95% CI
Gamma-statistic	−25.312	−34.754	[−124.382, 71.212]
Colless' index	1553.154	1484	[948, 2539]
Sackin's index	2603.978	2534	[2052, 3549]

negative. However, the confidence intervals became narrower. When the sampled trees were increased from 5000 to 15,000 trees, while kept number of sampled tips to 200, the mean values for gamma-statistic became more negative under structured population though remained almost constant under non structured population. The median values remained almost constant.

For Colless index, the mean and median values were higher under structured population compared to non-structured population. Increasing the sampled tips from 200 to 500 resulted in higher mean and median Colless index values. In-

creasing simulated trees from 5000 to 15,000 resulted in no big difference in the mean and median values. Sackin index values have the same pattern as the Colless index values.

The densities for the three tree statistics are shown in **Figure 8**. From these densities, there was no significant difference between structured and non-structured populations. The only difference is in the densities for Colless index as the median value for a structured population is higher than that for a non-structured one.

Figure 8. Densities for three tree statistics for 15,000 simulated trees with 200 sampled tips.

4. Discussion and Conclusion

From the results shown in **Tables 1-6** and **Figures 3-6**, parameter estimation was better when sampling probability was fixed at 0.2 compared to when it was set at 0.05 to depict sparse sampling. It was also observed that although parameter estimation for sparse sampling ($\psi_1 = \psi_2 = 0.05$) resulted in wider parameter intervals, parameters were fairly estimated. The results for sparse sampling being less accurate may suggest that we need to be mindful when making inference in situations with sparse sampling of DNA sequences. It was also observed that the second death parameter (μ_2) was poorly estimated. It is only the second death parameter (μ_2) that is worst estimated with the lower limit for the intervals being negative. The first death rate parameter (μ_1) is fairly estimated with all estimates being positive. The mean and median estimates for μ_2 were however fairly estimated. Poor extinction rate parameter estimates may be due to the fact that generally the extinction (death) rate parameters are poorly estimated by general constant birth-death models [29] [30] [31]. When the probability of re-

removal (r) was varied from 0.2 to 0.8, narrower 95% confidence intervals for parameter estimates were realized. We defined r as the probability that an individual becomes non-infectious upon sampling (removed from infectious pool). For $r = 1$, our model reduces to that of [14] which the authors employed for their parameter inference. When $r = 0$, it implies that an individual i remains infectious upon sampling with the same rate ($\lambda_{i,j}$). Our model resulted in less accurate results when $r = 0.2$ simply because it had an extra parameter that affected the identifiability of the parameters. The estimates were however better when $r = 0.8$ because as r approached 1, the model approached that used in [14] and it was a parameter less compared to the model used in our study. Our model is handy in making inference that depicts an African setting which is characterized by both sparse sampling of DNA sequences and a large proportion of individuals that always remain infectious even after diagnosis with HIV.

Our proposed model distinguished between heterogeneous and homogeneous populations. This is because 41 trees out of 50 supported the heterogeneous model dynamics, rejecting the homogeneous type of model inference. For homogeneous dynamics, all the 50 trees accepted the homogeneous dynamics and rejected the heterogeneous type dynamics. We employed the likelihood test statistic for this model selection using a level of significance (α) of 0.1. This was close to that used in [14] which was 0.2. In their work, a χ^2 distribution at 0.8 level was used for the inference. It should be noted that even after incorporating a new parameter (probability of removal), the model still accounted for a structured host population. The probability of removal was included to reflect a situation where individuals fail to change behaviour upon diagnosis with HIV. Whereas studies from Europe have reported that individuals become non-infectious upon diagnosis with HIV as reported in [14], it is still a challenge in many African communities. It is therefore imperative to emphasize that structured populations characterize HIV epidemics in many African countries due to the fact that there are some groups who are more at risk compared to the general population in many African countries.

For the investigation between structuring of a population and tree balance, it was concluded from the results shown in **Table 7** that simulated trees from a non-structured host population are more balanced compared to those from a structured host population. This conclusion was based on Colless and Sackin indices. It is not surprising that when we increased the number of simulated trees from 5000 to 15,000, we did not realise a significant change. This is because the definitions of Colless and Sackin indices use internal nodes of a tree. This is the reason we observed a change in the value of indices when we increased the number of sampled tips from 200 to 500. When tips of a tree are increased, the internal nodes are increased as well. For a structured population, we had only 2 demes, it might be interesting to analyse tree balance in a structured population with more than 2 demes. The relationship between timing of bifurcation events and tree structure was assessed using the gamma-statistic values. From densities

shown in **Figure 8**, those for gamma-statistic are identical regarding to whether the underlying population is structured or not. A general conclusion from **Table 7** is that the gamma-statistic values could not be used to uncover the underlying population structure. This could be because the timing of the bifurcation events may be having very less influence on the overall tree shape of a phylogeny.

Acknowledgements

We thank the Editor and reviewers for their comments which greatly improved this paper. The authors thank Pan-African University Institute of Basic Sciences, Technology and Innovation (PAUSTI) for funding this research.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

- [1] UNAIDS (2017) Unaid Data. Technical Report, Joint United Nations Programme on HIV/AIDS.
- [2] UNDESA (2017) United Nations Department of Economic and Social Affairs, Population Division. World Population Prospects: The 2017 Revision, Key Findings and Advance Tables. Technical Report, United Nations, New York.
- [3] UNAIDS (2011) Unaid Terminology Guidelines. Technical Report, Joint United Nations Programme on HIV/AIDS.
- [4] UAC (2016) Uganda HIV and Aids Country Progress Report. Technical Report, Uganda AIDS Commission.
- [5] UAC (2014) Uganda HIV and Aids Country Progress Report. Technical Report, Uganda AIDS Commission.
- [6] Kiwanuka, N., Ssetaala, A., Nalutaaya, A., Mpendo, J., Wambuzi, M., Nanvubya, A., Sigirenda, S., Kitandwe, P.K., Nielsen, L.E. and Balyegisawa, A. (2014) High Incidence of HIV-1 Infection in a General Population of Fishing Communities around Lake Victoria, Uganda. *PLoS ONE*, **9**, e94932. <https://doi.org/10.1371/journal.pone.0094932>
- [7] Kribs-Zaleta, C.M., Lee, M., Rom an, C., Wiley, S. and Hernandez-Suarez, C.M. (2005) The Effect of the HIV/AIDS Epidemic on Africa's Truck Drivers. *Mathematical Biosciences and Engineering*. *MBE*, **2**, 771-788. <https://doi.org/10.3934/mbe.2005.2.771>
- [8] Gouws, E. and Cuchi, P. (2012) Focusing the HIV Response through Estimating the Major Modes of HIV Transmission: A Multi-Country Analysis. *Sexually Transmitted Infections*, **88**, i76-i85. <https://doi.org/10.1136/sextrans-2012-050719>
- [9] Bbosa, N., Ssemwanga, D., Nsubuga, R.N., Salazar-Gonzalez, J.F., Salazar, M.G., Nanyonjo, M., Kuteesa, M., Seeley, J., Kiwanuka, N., Bagaya, B.S., *et al.* (2019) Phylogeography of HIV-1 Suggests that Ugandan Fishing Communities Are a Sink for, Not a Source of, Virus from General Populations. *Scientific Reports*, **9**, 1051. <https://doi.org/10.1038/s41598-018-37458-x>
- [10] UPHIA (2017) Uganda Population Based HIV Impact Assessment. Technical Report, Government of Uganda.

- [11] Pillay, D., Herbeck, J., Cohen, M.S., Tulio de Oliveira, Fraser, C., Ratmann, O., Leigh Brown, A. and Kellam, P. (2015) The Pangea-HIV Consortium: Phylogenetics and Networks for Generalised HIV Epidemics in Africa. *The Lancet Infectious Diseases*, **15**, 259. [https://doi.org/10.1016/S1473-3099\(15\)70036-8](https://doi.org/10.1016/S1473-3099(15)70036-8)
- [12] Stadler, T., Kouyos, R., von Viktor, W., Yerly, S., Boni, J., Burgisser, P., Klimkait, T., Joos, B., Rieder, P., Xie, D., *et al.* (2011) Estimating the Basic Reproductive Number from Viral Sequence Data. *Molecular Biology and Evolution*, **29**, 347-357. <https://doi.org/10.1093/molbev/msr217>
- [13] Kuhnert, D., Stadler, T., Vaughan, T.G. and Drummond, A.J. (2014) Simultaneous Reconstruction of Evolutionary History and Epidemiological Dynamics from Viral Sequences with the Birth-Death Sir Model. *Journal of the Royal Society Interface*, **11**, Article ID: 20131106. <https://doi.org/10.1098/rsif.2013.1106>
- [14] Stadler, T. and Bonhoeffer, S. (2013) Uncovering Epidemiological Dynamics in Heterogeneous Host Populations Using Phylogenetic Methods. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **368**, Article ID: 20120198. <https://doi.org/10.1098/rstb.2012.0198>
- [15] Gavryushkina, A., Welch, D., Stadler, T. and Drummond, A.J. (2014) Bayesian Inference of Sampled Ancestor Trees for Epidemiology and Fossil Calibration. *PLoS Computational Biology*, **10**, e1003919. <https://doi.org/10.1371/journal.pcbi.1003919>
- [16] Stadler, T. (2012) How Can We Improve Accuracy of Macroevolutionary Rate Estimates? *Systematic Biology*, **62**, 321-329. <https://doi.org/10.1093/sysbio/sys073>
- [17] Stadler, T. (2014) Treesim: Simulating Trees under the Birth-Death Model. R Package Version 2.0.
- [18] Stadler, T. and Maintainer Stadler, T. (2015) Estimating Birth and Death Rates Based on Phylogenies. R Package Version 3.3.
- [19] Gascuel, O. (2005) Mathematics of Evolution and Phylogeny. OUP, Oxford.
- [20] Ota, R., Waddell, P.J., Hasegawa, M., Shimodaira, H. and Kishino, H. (2000) Appropriate Likelihood Ratio Tests and Marginal Distributions for Evolutionary Tree Models with Constraints on Parameters. *Molecular Biology and Evolution*, **17**, 798-803. <https://doi.org/10.1093/oxfordjournals.molbev.a026358>
- [21] Yebra, G., Hodcroft, E.B., Ragonnet-Cronin, M.L., Pillay, D., Leigh Brown, A.J., Fraser, C., Kellam, P., De Oliveira, T., Dennis, A., Hoppe, A., *et al.* (2016) Using Nearly Full-Genome HIV Sequence Data Improves Phylogeny Reconstruction in a Simulated Epidemic. *Scientific Reports*, **6**, 39489. <https://doi.org/10.1038/srep39489>
- [22] Drummond, A.J. and Rambaut, A. (2007) Beast: Bayesian Evolutionary Analysis by Sampling Trees. *BMC Evolutionary Biology*, **7**, 214. <https://doi.org/10.1186/1471-2148-7-214>
- [23] Blum, M.G.B. and Francois, O. (2005) On Statistical Tests of Phylogenetic Tree Imbalance: The Sackin and Other Indices Revisited. *Mathematical Biosciences*, **195**, 141-153. <https://doi.org/10.1016/j.mbs.2005.03.003>
- [24] Pybus, O.G. and Harvey, P.H. (2000) Testing Macro-Evolutionary Models Using Incomplete Molecular Phylogenies. *Proceedings of the Royal Society of London B: Biological Sciences*, **267**, 2267-2272. <https://doi.org/10.1098/rspb.2000.1278>
- [25] Blum, M.G.B., Francois, O. and Janson, S. (2006) The Mean, Variance and Limiting Distribution of Two Statistics Sensitive to Phylogenetic Tree Balance. *The Annals of Applied Probability*, **16**, 2195-2214. <https://doi.org/10.1214/105051606000000547>
- [26] Shao, K.T. (1990) Tree Balance. *Systematic Zoology*, **39**, 266-276.

- <https://doi.org/10.2307/2992186>
- [27] Bortolussi, N., Durand, E., Blum, M. and Francois, O. (2005) Aptreeshape: Statistical Analysis of Phylogenetic Tree Shape. *Bioinformatics*, **22**, 363-364.
<https://doi.org/10.1093/bioinformatics/bti798>
- [28] Maddison, W.P., Midford, P.E. and Otto, S.P. (2007) Estimating a Binary Character's Effect on Speciation and Extinction. *Systematic Biology*, **56**, 701-710.
<https://doi.org/10.1080/10635150701607033>
- [29] Rabosky, D.L. (2010). Extinction Rates Should Not Be Estimated from Molecular Phylogenies. *Evolution: International Journal of Organic Evolution*, **64**, 1816-1824.
<https://doi.org/10.1111/j.1558-5646.2009.00926.x>
- [30] Beaulieu, J.M. and O'Meara, B.C. (2015) Extinction Can Be Estimated from Moderately Sized Molecular Phylogenies. *Evolution*, **69**, 1036-1043.
<https://doi.org/10.1111/evo.12614>
- [31] Rabosky, D.L. (2016) Challenges in the Estimation of Extinction from Molecular Phylogenies: A Response to Beaulieu and O'Meara. *Evolution*, **70**, 218-228.
<https://doi.org/10.1111/evo.12820>