Virus Evolution, 2018, 4(2): vey037

doi: 10.1093/ve/vey037
Research article

Tracking external introductions of HIV using phylodynamics reveals a major source of infections in rural KwaZulu-Natal, South Africa

David A. Rasmussen[1,2,*,†], Eduan Wilkinson[3], Alain Vandormael[3,4], Frank Tanser[4,5,6], Deenan Pillay[5,7], Tanja Stadler[8,9], and Tulio de Oliveira[3,10,11]

[1]Department of Entomology and Plant Pathology, North Carolina State University, Raleigh, NC, USA, [2]Bioinformatics Research Center, North Carolina State University, Raleigh, NC, USA, [3]KwaZulu-Natal Research Innovation and Sequencing Platform (KRISP), College of Health Sciences, University of KwaZulu-Natal, Durban, South Africa, [4]School of Nursing and Public Health, University of KwaZulu-Natal, Durban, South Africa, [5]Africa Health Research Institute, Durban, South Africa, [6]Research Department of Infection & Population Health, University College London, UK, [7]Division of Infection and Immunity, University College London, UK, [8]Department of Biosystems Science and Engineering, ETH Zürich, Basel, Switzerland, [9]Swiss Institute of Bioinformatics, Lausanne, Switzerland, [10]Centre for the AIDS Programme of Research in South Africa (CAPRISA), Durban, South Africa and [11]Department of Global Health, University of Washington, Seattle, USA

*Corresponding author: E-mail: drasmus@ncsu.edu

†http://orcid.org/0000-0001-9457-7561

Abstract

Despite increasing access to antiretrovirals, HIV incidence in rural KwaZulu-Natal remains among the highest ever reported in Africa. While many epidemiological factors have been invoked to explain such high incidence, widespread human mobility and viral movement suggest that transmission between communities may be a major source of new infections. High cross-community transmission rates call into question how effective increasing the coverage of antiretroviral therapy locally will be at preventing new infections, especially if many new cases arise from external introductions. To help address this question, we use a phylodynamic model to reconstruct epidemic dynamics and estimate the relative contribution of local transmission versus external introductions to overall incidence in KwaZulu-Natal from HIV-1 phylogenies. By comparing our results with population-based surveillance data, we show that we can reliably estimate incidence from viral phylogenies once viral movement in and out of the local population is accounted for. Our analysis reveals that early epidemic dynamics were largely driven by external introductions. More recently, we estimate that 35 per cent (95% confidence interval: 20–60%) of new infections arise from external introductions. These results highlight the growing need to consider larger-scale regional transmission dynamics when designing and testing prevention strategies.

Key words: HIV; molecular epidemiology; phylodynamics; migration

1. Introduction

While the HIV epidemic hit South Africa relatively late compared with other southern African nations, the epidemic grew explosively in the 1990s from an estimated 0.8 per cent prevalence in 1990 to over 20 per cent in 2000 (UNAIDS 2017). Prevalence nationwide stabilized at around 20 per cent in 2000, but an estimated 7.1 million people are still currently living with HIV in South Africa, more than any other country in the world. Prevalence is highest in the province of KwaZulu-Natal (KZN), where 25–40 per cent of the population is HIV positive in some communities (Zaidi et al. 2013; Kharsany et al. 2015; Shisana et al. 2015). While the increasing lifespan of infected individuals on antiretroviral treatment (ART) can partly explain why prevalence has remained high, incidence also remains alarmingly high at between 3 and 6 per cent per year in KZN (Karim et al. 2011; Nel et al. 2012; Vandormael et al. 2018) and may be even higher in certain high-risk groups such as young women (de Oliveira et al. 2017).

Many factors have been implicated in the explosive growth of the HIV epidemic in southern Africa and KZN in particular, but patterns of human movement in the region have long received special attention (Jochelson et al. 1991; Quinn 1994; Lurie et al. 1997). Migration rates have historically been high in the region, fueled by labor-intensive industries such as mining (Hargrove 2008; Corno and De Walque 2012). Mobility has also increased significantly since the end of Apartheid, and sexual networks are known to be geographically well-connected across long distances (Harrison et al. 2008). But the exact role human movement has played in the dynamics of the epidemic has been debated (Coffee et al. 2007; Deane et al. 2010). Given the rural and rather isolated nature of many KZN communities, it is apparent that some spatial mixing would have been necessary to seed the epidemic in different geographic locations. Beyond seeding local epidemics, increasing mobility and migrant labor may have also played a larger role by placing individuals at an increased risk of infection (Lurie et al. 2003; Welz et al. 2007; Camlin et al. 2010), possibly due to migrants engaging in riskier sexual behavior outside of their home communities (Coffee et al. 2007).

Resolving the contribution of human movement to the HIV epidemic has been challenging due to the difficulty of quantifying the extent to which transmission occurs between individuals within local communities versus new cases being imported through external introductions. While the geographic source of new infections cannot typically be resolved using traditional surveillance data, viral phylogenies can help reveal the source of new infections (Holmes 2004; Pybus and Rambaut 2009; Faria et al. 2011). Broadly speaking, if new infections primarily arise from transmission within local communities, viral samples collected within the community should be more closely phylogenetically related to one another than to samples taken from outside the community; whereas if many new infections are being externally introduced, then local samples will tend to cluster with external samples throughout the tree. Phylogenetics can therefore help reveal the movement of viral lineages within and between different communities. For example, one recent study revealed that a high proportion (~40%) of new infections in a rural Ugandan community arose from external introductions (Grabowski et al. 2014). Recent phylogenetic studies have also revealed widespread viral movement across larger spatial scales and even across national borders (Gray et al. 2009; Wilkinson et al. 2016).

Here, we explore the role of external introductions in the Africa Health Research Institute's study population in rural KwaZulu-Natal, where HIV prevalence is ~30 per cent. Using a phylodynamic approach that couples phylogenetic methods together with epidemiological modeling, we reconstructed epidemic dynamics consistent with estimates from population-based surveillance data. By tracking the movement of viral lineages between populations, we also directly quantified the contribution of local transmission versus external introductions to overall HIV incidence. We found that far from just seeding the local epidemic, external introductions played a large role in sustaining high HIV incidence, thus confirming the important role human mobility and migration have played in the hyper-epidemic setting of KZN.

2. Methods

2.1 Study population

We focused on the Africa Health Research Institute (AHRI) study population in KwaZulu-Natal, South Africa. The AHRI demographic surveillance area (DSA) is located 200 km north of Durban with a mostly rural or peri-urban population (Fig. 1). Demographic data were compiled from the Africa Centre Demographic Information System between 2000 and 2015 (Tanser et al. 2008). There were approximately 87,000 people under surveillance in this population at any given time, with an adult population of males aged 15–54 years and females aged 15–49 years of 55,000. Prevalence and incidence data were also made available from ongoing surveillance.

2.2 Population-based HIV surveillance

A population-based HIV survey has been performed within the DSA since 2004, and we compared our phylodynamic estimates of prevalence and incidence against data from this survey. Trained field-workers visit households every 12 months and identify eligible participants aged 15–49 years. After obtaining consent, the field workers then extract blood according to the UNAIDS and WHO Guidelines for Using HIV Testing Technologies in Surveillance. Of the eligible participants contacted, approximately 80 per cent agree to be tested at least once during the study period. Prevalence was determined from the number of individuals who tested HIV-positive during these yearly visits (Vandormael et al. 2018).

We calculated incidence rates using the surveillance data from the subset of eligible participants who had two or more HIV tests, of which the first was a valid HIV-negative result (Vandormael et al. 2014). During the 2004–14 surveillance period, we observed 2,557 seroconversion events for the 17,417 participants in the incidence cohort, irrespective of their residency status or time spent in the DSA. Due to periodic HIV testing, however, we did not know the exact dates of seroconversion events. We therefore used an Monte Carlo-based approach to impute a random seroconversion event between the ith participant's latest-negative and earliest-positive test dates (Vandormael et al. 2018).

2.3 Sequence data and phylogenetic analysis

Partial HIV-1 polymerase (*pol*) sequences were collected as part of population-based surveillance conducted in 2011 and 2014. HIV-1 viral load tests were done on all dried blood spot samples that tested positive by serology (SD Bioline ELISA). Only samples from ART-naive participants with viral loads greater than 10,000 RNA copies/ml were genotyped. More information about

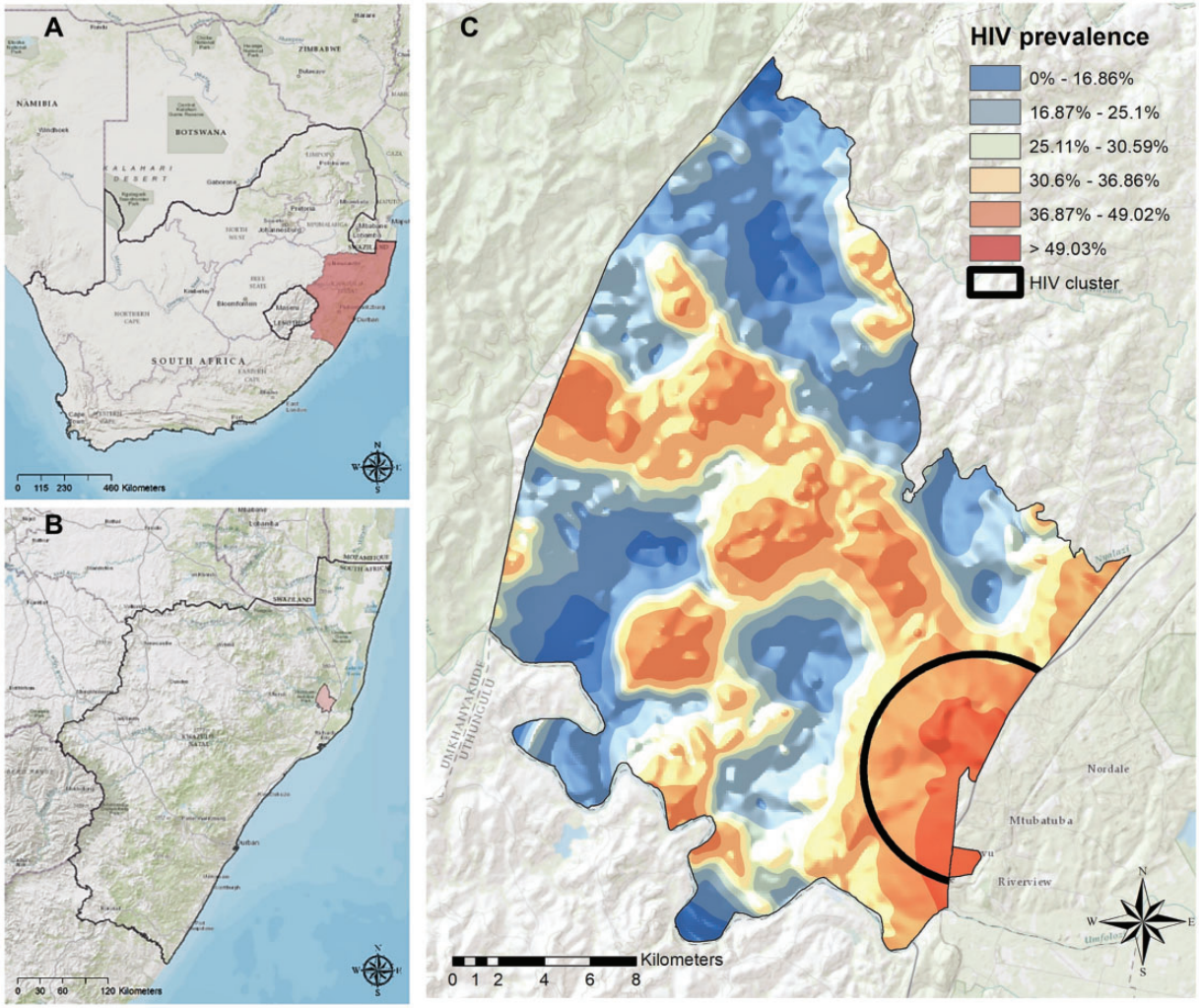

Figure 1. Location of the AHRI study population and prevalence within the area. (A and B) The location of KwaZulu-Natal province and the study area within the province. (C) The spatial variability of HIV prevalence within the study area based on population-based surveillance.

the genotyping process is described by Manasa et al. (2016). In total, 1,068 sequence samples from the DSA were included in our analysis. In addition to these samples, we used a large dataset containing 11,289 unique subtype C *pol* sequences described previously by Wilkinson et al. (2016) as a regional background dataset to help identify external introductions. This dataset contained other sequences from South Africa (n = 7,739) as well as sequences sampled between 1989 and 2014 from Angola (n = 9), Botswana (n = 863), the DRC (n = 25), Malawi (n = 352), Mozambique (n = 342), Swaziland (n = 47), Tanzania (n = 168), Zambia (n = 1,476), and Zimbabwe (n = 268).

Maximum likelihood (ML) phylogenetic trees were reconstructed using FastTree2 (Price et al., 2010). The ML trees were then dated using Least Squares Dating (To et al. 2015), so that branch lengths were given in units of real calendar time. For dating, we assumed a molecular clock rate of 2.0×10^{-3}, which falls in the center of previously estimated clock rates for subtype C (Wilkinson et al. 2016).

In a preliminary analysis, ML trees were reconstructed from an alignment containing all sequences in the background dataset together with all samples from the AHRI. To identify potential external introductions into the local population, we reconstructed the ancestral location of each internal node in the ML trees using the Fitch parsimony algorithm (Sankoff 1975). External introductions were assumed to occur whenever a child node reconstructed to be in the local population had a parent node reconstructed to be in the external population. The midpoint time between the parent and child node was then used as a proxy for the probable time of introduction.

For the phylodynamic analysis, phylogenetic trees were reconstructed from all AHRI sequences and an equal number of sequences randomly sampled from the background dataset. This was done to reduce the computational cost of fitting the phylodynamic model. To take into account phylogenetic uncertainty and variability across sub-sampled datasets, the phylodynamic analysis was replicated on ten phylogenies each reconstructed from a different set of sequences sub-sampled from the full regional background dataset. Estimates provided in the Results represent an average over these ten phylogenies and sampling replicates (see MCMC details below).

2.4 Ethics statement

Ethics permission for the population-based HIV surveillance at the Africa Health Research Institute was obtained from the Biomedical Research Ethics Committee of the College of Health

Sciences, University of KwaZulu-Natal (Ethics Nos. BF233/09 and E134/06). All participants in the study provided written informed consent for the analysis of their samples. Sequence samples obtained from HIV positive individuals were obtained from this already-existing study.

2.5 Data availability

A representative subset of the HIV pol sequences have been deposited on GenBank with accession numbers MH920641–MH920852. This subset represents 212 of the 1068 total sequences used in this study. The full dataset has been deposited in a secure repository: https://doi.org/10.23664/AC_HIVpol_full1068. Researchers with a clearly demonstrated scientific need may request access to the full dataset by contacting AHRI's Chief Information Officer, Kobus Herbst at Kobus.Herbst@ahri.org.

2.6 Phylodynamic model

Our phylodynamic model divides the host population into a local and an external population. Within the local population, transmission dynamics are modeled using an SIR-type epidemiological model where the number of susceptible (S), infected (I) or removed (R) individuals change over time according to the differential equations:

$$\begin{aligned}\frac{dS_l}{dt} &= \mu N_l - \beta_{e\to l}(t)\frac{S_l}{N_l}I_e - \beta_{l\to l}(t)\frac{S_l}{N_l}I_l - \mu S_l \\ \frac{dI_l}{dt} &= \beta_{e\to l}(t)\frac{S_l}{N_l}I_e + \beta_{l\to l}(t)\frac{S_l}{N_l}I_l - (\nu+\mu)I_l \\ \frac{dR_l}{dt} &= \nu I_l - \mu R_l\end{aligned} \tag{1}$$

The subscripts denote whether a host resides in the local (l) or external (e) population. Two sources of transmission contribute to new infections within the local population. The first source, the $\beta_{e\to l}(t)\frac{S_l}{N_l}I_e$ term in (1), represents external transmissions into the local population from the external infected population and corresponds to the $\alpha(t)$ terms in Fig. 3. The second source, the $\beta_{l\to l}(t)\frac{S_l}{N_l}I_l$ term in (1), represents transmission within the local population and corresponds to the $\beta(t)$ terms in Fig. 3. All transmission rates $\beta(t)$ are allowed to vary over time to accommodate changes in risk behavior, treatment, or other non-modeled factors. The birth rate μ and removal rate ν are assumed to be constant over time. Upon removal, infected individuals are assumed to no longer be infectious or sexually active.

We assume that the external infected population size I_e grows logistically over time according to an SIS model. The number of susceptible (S_e) and infected (I_e) hosts in the external population change over time as follows:

$$\begin{aligned}\frac{dS_e}{dt} &= -\beta_{e\to e}\frac{S_e}{N_e}I_e + \nu I_e \\ \frac{dI_e}{dt} &= \beta_{e\to e}\frac{S_e}{N_e}I_e - \nu I_e\end{aligned} \tag{2}$$

We set the initial infected population size $I_e^{\text{init}} = 1$ in 1975. While this model ignores the considerable spatiotemporal complexity in HIV dynamics within southern Africa, it captures the main trend in HIV dynamics over time—rapidly increasing growth followed by stabilization in prevalence. We therefore view I_e as the effective number of infected individuals living in the region who could have transmitted to an individual living in the local population.

We assume that the time-dependent transmission rates, $\beta_{e\to l}(t)$ and $\beta_{l\to l}(t)$, are constant within a given time interval but can change between intervals in a piecewise-constant manner. To obtain a smoothed estimate of how these parameters change over time, we placed a Gaussian Markov Random Field prior on each parameter. Such GMRF models have been used previously in phylodynamics to prevent large fluctuations in parameter values between neighboring time intervals by penalizing against overly large changes (Minin et al. 2008). The prior probability on a given sequence of time-dependent parameters $\gamma_{1:T}$ is computed as:

$$Pr(\gamma_{1:T}|\tau) \propto \tau^{\frac{T-2}{2}} \exp\left[-\frac{\tau}{2}\sum_{t=2}^{T-1}\frac{(\gamma_{t+1}-\gamma_t)^2}{\Delta t}\right], \tag{3}$$

where τ is the precision parameter that controls the expected autocorrelation in parameter values between neighboring time intervals. The precision parameter τ was inferred separately for each time-dependent parameter. The time interval Δt between change points was fixed at two years between 1990 and 2006. Time-dependent parameters were assumed to be constant after 2006 to prevent overfitting during time periods when the phylogeny was relatively uninformative about changes in epidemic dynamics.

The structured coalescent framework of (Volz 2012) was used to compute the likelihood of the reconstructed phylogenies under our phylodynamic model, which has previously been used to analyze HIV transmission dynamics (Volz et al. 2013; Rasmussen et al. 2014b). The likelihood of a given phylogeny under a general structured coalescent model is:

$$\mathcal{L}(\mathcal{T}) = \prod_{m=1}^{M-1}\lambda_{ij}(t_m)\exp\left[-\int_{s=t_m}^{s=t_{m+1}}\sum_{i}^{i\in\mathcal{A}(s)}\sum_{j>i}^{j\in\mathcal{A}(s)}\lambda_{ij}(s)ds\right]. \tag{4}$$

For a tree containing M samples, the total likelihood is the product of the likelihood of each of the $M-1$ coalescent events and the waiting times between events. The likelihood depends on the pairwise coalescent rate $\lambda_{ij}(t)$ at which two lineages i and j coalesce at time t. The total coalescent rate is then computed by summing over all pairs of lineages in the set of lineages $\mathcal{A}(s)$ present in the phylogeny at time s, which is allowed to change within coalescent intervals due to sampling.

As shown in Volz (2012), under a structured epidemiological model with more than one type of infected individual, the pairwise coalescent rate is:

$$\lambda_{ij}(t) = \sum_{k}^{m}\sum_{l}^{m}\frac{f_{kl}}{I_k I_l}(p_{ik}p_{jl} + p_{il}p_{jk}), \tag{5}$$

where p_{ik} is the probability that lineage i is in state k and p_{jl} is the probability that lineage j is in state l. I_k and I_l are the number of infected individuals in populations k and l, respectively.

The lineage state probabilities can be computed using a system of differential equations that describe how the probability of lineage i being in state k evolves backwards in time based on its sampling location:

$$\frac{dp_{ik}}{dt} = \sum_{l}\left(\frac{f_{kl}}{I_l}p_{il} - \frac{f_{lk}}{I_k}p_{ik}\right) - p_{ik}\sum_{j\neq i}^{j\in\mathcal{A}(t)}\sum_{l}^{m}\frac{f_{kl}+f_{lk}}{I_k I_l}p_{jl}. \tag{6}$$

Here we have modified the equations originally introduced in Volz (2012) by adding the final term in (6), which takes into account how the relative probability of lineage i being in state k changes conditional on the observation that the lineage has not coalesced with any other lineage j in the phylogeny. This was shown by Müller et al. (2017) to improve parameter estimates under asymmetric population dynamics and sampling.

Under our two population HIV model, f_{kl} represents the transmission rate between populations k and l. In matrix notation, we have the transmission rate matrix:

$$F(t) = \begin{pmatrix} \beta_{e\to e}\frac{S_e}{N_e}I_e & \beta_{e\to l}(t)\frac{S_l}{N_l}I_e \\ \beta_{l\to e}I_l & \beta_{l\to l}(t)\frac{S_l}{N_l}I_l \end{pmatrix}. \quad (7)$$

Transmission from the local to the external population allows lineages in the local population to move back into the external population, but the rate $\beta_{l\to e}$ is assumed to be constant over time. Transmission in this direction does not affect the number of infected individuals in the external population, as this is assumed to be much larger than the number of cases exported from the local populations.

We implemented this phylodynamic model in BEAST 2 (Bouckaert et al. 2014) as an add-on package named Marula (Rasmussen 2017). Source code and input XML files that can be used to replicate our analysis are freely available at https://github.com/davidrasm/Marula. We first validated the implementation of our model on mock phylogenies simulated under a stochastic, individual-based version of our base epidemiological model. During these simulations, the full transmission tree was recorded and then subsequently subsampled to produce mock phylogenies. Mock phylogenies were simulated under two parameter regimes where either ~25 or ~2.5 per cent of all infections in the local population were due to external introductions, corresponding to the positive and negative controls in Fig. 3D and E, respectively. We verified that the true simulated prevalence, incidence and fraction attributable to external introductions fell within the estimated 95 per cent credible intervals in at least 95 per cent of all years across ten simulations under both parameter regimes.

The posterior distribution of all model parameters and epidemic dynamics were inferred using BEAST's built-in MCMC sampler. The marginal likelihood of the phylogeny under different models was then computed by taking the harmonic mean of the posterior probabilities sampled by the MCMC algorithm. Due to the large size of the phylogenies, replicate MCMC runs were performed on ten different fixed ML phylogenies reconstructed from different sub-sampled sequence datasets (see above) rather than jointly estimating the full phylogeny while simultaneously fitting the phylodynamic model. Samples from each MCMC replicate were then pooled to obtain posterior estimates averaged over the different ML phylogenies. Each MCMC replicate was run for at least one million iterations.

For inference, weakly informative priors were placed on all estimated parameters (Table 1). In addition, a few demographic parameters were assumed to be known from the AHRI DIS (Table 2). The local population size N_l reflects the adult population size of males 15–54 and females 15–49 years old. The population was assumed to be at demographic equilibrium with a constant birth/death rate of 2.8 per cent per capita each year.

2.7. Model variants

In addition to our base phylodynamic model, we consider four other model variants in order to check the robustness of our results to simplifying assumptions made in the base model. In each of these models, the epidemic dynamics in the external population were modeled using (2), but may include different infected classes in the local population.

2.7.1 *Hot spot model*

The first model variant considers subpopulation structure within the AHRI population due to the clustering of individuals into hot and cold spots of prevalence (Fig. 1C). Spatial scan statistics were used to identify areas with an excess number of HIV cases when compared against a null model assuming a random spatial distribution of cases (Tanser et al. 2009). All other areas were categorized as cold spots. On average, prevalence was >25 per cent in hot spots and <10 per cent in cold spots. The epidemic dynamics in the local population are described by the following system of differential equations:

$$\begin{aligned}
\frac{dS_h}{dt} &= \mu N_h - \left(\beta_{e\to h}(t)I_e + \beta_{h\to h}(t)I_h + \beta_{c\to h}(t)I_c\right)\frac{S_h}{N_h} - \mu S_h \\
\frac{dS_c}{dt} &= \mu N_c - \left(\beta_{e\to c}(t)I_e + \beta_{c\to c}(t)I_c + \beta_{h\to c}(t)I_h\right)\frac{S_c}{N_c} - \mu S_c \\
\frac{dI_h}{dt} &= \left(\beta_{e\to h}(t)I_e + \beta_{h\to h}(t)I_h + \beta_{c\to h}(t)I_c\right)\frac{S_h}{N_h} - (\nu+\mu)I_h \\
\frac{dI_c}{dt} &= \left(\beta_{e\to c}(t)I_e + \beta_{c\to c}(t)I_c + \beta_{h\to c}(t)I_h\right)\frac{S_c}{N_c} - (\nu+\mu)I_c \\
\frac{dR_h}{dt} &= \nu I_h - \mu R_h \\
\frac{dR_c}{dt} &= \nu I_c - \mu R_c
\end{aligned} \quad (8)$$

Here, the subscripts h and c indicate whether an individual resides in a hot or cold spot. Based on the geographic location of households within the DSA, approximately 40 per cent of the population lives in hot spots, corresponding to population sizes of about N_h = 22,000 and N_c = 33,000.

The same structured coalescent framework as used for the base model can be used to fit this model to the viral phylogenies, but we need to expand the transmission rate matrix F used to track the movement of lineages between populations to include hot and cold spots:

$$F(t) = \begin{pmatrix} \beta_{e\to e}\frac{S_e}{N_e}I_e & \beta_{e\to h}(t)\frac{S_h}{N_h}I_e & \beta_{e\to c}(t)\frac{S_c}{N_c}I_e \\ \beta_{h\to e}I_h & \beta_{h\to h}(t)\frac{S_h}{N_h}I_h & \beta_{h\to c}(t)\frac{S_c}{N_c}I_h \\ \beta_{c\to e}I_c & \beta_{c\to h}(t)\frac{S_h}{N_h}I_c & \beta_{c\to c}(t)\frac{S_c}{N_c}I_c \end{pmatrix}. \quad (9)$$

The location of lineages at the tips of phylogenies was assigned based on whether the sampled individual resided in a hot or cold spot.

Parameters and prior distributions for the hotspot model that differ from those used in the base model are given in Table 3. We estimated the time-varying transmission rates $\beta_{e\to c}(t)$ and $\beta_{c\to c}(t)$. The transmission rate from the external population into hot spots was parameterized in terms of a time-invariant scalar $\kappa_{e\to h}$ of the time-varying transmission rate from the external population into cold spots: $\beta_{e\to h}(t) = \kappa_{e\to h}\beta_{e\to c}(t)$. Likewise, the transmission rates within the local population were parameterized in terms of time-invariant scalars of the transmission rate within cold spots, for instance $\beta_{h\to h}(t) = \kappa_{h\to h}\beta_{c\to c}(t)$.

Table 1. Prior distributions on all estimated parameters.

Parameter	Name	Prior distribution	Prior values
βl → l	Local transmission rate	Log-normal	$\mu = -2.3026$; $\sigma = 1.0$
βe → e	External transmission rate	Log-normal	$\mu = -0.5065$; $\sigma = 1.0$
βe → l	External introduction rate	Log-normal	$\mu = -2.3026$; $\sigma = 1.0$
βl → e	Local export rate	Log-normal	$\mu = -2.3026$; $\sigma = 1.0$
Ne	External pop size	Uniform	min = 55,000 max = 250,000
ν	Removal rate	Log-normal	$\mu = -2.2769$; $\sigma = 0.1$
τ	GMRF precision	Gamma	$\alpha = 0.01$; $\beta = 0.01$

Table 2. Demographic parameters fixed at constant values.

Parameter	Name	Value
Nl	Local pop size	55,000
μ	Birth/death rate	2.8% per year

2.7.2 Time-varying removal model

The second model is very similar to the base model but allows for individuals to remain in the infected class longer as the epidemic progresses due to decreasing removal rates, which mimics improvements in clinical care or treatment over time. We allowed the duration of infection to gradually lengthen by shifting our prior on the removal rate ν to lower values over time (Table 4).

2.7.3 ART model

Our third model variant allows infected individuals in the local population to initiate ART after 2004, when widespread ART coverage began in South Africa. Under this model, the epidemic dynamics in the local population are described by the following system of differential equations:

$$\begin{aligned}
\frac{dS_l}{dt} &= \mu N_l - \left(\beta_{e\to l}(t)I_e + \beta_{l\to l}(t)I_l + \beta_{t\to l}(t)I_t\right)\frac{S_l}{N_l} - \mu S_l \\
\frac{dI_l}{dt} &= \left(\beta_{e\to l}(t)I_e + \beta_{l\to l}(t)I_l + \beta_{t\to l}(t)I_t\right)\frac{S_l}{N_l} - (\nu + \mu + \gamma_t)I_l \\
\frac{dI_t}{dt} &= \gamma_t I_l - (\nu_t + \mu)I_t \\
\frac{dR_l}{dt} &= \nu I_l + \nu_t I_t - \mu R_l.
\end{aligned} \tag{10}$$

Here, I_t represents the number of infected individuals on ART in the local population. The parameter γ_t controls the rate at which infected individuals initiate ART and ν_t the rate at which individuals on ART are removed from the infectious population, which is allowed to differ from the base removal rate ν.

The structured coalescent model corresponding to this epidemiological model has the transmission rate matrix:

$$F(t) = \begin{pmatrix} \beta_{e\to e}\frac{S_e}{N_e}I_e & \beta_{e\to l}(t)\frac{S_l}{N_l}I_e & 0 \\ \beta_{l\to e}I_l & \beta_{l\to l}(t)\frac{S_l}{N_l}I_l & 0 \\ \beta_{t\to e}I_t & \beta_{t\to l}(t)\frac{S_l}{N_l}I_t & 0 \end{pmatrix}. \tag{11}$$

The elements in F describe how lineages move between infected individuals in different populations through transmission events but does not take into account the movement of a lineage into the ART-infected compartment when a host initiates treatment. We therefore need to modify (6) above for tracking the lineage state probabilities to include the rates g_{kl} at which lineages transition between states k and l independent of transmission events:

$$\frac{dp_{ik}}{dt} = \sum_l \left(\frac{f_{kl}}{I_l}p_{il} + \frac{g_{kl}}{I_l}p_{il} - \frac{f_{lk}}{I_k}p_{ik} - \frac{g_{lk}}{I_k}p_{ik}\right) - p_{ik}\sum_{j\neq i}^{j\in\mathcal{A}(t)}\sum_l^m \frac{f_{kl}+f_{lk}}{I_k I_l}p_{jl}. \tag{12}$$

Including these types of transitions is discussed in more detail by Volz (2012).

For our model with ART, all $g_{kl} = 0$ except g_{lt} which gives the rate at which lineages transition into the treatment class:

$$G(t) = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & \gamma_t \\ 0 & 0 & 0 \end{pmatrix}. \tag{13}$$

Parameters and prior distributions for the ART model are given in Table 5. In this model, we still estimate the time-varying transmission rates $\beta_{e\to l}(t)$ and $\beta_{l\to l}(t)$, but transmission from individuals on ART is parameterized in terms of a time-invariant scalar κ_t of the time-varying local transmission rate: $\beta_{t\to l}(t) = \kappa_t\beta_{l\to l}(t)$.

2.7.4 AIDS-related deaths model

Our fourth model variant allows HIV-positive individuals to progress to an AIDS stage of infection with a higher death rate. Therefore, unlike in our other models where infected individuals are 'removed' from the infectious population but do not suffer increased mortality, AIDS-related deaths can permanently remove infected individuals from the entire population. This may be a more realistic way of modeling the demographic impact of the HIV epidemic, as there is strong evidence that AIDS-related deaths dramatically reduced adult lifespan before ART become widely available (Herbst et al. 2011; Bor et al. 2013). However, demographic surveillance in the AHRI study area indicates that the adult population size did not substantially change over this same time period. We therefore let net immigration into the local population demographically compensate for increased adult mortality due to AIDS. In our model, AIDS-related deaths are therefore directly offset by immigration from the external population.

The epidemic dynamics under this model are described by the following system of differential equations:

$$\begin{aligned}
\frac{dS_l}{dt} &= \mu N_l + \mu_a(t)I_a\frac{S_e}{N_e} - \left(\beta_{e\to l}(t)I_e + \beta_{l\to l}(t)I_l + \beta_{a\to l}(t)I_a\right)\frac{S_l}{N_l} - \mu S_l \\
\frac{dI_l}{dt} &= \mu_a(t)I_a\frac{I_e}{N_e} + \left(\beta_{e\to l}(t)I_e + \beta_{l\to l}(t)I_l + \beta_{a\to l}(t)I_a\right)\frac{S_l}{N_l} - (\mu + \gamma_a)I_l \\
\frac{dI_a}{dt} &= \gamma_a I_l - \mu_a(t)I_a.
\end{aligned} \tag{14}$$

Table 3. Parameters and prior distributions for model with hot spots of prevalence.

Parameter	Name	Prior distribution	Prior values
$\beta c \rightarrow c$	Local transmission rate	Log-normal	$\mu = -2.3026$; $\sigma = 1.0$
$\beta e \rightarrow c$	External introduction rate	Log-normal	$\mu = -2.3026$; $\sigma = 1.0$
$\kappa e \rightarrow h$	Transmission scalar	Log-normal	$\mu = 0.6931$; $\sigma = 1.0$
$\kappa h \rightarrow h$	Transmission scalar	Log-normal	$\mu = 0.0$; $\sigma = 1.0$
$\kappa h \rightarrow c$	Transmission scalar	Log-normal	$\mu = 0.0$; $\sigma = 1.0$
$\kappa c \rightarrow h$	Transmission scalar	Log-normal	$\mu = 0.0$; $\sigma = 1.0$
Nh	Hot spot pop size	Fixed	22,000
Nc	Cold spot pop size	Fixed	33,000

Table 4. Prior distributions on time-varying removal rates and the implied mean duration of infection.

Parameter	Time period	Prior distribution	Prior values	Mean duration (years)
ν	<2002	Log-normal	$\mu = -2.2769$; $\sigma = 0.1$	9.75
ν	2002–4	Log-normal	$\mu = -2.6208$; $\sigma = 0.1$	13.75
ν	2004–6	Log-normal	$\mu = -2.8762$; $\sigma = 0.1$	17.75
ν	2006–14	Log-normal	$\mu = -3.0795$; $\sigma = 0.1$	21.75

Table 5. Parameters and prior distributions for model with ART.

Parameter	Name	Prior distribution	Prior values
κt	ART transmission scalar	Beta	$\alpha = 1.0$; $\beta = 1.0$
γt	ART initiation rate	Log-normal	$\mu = -1.0996$; $\sigma = 0.5$
νt	ART removal rate	Log-normal	$\mu = -2.2769$; $\sigma = 0.5$

Here, I_a represents the number of infected individuals in the AIDS stage. The parameter γ_a controls the rate at which infected individuals progress to AIDS. The AIDS-related death rate $\mu_a(t)$ is allowed to differ from the death rate μ in the general population and vary over time. AIDS-related deaths are instantaneously offset by immigration from the external population. What fraction of immigrants are susceptible or infected is determined by the susceptible and infected fractions in the external population.

The structured coalescent model corresponding to this epidemiological model has the transmission rate matrix:

$$F(t) = \begin{pmatrix} \beta_{e\rightarrow e}\frac{S_e}{N_e}I_e & \beta_{e\rightarrow l}(t)\frac{S_l}{N_l}I_e & 0 \\ \beta_{l\rightarrow e}I_l & \beta_{l\rightarrow l}(t)\frac{S_l}{N_l}I_l & 0 \\ \beta_{a\rightarrow e}I_t & \beta_{a\rightarrow l}(t)\frac{S_l}{N_l}I_a & 0 \end{pmatrix}. \quad (15)$$

To account for lineage movement between infected compartments that occurs independently of transmission, we can again use (12) to track the lineage state probabilities, with the state transitions rates:

$$G(t) = \begin{pmatrix} 0 & \mu_a(t)I_a\frac{I_e}{N_e} & 0 \\ 0 & 0 & \gamma_a I_l \\ 0 & 0 & 0 \end{pmatrix}. \quad (16)$$

The transition rate $\mu_a(t)I_a\frac{I_e}{N_e}$ accounts for the movement of lineages into the local population through immigration of already infected individuals. The rate $\gamma_a I_l$ accounts for lineage movement due the progression of infected individuals to the AIDS stage.

Parameters and prior distributions for the AIDS model are given in Table 6. As in the ART model, the transmission rate from individuals in the AIDS stage is parameterized in terms of a time-invariant scalar κ_a of the time-varying local transmission rate: $\beta_{a\rightarrow l}(t) = \kappa_a\beta_{l\rightarrow l}(t)$. Estimating the AIDS progression rate γ_a as free parameter, we obtained extremely high rates of progression to AIDS resulting in unrealistic estimates of several other parameters. We therefore fixed $\gamma_a = 0.25$, giving an average time from infection to AIDS of four years.

4. Results

To help situate the local epidemic in the AHRI study population within the larger context of the southern African HIV epidemic, an ML phylogeny was reconstructed from HIV-1 subtype C sequences sampled from 1,068 infected individuals in the AHRI population along with 11,289 sequences from a larger regional background dataset (Wilkinson et al. 2016) sampled throughout southern Africa (Fig. 2A). Branch lengths in the ML phylogeny were then rescaled into units of calendar time using least squares dating. Although there are some larger clades composed predominantly of local viral samples which likely represent locally evolving sub-epidemics, the majority of samples from the AHRI are interspersed throughout clades composed predominantly of external samples (i.e. from the regional background dataset), suggesting that many independent introduction events have occurred into the local population from elsewhere in South Africa or from other neighboring countries.

Table 6. Parameters and prior distributions for model with AIDS-related deaths.

Parameter	Name	Prior distribution	Prior values
Kt	ART transmission scalar	Beta	$\alpha = 1.0$; $\beta = 1.0$
γt	ART initiation rate	Log-normal	$\mu = -1.0996$; $\sigma = 0.5$
νt	ART removal rate	Log-normal	$\mu = -2.2769$; $\sigma = 0.5$

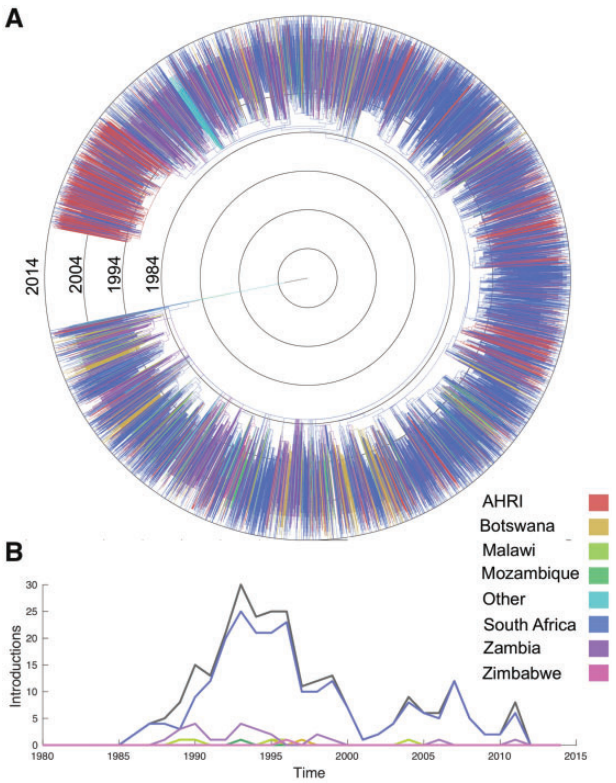

Figure 2. The local HIV epidemic in the AHRI study population within the larger phylogenetic context of the southern African subtype C epidemic. (A) ML phylogenetic tree reconstructed from HIV *pol* sequences from the AHRI (local) along with the regional background dataset. Tips are colored by sampling location, internal branches are colored according to their ancestral location reconstructed via maximum parsimony. (B) Time series showing the temporal distribution of external introductions from each country into the local population, as identified by maximum parsimony. The black line gives the total number of introductions summed over all countries.

The ancestral location of each lineage in the ML tree was then reconstructed using maximum parsimony to reveal how viral lineages have moved over time (Fig. 2A). This allowed us to identify potential external introductions at branches in the phylogeny where the most parsimonious ancestral location transitioned from the external to the local population. Note that we define an external introduction as transmission from an individual in the external population to an individual living in the local population independently of whether the transmission event occurred inside or outside the local population because the phylogeny contains no information about the exact location of transmission. In total, 248 external introductions were identified into the local population. When transitions in ancestral location occurred between parent and child nodes, we used the midpoint between the parent and child node as a proxy for the timing of external introductions. Most of these presumed introduction events occurred between 1990 and 2000 during the period of rapid epidemic growth in South Africa, with relatively fewer introductions after 2000 once the epidemic stabilized (Fig. 2B). Most introductions occurred from elsewhere in South Africa, although a few potential introductions occurred from Botswana, Malawi, Mozambique, Zambia, and Zimbabwe.

Because phylogenies only contain lineages ancestral to sampled viruses, the number of external introductions identified in the foregoing analysis likely represents only a small fraction of the total number. Moreover, due to the large size of the epidemic relative to the number of infected individuals sampled (~8%), it is extremely unlikely that we would have sampled both descendants of any recent transmission event between an external donor and a local recipient. In fact, the most recent common ancestor of most pairs of sampled viruses predates the early stages of the South African epidemic (Fig. 2A). The timing of reconstructed state changes may therefore not be a reliable proxy for the timing of introduction events since introductions may have occurred more recently (see Fig. 3C). It is highly likely then that our preliminary analysis based on parsimony both underestimates the true number of external introductions and skews their temporal distribution towards the more distance past.

4.1 Phylodynamic analysis

We therefore developed a phylodynamic model based on a simple but still realistic epidemiological model using a previously described structured coalescent framework (Volz 2012; Müller et al. 2017) in order to quantify the contribution of external introductions versus local transmission. The model tracks the number of infected individuals in the external population along with local epidemic dynamics (Fig. 3A and B). The model also probabilistically tracks how lineages in the tree move between populations based on their sampling location and the estimated transmission rates between populations (Fig. 3C). Tracking the movement of lineages in this way allows us to estimate whether new infections were derived from a local or external source. We validated our model using phylogenies simulated to reflect different epidemic scenarios. In the first scenario, external introductions play a large role in driving and sustaining the local epidemic (positive control). In the second scenario, external introductions only play a minor role in seeding the epidemic (negative control). In both scenarios, we were able to accurately estimate both the overall epidemic dynamics and the incidence attributable to internal and external transmission (Fig. 3D and E).

Using the phylodynamic model, we reconstructed epidemic dynamics in the local population from phylogenies reconstructed from the viral samples. As expected, both prevalence and incidence rapidly grew during the 1990s in the local population, and then grew more slowly after 2004 (Fig. 4A and B). Posterior estimates of the external infected population size through time and all inferred parameters are shown in Supplementary Figs. S1–S4. After 2004, independent estimates of prevalence and incidence based on population surveillance data are available from the AHRI. Prevalence estimates based on surveillance data were about 10 per cent higher than our phylodynamic estimates, although both methods estimate a

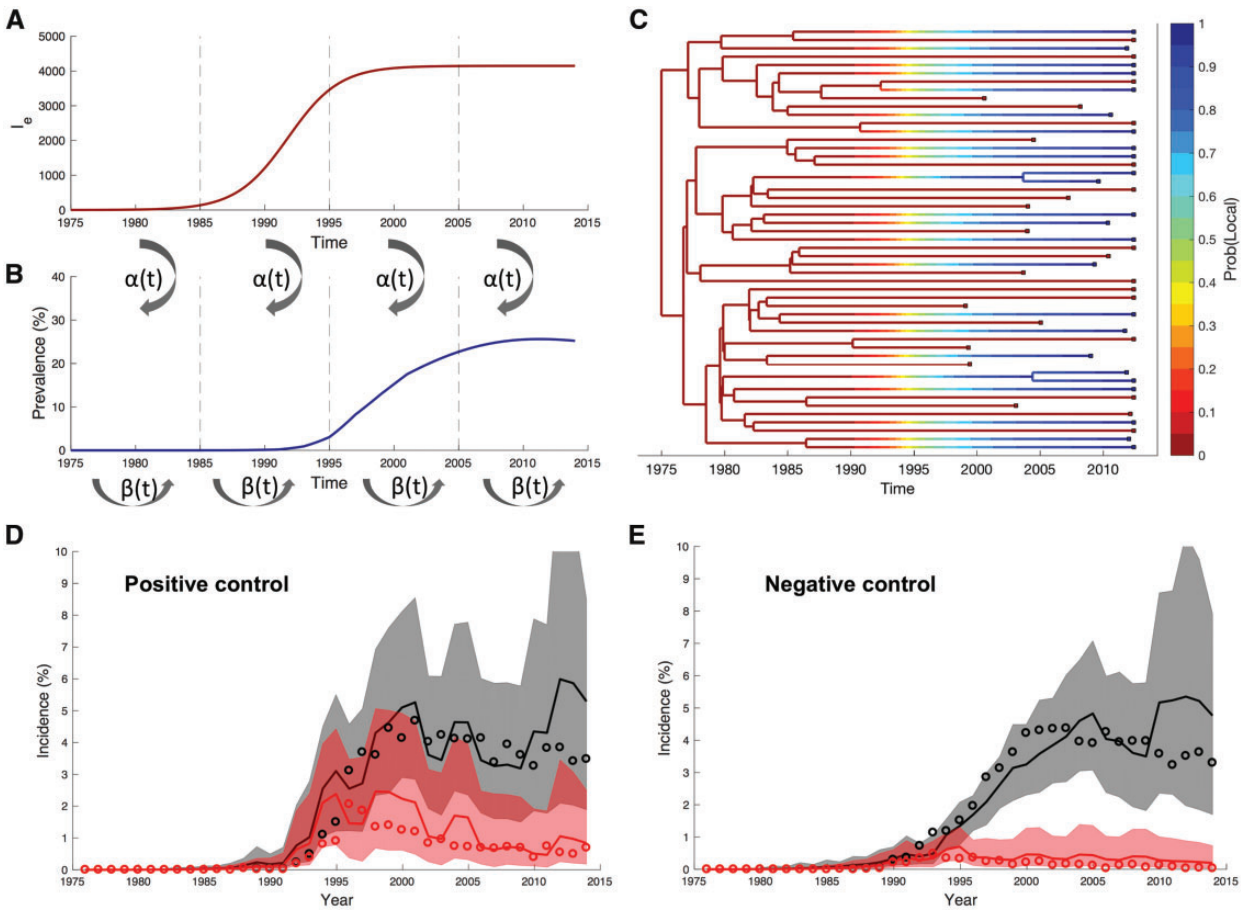

Figure 3. Schematic of the phylodynamic model and its validation on simulated data. Epidemic dynamics simulated under the model showing the number of infected individuals in the external population I_e (A) and the local population (B). Transmission events from the external to the local population occur at rate $\alpha(t)$ and within the local population at a rate proportional to $\beta(t)$. Both of these rates are time dependent and vary in a piecewise constant manner to accommodate changes in behavior, treatment or other interventions. Although not shown here, viral lineages can also be exported from the local population through transmission to the external population. (C) A simulated phylogeny generated under the same phylodynamic model. Each lineage is colored according to its probability of being in the local population (blue). These probabilities were computed under the model based on each lineage's sampling location and the estimated transmission rates between populations. (D and E) Total incidence (gray) and incidence attributable to external introductions (red) inferred from simulated phylogenies. Solid lines represent the posterior median estimate, shaded regions mark the 95 per cent credible intervals and open circles mark the true yearly incidence known from the simulations. In the positive control (D), we correctly infer that external introductions played a large role in driving and sustaining the local epidemic; whereas in the negative control (E), we correctly infer that external introductions only played a minor role in seeding the epidemic.

similar growth in prevalence since 2004 when widespread access to ART began (Fig. 4A), consistent with earlier reports by (Zaidi et al. 2013). Estimates of incidence since 2004 are in closer agreement, with both methods returning a median estimate of yearly incidence between 3 and 4 per cent with no detectable decline since ART coverage began increasing in 2004 (Fig. 4B).

Given that we could reliably reconstruct overall epidemic dynamics from simulated and empirical viral phylogenies, we used the phylodynamic model to quantify the relative contribution of external introductions versus local transmission to overall incidence. During the earliest stages of the local epidemic, the incidence attributable to external introductions was very high (Fig. 4B, red). While after 2004 the fraction attributable to external introductions declined, as of 2014 an estimated 35 per cent [95% confidence interval (CI): 20–60%] of all present day infections were due to external introductions (Fig. 4C).

4.2 Model testing and robustness

While our phylodynamic reconstruction of epidemic dynamics appears consistent with population-based surveillance, our model makes several simplifying assumptions about HIV's transmission dynamics and demography in the AHRI population. We therefore formulated four different variants of the base model used above to relax what we view as the most questionable of these assumptions. We then fit each variant to a single ML phylogeny to gauge how robust our phylodynamic estimates were to relaxing these assumptions. Informal comparisons of the marginal likelihood of the phylogeny under each model showed that all but one variant fit the phylogeny better than the base model (Table 7). There were also interesting differences in the epidemic dynamics reconstructed under each model as we discuss below.

The first model variant included geographic hot and cold spots of prevalence, relaxing the assumption in the base model of no population substructure within the AHRI population. Considering this form of subpopulation structure is a natural choice because the AHRI population is clustered into areas of high prevalence (>25%) along major roads and areas of low prevalence (<10%) in more inaccessible rural areas (Tanser et al. 2009) (Fig. 1C). While adding subpopulation structure substantially increased the marginal likelihood relative to the base

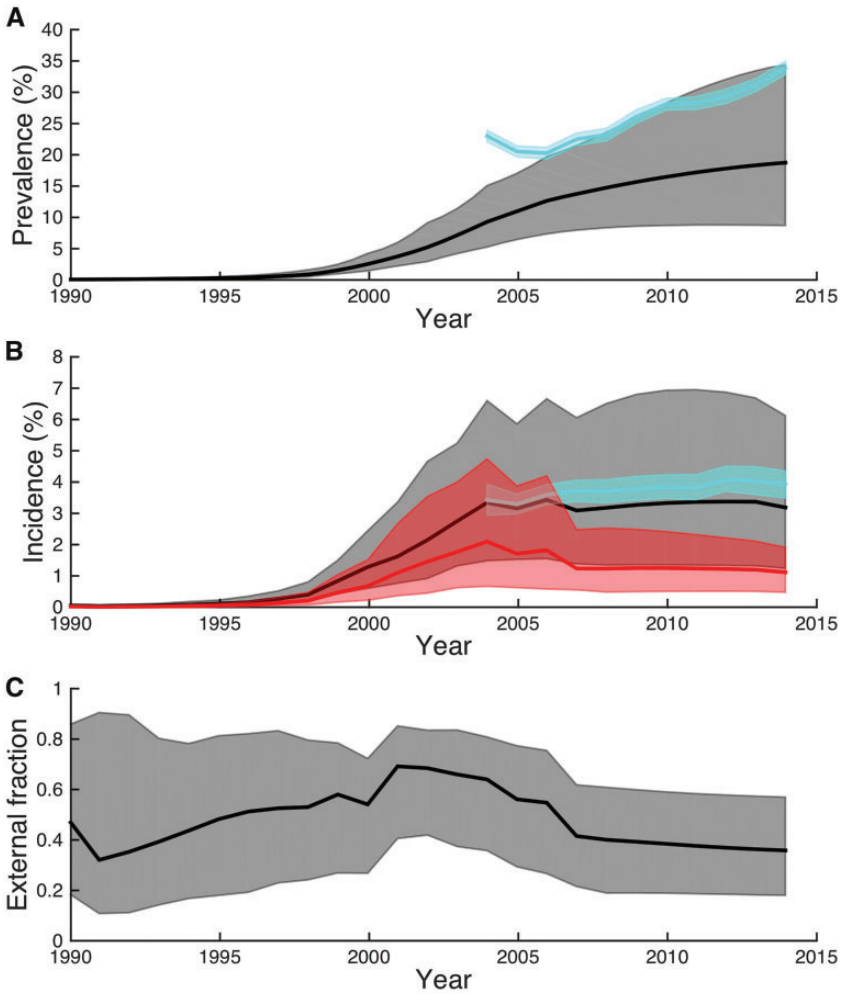

Figure 4. Epidemic dynamics reconstructed from viral phylogenies using the phylodynamic model. (A) Prevalence estimates from the phylogeny (gray) and independent surveillance data (blue). (B) Total incidence estimated from the phylogeny (gray) and surveillance data (blue). Incidence attributable to external introductions estimated from the phylogeny is shown in red. (C) The fraction of incidence attributable to external introductions over time. All solid lines represent the posterior median estimates while shaded regions mark the 95 per cent credible intervals. All estimates represent a posterior average over a set of phylogenies reconstructed from different sub-sampled datasets and thus take into account both phylogenetic uncertainty and sampling variance.

Table 7. Comparison of phylodynamic model fit and epidemiological estimates as of 2014.

Model	Marginal log likelihood	Prevalence (%)	Incidence (%)	Fraction external
Base	−18,278	21	2.37	0.35
Hot spots	−18,229	21.3	2.6	0.31
txRemoval	−18,282	27.6	3.61	0.21
ART	−18,255	35.1	4.82	0.2
AIDS	−18,239	47	5.1	0.05 (+0.31 migrants)

Values reported are median posterior estimates.

model (Table 7), prevalence is still underestimated relative to population-based surveillance and incidence is also estimated to be slightly lower (Fig. 5D and E). Nevertheless, our estimates of the fraction of incidence attributable to external introductions are very similar to those estimated under the base model (Fig. 5F).

Because prevalence was underestimated in the base and hot spot model, we considered a second model where individuals could remain in the infected class longer due to removal rates decreasing over time. Adding time-varying removal rates to the base model did result in a more realistic increase in prevalence towards present, but total prevalence was still underestimated relative to population-based surveillance (Fig. 5G). Adding time-varying removal rates did not increase the marginal likelihood relative to the base model, but unlike the other model variants this model did not add an additional population state to the phylodynamic model.

Because allowing for a longer duration of infection appeared to improve our estimates of prevalence, we considered a third model where infected individuals could initiate ART after 2004.

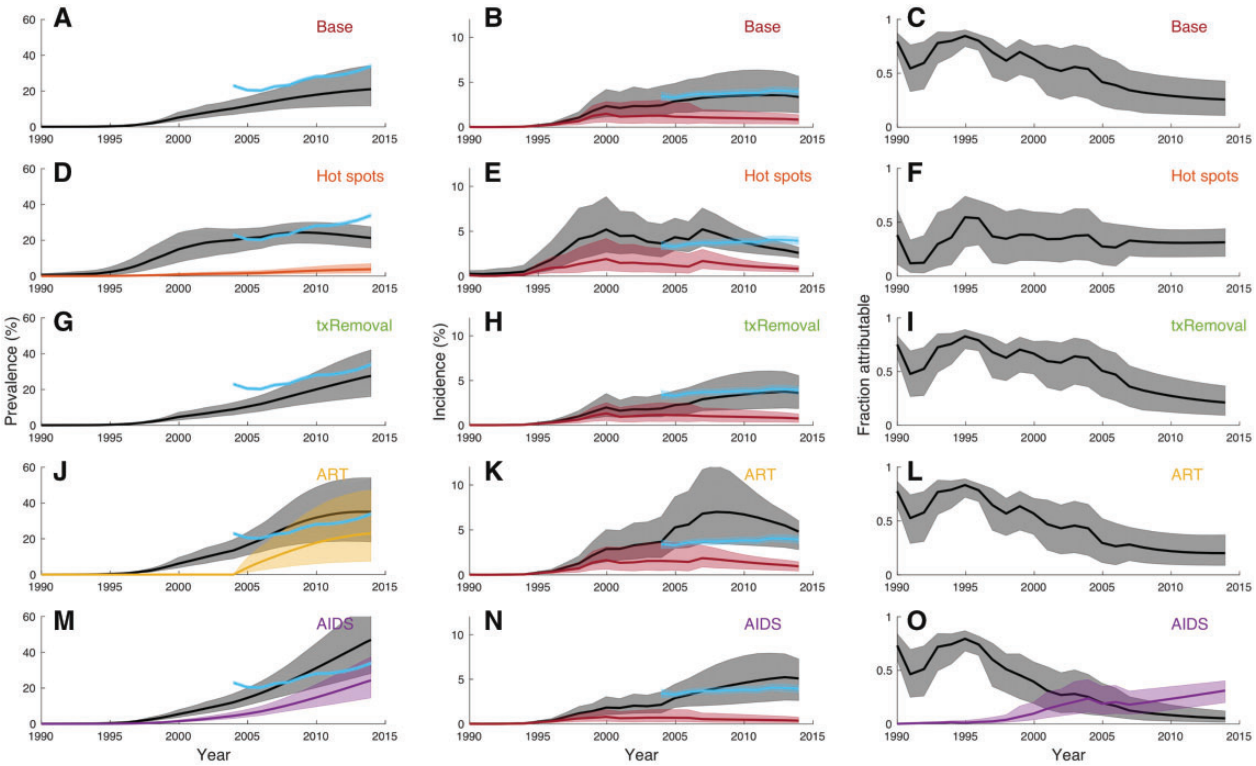

Figure 5. Epidemic dynamics reconstructed under different variants of the phylodynamic model. (A–C) Estimates from fitting the base model to the same ML phylogeny as the other model variants. (D–F) Model with geographic hot and cold spots of prevalence. Estimated prevalence in hot spots as a percentage of the total population size is shown in orange. (G–I) Model with time-varying removal rates. (J–L) Model that included antiretroviral treatment after 2004. Prevalence of infected individuals on ART is shown in gold. (M–O) Model with AIDS-related deaths and migration from external population. Prevalence of AIDS is shown in purple. (O) The fraction of new cases attributable to immigration of already infected individuals is also shown in purple. As before, the fraction of new cases attributable to transmission from the external population is shown in gray. Prevalence and incidence estimates from population-based surveillance (blue) are replicated here for easy comparison with Fig. 4.

Under this model, ART was allowed to not only increase the duration of infection but also reduce transmissibility for individuals on ART. Interestingly, prevalence estimates become more consistent with population-based surveillance under this model while total incidence was overestimated towards present (Fig. 5J and K). Estimates of the fraction attributable to external introductions were only slightly lower than under the base model (Fig. 5L, Table 7).

Finally, the fourth model included progression to a late stage of infection or AIDS with an increased death rate, consistent with observations that life expectancy dropped significantly during the early stages of the KZN epidemic before ART become widely available (Bor et al. 2013). However, an increased AIDS-related death rate would cause the adult population size to decline over time, inconsistent with demographic surveillance data from the AHRI study area. We therefore allowed AIDS-related deaths to be offset by immigration from the external population to keep the adult population size constant. Because there is no evidence that in-migration was significantly higher than out-migration this is a rather extreme assumption, but it allowed us to test our results against the assumption that there was no direct migration of already infected individuals into the local population. Under this model, phylodynamic estimates of prevalence were substantially higher than under other models, largely due to already infected individuals migrating into the local population (Fig. 5M). We also estimated a relatively lower fraction of incidence due to transmission from the external population (Fig. 5N, Table 7). However, this is not necessarily inconsistent with our other estimates. Fewer external introductions due to transmission from the external population were simply offset by immigration of already infected individuals into the local population (Fig. 5O), which can be viewed as another source of external introductions.

5. Discussion

Our phylodynamic results on external introductions fit within a growing body of evidence that transmission networks are highly interconnected across communities even at relatively large geographic distances due to human movement patterns. Historians and social scientists have long noted that both mass migration and increased mobility may have played an important role in the rapid spread of HIV through southern Africa (Iliffe 2006). In South Africa, in particular, historical patterns of circular labor migration among neighboring countries and increased mobility following the end of the Apartheid system may have contributed to the rapid growth of the epidemic (Jochelson et al. 1991; Lurie et al. 1997; Hargrove 2008). More recently, phylogenetic studies of HIV in Africa have provided support for frequent viral movement at spatial scales ranging from local communities (Grabowski et al. 2014) to across national borders (Wilkinson et al. 2016). This frequent movement has been linked to highly mobile individuals such as economic migrants, soldiers, sex workers, and truck drivers (Gray et al. 2009; Wilkinson et al. 2016).

Our phylogenetic and phylodynamic analysis of the HIV epidemic in the Africa Health Research Institute study population revealed that external introductions via human movement

played and continue to play a vital role in driving the epidemic in rural KZN. A preliminary phylogenetic analysis based on maximum parsimony suggested that a large wave of introductions occurred in the 1990s during the early stages of the South African epidemic. A subsequent analysis using a more realistic phylodynamic model confirmed that the earliest stages of the epidemic were indeed largely driven by external introductions, although these early introductions may have been less wave-like.

Our phylodynamic analysis also suggested that, far from just seeding the local epidemic, external introductions continue to play an important role in sustaining the high incidence of HIV in local KZN populations. This has direct relevance for ART as prevention (TasP) trials and programs. If most transmission occurs locally, TasP programs should be able to efficiently and cost-effectively reduce incidence (Granich et al. 2009) and could selectively target communities with higher incidence (Tanser et al. 2009). However, if most new infections are acquired externally, then increasing local ART coverage may not substantially reduce incidence. Our phylodynamic estimates for one KZN population indicate that the situation likely lies between these two extremes. Our median estimate is that presently 35 per cent (95% CI: 20–60%) of new infections in the AHRI population are attributable to external introductions, suggesting a substantial number of new infections could be prevented if the source of these infections could also be targeted. Nevertheless, these results imply that the majority of new infections may be attributable to local transmission, and increasing local ART coverage may still prevent many future infections. However, a recent TasP trial in a population immediately adjacent to the AHRI study area showed no incidence reduction despite universal access to ART (Iwuji et al. 2017). Identifying the source of new infections and whether they arose locally from untreated infections or from external introductions will be key to understanding why this trial failed to decrease incidence.

While phylogenies can reveal the movement of viruses between populations, they cannot necessarily reveal where transmission events occurred or how viral lineages first entered a population without additional information about human movement. An external introduction event may result from a visitor transmitting to an individual residing in the local population or while an individual currently residing in the local population was living away or traveling. In our phylodynamic model, we cannot distinguish between these two scenarios because they result in identical phylogenetic patterns. The AHRI does however collect data on the residence and migration status of individuals in the study area. Younger men and women frequently leave and return to the area, and those who leave spend a substantial amount of their time (30.8%) outside the area (Dobra et al. 2017). About 55 per cent of these migration events occur within a 300-km radius of the study area and generally fall within the province of KZN, including Richards Bay and Durban (Dobra et al. 2017). Furthermore, these data show that prevalence is higher among residents with histories of recent migration (McGrath et al. 2015). Individuals who spend more time away and travel longer distances outside their community also have a significantly higher risk of acquiring HIV infection (Dobra et al. 2017). Both of these observations support the idea that local residents may be infected while living or traveling outside of the study area and then return to the area, highlighting the potential importance of circulatory migration patterns. The extent to which external introductions are occurring outside local communities could be further teased apart by combining viral phylogenetics with detailed sociological data on the migration history of sampled individuals to infer the location and migration status of individuals at the time of infection.

Although phylodynamics is increasingly used to study epidemic dynamics, the validity of estimates derived entirely from viral phylogenies can of course be questioned and have their own limitations as just discussed. We however feel confident that our main results about external introductions are reliable for several reasons. First, we validated our model on simulated phylogenies and found that we could accurately reconstruct total incidence and the fraction of incidence attributable to external introductions in scenarios where introductions both did and did not play a large role in driving epidemic dynamics. Second, we were able to reconstruct dynamics consistent with prevalence and incidence trends estimated from independent surveillance data. Third, while there is no ‘gold standard’ to which we can compare our phylodynamic estimates of external introductions, our estimated fraction of external introductions is consistent with known mobility patterns in the AHRI population where 38 per cent of males and 32 per cent of females are recent migrants or frequently leave the area (Camlin et al. 2010; Muhwava et al. 2013). Finally, our estimates of external introductions appear robust to the exact formulation of the epidemiological model assumed. Different variants of our model including prevalence hot spots, time-varying removal rates, ART or AIDS-related deaths all returned similar estimates of external introductions.

In addition to showing that external introductions play an important role in sustaining high HIV incidence in this hyper-epidemic setting, we were also able to accurately estimate population-level incidence. We believe that this is also an important result of our study, as there is currently great interest in using phylodynamics to quantify epidemic dynamics, especially changes in incidence, without the need for expensive longitudinal cohort studies (Dennis et al. 2014; Cohen 2015). A recent simulation study has demonstrated that it may be possible to estimate incidence dynamics from phylogenies in the context of African HIV epidemics (Ratmann et al. 2017), but the accuracy of estimates varied considerably depending on the choice of phylodynamic model, with the most complex structured coalescent model vastly outperforming other models. In contrast, we accurately estimated incidence and, to a lesser extent, prevalence from empirical phylogenies using relatively simple models.

In this respect, we view the simplicity of our model as one of its strengths, even though far more detailed epidemiological models have been developed in recent years for the purposes of predicting and evaluating the impact of HIV prevention efforts (see Eaton et al. 2012 for a review). Our desire for simplicity was motivated by the fact that pathogen phylogenies are only a proxy for the unobserved transmission tree, and only contain information about a subset of transmission events from which we can draw inferences about epidemic dynamics. By itself then, a pathogen phylogeny contains no information about the underlying drivers of transmission. We therefore opted to make minimal mechanistic assumptions about how changes in risk behavior, prevention efforts, and clinical care all interact to shape HIV’s complex transmission dynamics. Adding parameters or state variables to our model in order to capture these processes may reduce certain biases, but will inevitably increase the variance of our phylodynamic estimates since the phylogeny contains no information about them.

Nevertheless, it remains largely unexplored what epidemiological factors need to be included in phylodynamic models in order for estimates to be accurate and robust to slight model

misspecifications. This is especially true when we try to relate population-level variables of interest like prevalence or incidence back to the branching structure of a phylogeny since many processes simultaneously shape the phylogeny and may confound analysis if not properly accounted for. Previous work has shown that major subdivisions within a host population can distort inferences because lineages in different subpopulations are no longer exchangeable as standard coalescent theory assumes (Heller et al. 2013; Rasmussen et al. 2014a). We therefore accounted for subdivision between local and external populations using a structured coalescent model (Volz 2012). But it is less clear what other forms of host heterogeneity need to be considered for HIV. For example, contact heterogeneity in sexual networks can also shape pathogen phylogenies, increasing the coalescent rate during the early stages of an epidemic as the pathogen spreads though highly connected parts of the network (O'Dea and Wilke 2011; Leventhal et al. 2012; Rasmussen et al. 2017). It is therefore possible that if we had accounted for contact heterogeneity, we may have estimated a higher incidence during the early stages of the epidemic in the AHRI population, bringing our estimates of prevalence closer to those observed when population-based surveillance began in 2004.

For a chronic infection like HIV, it may also be important to correctly model the progression of individuals though different stages of infection or clinical care. Phylodynamic models for HIV that allow for disease progression have been shown to accurately reconstruct epidemic dynamics from real and simulated data (Volz et al. 2013; Ratmann et al. 2017). In our case, adding an infection class for people on ART corrected our underestimates of prevalence by allowing individuals on ART to remain infected longer. But while it is conceptually straightforward to model multiple stages of infection, transmission and progression rates for each of these stages can also vary over time, leading to very parameter-rich models. As a first-order approximation, we therefore tried to circumvent this complexity by only modeling temporal variability in transmission rates independent of disease progression. However, this strategy proved insufficient to capture realistic changes in prevalence and it was necessary to include either time-varying removal rates or an ART-infected class to account for increasing durations of infection after the rollout of ART in 2004. Encouragingly though, our estimates of incidence were surprisingly accurate even under the base model, suggesting that highly detailed models may not be necessary to estimate incidence accurately. We suggest that this may be due to the fact that, while simple, our base model already includes two of the most relevant features of HIV's transmission dynamics in rural Africa: time-varying transmission rates and local versus external contact structure.

In conclusion, we believe our results demonstrate the power of using phylodynamics to study HIV transmission dynamics in large, generalized African epidemics. Using a relatively simple phylodynamic model, we were able to quantify the relative contribution of external introductions to address a longstanding question about the role human mobility plays in local HIV epidemics, while showing for the first time that it is possible to accurately estimate HIV incidence from empirical data in the context of generalized African epidemics. Given the promising performance of our phylodynamic model, we have made it freely available as an add-on to the widely used phylogenetic software package BEAST 2 (Bouckaert et al. 2014; Rasmussen 2017). Finally, we note that as HIV sequence datasets expand to encompass larger spatial regions, it should be possible to use similar phylodynamic approaches to quantify transmission rates between multiple different communities and thereby pinpoint the geographic source of new infections—providing detailed knowledge that can be used directly to prevent new infections.

Acknowledgments

We would like to thank Andrew Tomita for providing help and data from the Africa Centre Demographic Information Service. D.A.R. was funded by the ETH Zürich Postdoctoral Fellowship Program, the Marie Curie Actions for People COFUND, and a travel fellowship from the SAMRC to visit AHRI and KRISP in South Africa. T.d.O., E.W., and A.V. are supported by a research Flagship grant from the South African Medical Research Council (MRC-RFA-UFSP-01-2013/UKZN HIVEPI), a Royal Society Newton Advanced Fellowship (T.d.O.), and the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 634650. D.P. and AHRI are funded by the Wellcome Trust. T.S. is supported in part by the European Research Council under the Seventh Framework Programme of the European Commission (PhyPD: Grant Agreement Number 335529, http://erc.europa.eu).

Supplementary data

Supplementary data are available at *Virus Evolution* online.

Conflict of interest: None declared.

References

Bor, J. et al. (2013) 'Increases in Adult Life Expectancy in Rural South Africa: Valuing the Scale-up of HIV Treatment', *Science*, 339: 961–5.

Bouckaert, R. et al. (2014) 'BEAST 2: A Software Platform for Bayesian Evolutionary Analysis', *PLoS Computational Biology*, 10: e1003537.

Camlin, C. S. et al. (2010) 'Gender, Migration and HIV in Rural KwaZulu-Natal, South Africa', *PLoS One*, 5: e11539.

Coffee, M., Lurie, M. N., and Garnett, G. P. (2007) 'Modelling the Impact of Migration on the HIV Epidemic in South Africa', *Aids (London, England)*, 21: 343–50.

Cohen, J. (2015) 'HIV Family Trees Reveal Viral Spread', *Science (New York, N.Y.)*, 348: 1188–9.

Corno, L., and De Walque, D. (2012) *Mines, migration and HIV/AIDS in Southern Africa*. The World Bank.

de Oliveira, T. et al. (2017) 'Transmission Networks and Risk of HIV Infection in KwaZulu-Natal, South Africa: A Community-Wide Phylogenetic Study', *The Lancet HIV*, 4: e41–50.

Deane, K. D., Parkhurst, J. O., and Johnston, D. (2010) 'Linking Migration, Mobility and HIV', *Tropical Medicine & International Health*, 15: 1458–63.

Dennis, A. M. et al. (2014) 'Phylogenetic Studies of Transmission Dynamics in Generalized HIV Epidemics: An Essential Tool Where the Burden Is Greatest?', *Journal of AIDS*, 67: 181.

Dobra, A. et al. (2017) 'Space-Time Migration Patterns and Risk of HIV Acquisition in Rural South Africa', *Aids (London, England)*, 31: 137.

Eaton, J. W. et al. (2012) 'HIV Treatment as Prevention: Systematic Comparison of Mathematical Models of the Potential Impact of Antiretroviral Therapy on HIV Incidence in South Africa', *PLoS Medicine*, 9: e1001245.

Faria, N. R. et al. (2011) 'Toward a Quantitative Understanding of Viral Phylogeography', *Current Opinion in Virology*, 1: 423–9.

Grabowski, M. K. et al. (2014) 'The Role of Viral Introductions in Sustaining Community-Based HIV Epidemics in Rural Uganda: Evidence from Spatial Clustering, Phylogenetics, and Egocentric Transmission Models', *PLoS Medicine*, 11: e1001610.

Granich, R. M. et al. (2009) 'Universal Voluntary HIV Testing with Immediate Antiretroviral Therapy as a Strategy for Elimination of HIV Transmission: A Mathematical Model', *The Lancet*, 373: 48–57.

Gray, R. R. et al. (2009) 'Spatial Phylodynamics of HIV-1 Epidemic Emergence in East Africa', *Aids (London, England)*, 23: F9.

Hargrove, J. (2008) 'Migration, Mines and Mores: The HIV Epidemic in Southern Africa', *South African Journal of Science*, 104: 53–61.

Harrison, A., Cleland, J., and Frohlich, J. (2008) 'Young People's Sexual Partnerships in KwaZulu/Natal, South Africa: Patterns, Contextual Influences, and HIV Risk', *Studies in Family Planning*, 39: 295.

Heller, R., Chikhi, L., and Siegismund, H. R. (2013) 'The Confounding Effect of Population Structure on Bayesian Skyline Plot Inferences of Demographic History', *PLoS One*, 8: e62992.

Herbst, A. J., Mafojane, T., and Newell, M.-L. (2011) 'Verbal Autopsy-Based Cause-Specific Mortality Trends in Rural KwaZulu-Natal, South Africa, 2000-2009', *Population Health Metrics*, 9: 47.

Holmes, E. C. (2004) 'The Phylogeography of Human Viruses', *Molecular Ecology*, 13: 745–56.

Iliffe, J. (2006) *The African AIDS Epidemic: A History*. Athens: Ohio University Press.

Iwuji, C. C. et al. (2017) 'Universal Test and Treat and the HIV Epidemic in Rural South Africa: A Phase 4, Open-Label, Community Cluster Randomised Trial', *The Lancet HIV*, 5: e116–25.

Jochelson, K., Mothibeli, M., and Leger, J.-P. (1991) 'Human Immunodeficiency Virus and Migrant Labor in South Africa', *International Journal of Health Services*, 21: 157–73.

Karim, Q. A. et al. (2011) 'Stabilizing HIV Prevalence Masks High HIV Incidence Rates Amongst Rural and Urban Women in KwaZulu-Natal, South Africa', *International Journal of Epidemiology*, 40: 922–30.

Kharsany, A. B. et al. (2015) 'Strengthening HIV Surveillance in the Antiretroviral Therapy Era: Rationale and Design of a Longitudinal Study to Monitor HIV Prevalence and Incidence in the uMgungundlovu District, KwaZulu-Natal, South Africa', *BMC Public Health*, 15: 1.

Leventhal, G. E. et al. (2012) 'Inferring Epidemic Contact Structure from Phylogenetic Trees', *PLoS Computational Biology*, 8: e1002413.

Lurie, M. et al. (1997) 'Circular Migration and Sexual Networking in Rural KwaZulu/Natal: Implications for the Spread of HIV and Other Sexually Transmitted Diseases', *Health Transition Review*, 7: 17–27.

Lurie, M. N. et al. (2003) 'The Impact of Migration on HIV-1 Transmission in South Africa: A Study of Migrant and Nonmigrant Men and Their Partners', *Sexually Transmitted Diseases*, 30: 149–56.

Manasa, J. et al. (2016) 'Increasing HIV-1 Drug Resistance between 2010 and 2012 in Adults Participating in Population-Based HIV Surveillance in Rural KwaZulu-Natal, South Africa', *AIDS Research and Human Retroviruses*, 32: 763–9.

McGrath, N. et al. (2015) 'Migration, Sexual Behaviour, and HIV Risk: A General Population Cohort in Rural South Africa', *The Lancet HIV*, 2: e252–9.

Minin, V. N., Bloomquist, E. W., and Suchard, M. A. (2008) 'Smooth Skyride through a Rough Skyline: Bayesian Coalescent-Based Inference of Population Dynamics', *Molecular Biology and Evolution*, 25: 1459–71.

Muhwava, W. et al. (2013) 'Levels and Determinants of Migration in Rural KwaZulu-Natal, South Africa', *African Population Studies*, 24: 259–80.

Müller, N. F., Rasmussen, D. A., and Stadler, T. (2017) 'The Structured Coalescent and Its Approximations', *Molecular Biology and Evolution*, 34(11): 2970–81.

Nel, A. et al. (2012) 'HIV Incidence Remains High in KwaZulu-Natal, South Africa: Evidence from Three Districts', *PLoS One*, 7: e35278.

O'Dea, E. B., and Wilke, C. O. (2011) 'Contact Heterogeneity and Phylodynamics: How Contact Networks Shape Parasite Evolutionary Trees', *Interdisciplinary Perspectives on Infectious Diseases*, 2011: 1.

Price, M. N., Dehal, P. S., and Arkin, A. P. (2010) 'Fasttree 2–Approximately Maximum-Likelihood Trees for Large Alignments', *PLoS One*, 5: e9490.

Pybus, O. G., and Rambaut, A. (2009) 'Evolutionary Analysis of the Dynamics of Viral Infectious Disease', *Nature Reviews Genetics*, 10: 540–50.

Quinn, T. C. (1994) 'Population Migration and the Spread of Types 1 and 2 Human Immunodeficiency Viruses', *Proceedings of the National Academy of Sciences*, 91: 2407–14.

Rasmussen, D. A. (2017) *Marula BEAST 2 Package*. <https://github.com/davidrasm/Marula> accessed Nov 2018.

——, Boni, M. F., and Koelle, K. (2014) 'Reconciling Phylodynamics with Epidemiology: The Case of Dengue Virus in Southern Vietnam', *Molecular Biology and Evolution*, 31: 258–71.

——, Volz, E. M., and —— (2014) 'Phylodynamic Inference for Structured Epidemiological Models', *PLoS Computational Biology*, 10: e1003570.

—— et al. (2017) 'Phylodynamics on Local Sexual Contact Networks', *PLoS Computational Biology*, 13: e1005448.

Ratmann, O. et al. (2017) 'Phylogenetic Tools for Generalized HIV-1 Epidemics: Findings from the PANGEA-HIV Methods Comparison', *Molecular Biology and Evolution*, 34: 185–203.

Sankoff, D. (1975) 'Minimal Mutation Trees of Sequences', *SIAM Journal on Applied Mathematics*, 28: 35–42.

Shisana, O. et al. (2015) *South African National HIV Prevalence, Incidence and Behaviour Survey, 2012*. http://repository.hsrc.ac.za/handle/20.500.11910/2490.

Tanser, F. et al. (2008) 'Cohort Profile: Africa Centre Demographic Information System (ACDIS) and Population-Based HIV Survey', *International Journal of Epidemiology*, 37: 956–62.

—— et al. (2009) 'Localized Spatial Clustering of HIV Infections in a Widely Disseminated Rural South African Epidemic', *International Journal of Epidemiology*, 38: 1008–16.

To, T.-H. et al. (2015) 'Fast Dating Using Least-Squares Criteria and Algorithms', *Systematic Biology*.

UNAIDS (2017). *UNAIDS 2017 Estimates*. http://www.unaids.org/en/resources/documents/2017/2017_data_book.

Vandormael, A. et al. (2014) 'Use of Antiretroviral Therapy in Households and Risk of HIV Acquisition in Rural KwaZulu-Natal, South Africa, 2004–12: A Prospective Cohort Study', *The Lancet Global Health*, 2: e209–15.

—— et al. (2018) 'Incidence Rate Estimation, Periodic Testing, and the Limitations of the Mid-point Imputation Approach', *International Journal of Epidemiology*, 47: 236–45.

—— et al. (2018) 'High Percentage of Undiagnosed HIV Cases within a Hyperendemic South African Community: A Population-Based Study', *Journal of Epidemiology and Community Health*, 72: 168–72.

Volz, E. M. (2012) 'Complex Population Dynamics and the Coalescent under Neutrality', *Genetics*, 190: 187–201.

—— et al. (2013) 'HIV-1 Transmission during Early Infection in Men Who Have Sex with Men: A Phylodynamic Analysis', *PLoS Medicine*, 10: e1001568.

Welz, T. et al. (2007) 'Continued Very High Prevalence of HIV Infection in Rural KwaZulu-Natal, South Africa: A Population-Based Longitudinal Study', *AIDS*, 21: 1467–72.

Wilkinson, E. et al. (2016) 'Origin, Imports and Exports of HIV-1 Subtype C in South Africa: A Historical Perspective', *Infection, Genetics and Evolution*.

Zaidi, J. et al. (2013) 'Dramatic Increases in HIV Prevalence after Scale-up of Antiretroviral Treatment: A Longitudinal Population-Based HIV Surveillance Study in Rural Kwazulu-Natal', *AIDS (London, England)*, 27: 2301.