# *Phylogenetics and Networks for Generalised HIV Epidemics in Africa*

## Phylodynamic Methods Comparison Exercise – Appendix Regional Simulation

*Introduction*

The regional simulation captures individual-level HIV transmission dynamics in a larger regional population that is broadly similar to a site (cluster) of the HPTN071/PopART HIV prevention trial in South Africa (Hayes et al., 2014).

The model builds on the results of the model comparison exercise of the HIV modelling consortium, which compared mathematical model prediction on the potential impact of antiretroviral therapy on HIV incidence in South Africa (Eaton et al., 2012). This agent-based model is currently unpublished, however it is developed based in part on findings from the compartment model that was used to inform the setup of the HPTN071/PopART trial (Cori et al., 2014), and an extended version of this model will be used to help evaluate results from the HPTN071/PopART trial.

Table 1 gives an overview of the data files that are generated by the model.

### Table 1 Description of data files

| File name | Description |
|---|---|
| *.fa | Fasta file of aligned, simulated sequences. One for each gene. |
| *_metadata.csv | Information on sampled individuals. |
| *_SURVEY.csv | Counts of surveys conducted on a random subset of the simulated study population. |
| *_DATEDTREE.newick | Trees in newick format. Each tree corresponds to the simulated viral phylogeny among sampled individuals of one simulated transmission chain. One tree per line. |

### *Epidemiological component of the model*

The epidemiological model simulates a population of individuals, growing at a rate of approximately 2.5% per year. For each individual in the epidemiological model, when entering the sexually active population aged 13 they are assigned a set of characteristics. These include among others date of birth, gender, HIV status and ART status (if applicable). Characteristics of individuals change at events. Events include demographics (birth, death, ageing), partnership formation and breakup, HIV infection, HIV disease progression, HIV testing, voluntary male medical circumcision, and the ART cascade. All individuals within the population are heterosexual. Each individual is assigned to one of three risk categories that determine the maximum number of sexual partners they may have at one time. Individuals are able to form partnerships until they reach this limit. HIV transmission occurs only through sexual contact between individuals currently in a partnership, and depends on clinical characteristics (acute infection, CD4 count and set-point viral load) of the infecting partner, as well as the circumcision status of the transmitter (if male).

### *Evolutionary component of the model*

The evolutionary model describes the evolution of the *gag*, *pol* and *env* genes of the HIV-1 subtype C genome from the point of separate introductions into the study population.

An observation model specifies the infected individuals whose sequence is to be simulated. First, the number of sequences to be sampled in each year of the simulation is determined. This varies between data sets. The specified number of individuals is sampled at random from the diagnosed population.

The viral phylogeny evolves between and within individuals along simulated transmission chains through a coalescent model. The phylogeny is pruned to the sampled cases. This phylogeny has branch lengths in units of calendar time. Co-infections and viral recombination are not modelled.

Viral sequences are simulated according to a GTR-Gamma substitution model (Rambaut & Grassly, 1997). The model depends on codon position. Evolutionary rates depend on whether the virus is on a transmission lineage or not. The following genomic regions are simulated:

1. gag: p17 start to pol PROT start; length 1440 nucleotides. The simulated gag gene does not include the last 14 amino acids of p6, due to the overlap with pol.
2. pol: PROT start to Integrase end; length 2844 nucleotides.
3. env: CDS signal peptide start to gp41 end; length 2523 nucleotides.

### *Epidemiological simulation*

Epidemic simulations were calibrated to reflect key aspects of the HIV epidemic in South Africa (Cori et al., 2014).

After a period of burn-in to allow partnerships to settle, HIV is introduced into the population in 1980. After 1980, viral introductions from outside the simulated population occur at a non-negligible rate (Cori et al., 2014;

Grabowski et al., 2014). Several plausible epidemic scenarios were simulated (Cohen et al., 2012). In all scenarios, the simulated population of sexually active individuals aged 13 or over grows to approximately 80,000 individuals by 2020.

Active surveillance with first HIV diagnoses starts in 2000. Antiretroviral therapy begins in 2004, with ART CD4 eligibility criteria following the historical changes in South African national guidelines (CD4<200cells/mm$^3$ before 2011, <350cells/mm$^3$ until 2015). In 2015, a comprehensive HIV prevention campaign similar in spirit to the HPTN071/PopART intervention, with annual HIV testing and universal access to ART, is introduced in the community. The campaign lasts for three years. Several intervention scenarios were simulated from 2015 onwards. After the campaign, the simulation continues for several years at standard of care.

Data on population growth, diagnoses, and ART uptake is available through a cross-sectional survey on a random sample of less than 10% of the simulated population. Table 2 lists the available variables.

Individual-level data are available for those individuals with a sequence. Table 3 list the available information.

**Table 2 Variables in simulated survey in file "*_SURVEY.csv"**

| **Group variable** | **Description** |
|---|---|
| YR | Time at which the survey was taken |
| Gender | Gender |
| AGE | Age |
| | |
| **Count by group** | |
| ALIVE_N | Number of sexually active individuals |
| ALIVE_AND_DIAG_N | Number of diagnosed individuals |
| ALIVE_AND_ART_N | Number of individuals that started ART |
| ALIVE_AND_SEQ_N | Number of individuals with a sequence |

**Table 3 Variables of simulated individuals in file "*_metadata.csv"**

| **Variable** | **Description** |
|---|---|
| IDPOP | Identifier of individual that can be linked to sequences or tips in the viral phylogeny |
| GENDER | Gender<br>NA if archival sequence |
| DOB | Date of birth<br>NA if archival sequence |
| DOD | Date of death<br>NA if alive at end of simulation |
| DIAG_T | Time of diagnosis<br>NA if archival sequence |
| DIAG_CD4 | CD4 count at diagnosis<br>NA if archival sequence |

| | |
|---|---|
| ART1_T | ART start date<br>NA if ART not started |
| ART1_CD4 | CD4 count at ART start<br>NA if ART not started |
| TIME_SEQ | Date sequence taken |
| RECENT_TR | Y if transmission occurred at most 6 months after diagnosis, N otherwise |

### *Evolutionary simulation*

Available full genome subtype C sequences from sub Saharan Africa in the Los Alamos Sequence database were used to seed the simulation of HIV-1C viral variants in the study population. Separate viral introductions occurred at the introduction of HIV into the population in 1980, and thereafter at a constant rate over time. Viral introductions are simulated from a seed sequence. The date of the seed sequence was chosen so that the TMRCA of the simulated sequences falls is consistent with current phylogenetic estimates (Walker, Pybus, Rambaut, & Holmes, 2005).

Within-host lineages that are not part of transmission lineages have a higher evolutionary rate in the simulation (Alizon & Fraser, 2013; Vrancken et al., 2014).

A few 'archival' sequences are sampled for the period 1985 to 1999. Since 2000, sequences are more systematically sampled from infected individuals as part of HIV surveillance. Since 2015 until the end of the simulation, the population is more intensely sampled, with roughly the same number of sequences per year. Except of a few simulations, the sequence coverage of the HIV epidemic at the time point 2020.0 is below 10%. Sequence coverage may vary between some data sets. The number of sampled individuals is fixed to 1600 (or 3200 in a few data sets).

For some data sets, the simulated viral phylogenies are provided for each transmission chain with at least one sampled individual in file "*_DATEDTREE.newick". Each transmission chain phylogeny is in newick format, one phylogeny per line. Tip names match to individual level identifiers in file "*_metadata.csv". For the remaining data sets, simulated sequences are provided. Sequence names match to individual level identifiers in file "*_metadata.csv".

### *Scenarios and Objectives*

To address the primary objectives, basic epidemiological and prevention parameters of the epidemiological model component were varied. To address the secondary objectives, the sampling frame and the proportion of transmission that originate from outside the simulated population were varied.

The outcome measures described in the general information for participants are challenging to estimate, and may be confounded by several effects. Population growth changes the number of individuals at risk of HIV infection in the simulation. Roll-out of ART changes the generation time in the simulation.

Changes in sampling intensity over time impacts on the distribution of coalescent times in the simulation. Faster within-host evolution influences the evolutionary rate of particular lineages in the simulation.

The start time of the evaluation period is in all simulations January 2015. The end time of the evaluation period is December 2018, one year before the end of the simulation. In a small subset of scenarios, the simulation ends in December 2017 so that sequence data are intensely sampled only for a three-year period. In this subset of scenarios, the evaluation period ends in December 2016. The end of the evaluation period does not coincide with the end of the simulation to facilitate estimation of the objectives with phylogenetic methods (de Silva, Ferguson, & Fraser, 2012).

### *References*

Alizon, S., & Fraser, C. (2013). Within-host and between-host evolutionary rates across the HIV-1 genome. *Retrovirology, 10*, 49. doi: 10.1186/1742-4690-10-49

Cohen, M. S., Dye, C., Fraser, C., Miller, W. C., Powers, K. A., & Williams, B. G. (2012). HIV treatment as prevention: debate and commentary--will early infection compromise treatment-as-prevention strategies? *PLoS Med, 9*(7), e1001232. doi: 10.1371/journal.pmed.1001232

PMEDICINE-D-11-03033 [pii]

Cori, A., Ayles, H., Beyers, N., Schaap, A., Floyd, S., Sabapathy, K., . . . Team, Hptn PopART Study. (2014). HPTN 071 (PopART): a cluster-randomized trial of the population impact of an HIV combination prevention intervention including universal testing and treatment: mathematical model. *PLoS One, 9*(1), e84511. doi: 10.1371/journal.pone.0084511

de Silva, E., Ferguson, N. M., & Fraser, C. (2012). Inferring pandemic growth rates from sequence data. *J R Soc Interface, 9*(73), 1797-1808. doi: rsif.2011.0850 [pii]

10.1098/rsif.2011.0850

Eaton, J. W., Johnson, L. F., Salomon, J. A., Barnighausen, T., Bendavid, E., Bershteyn, A., . . . Hallett, T. B. (2012). HIV treatment as prevention: systematic comparison of mathematical models of the potential impact of antiretroviral therapy on HIV incidence in South Africa. *PLoS Med, 9*(7), e1001245. doi: 10.1371/journal.pmed.1001245

Grabowski, M. K., Lessler, J., Redd, A. D., Kagaayi, J., Laeyendecker, O., Ndyanabo, A., . . . Rakai Health Sciences, Program. (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 Med, 11*(3), e1001610. doi: 10.1371/journal.pmed.1001610

Hayes, R., Ayles, H., Beyers, N., Sabapathy, K., Floyd, S., Shanaube, K., . . . Team, Hptn Study. (2014). HPTN 071 (PopART): rationale and design of a cluster-randomised trial of the population impact of an HIV combination prevention intervention including universal testing and treatment - a study protocol for a cluster randomised trial. *Trials, 15*, 57. doi: 10.1186/1745-6215-15-57

Rambaut, A., & Grassly, N. C. (1997). Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Comput Appl Biosci, 13*(3), 235-238.

Vrancken, B., Rambaut, A., Suchard, M. A., Drummond, A., Baele, G., Derdelinckx, I., . . . Lemey, P. (2014). The Genealogical Population Dynamics of HIV-1 in a Large Transmission Chain: Bridging within and among Host Evolutionary Rates. *Plos Computational Biology, 10*(4). doi: Artn E1003505
Doi 10.1371/Journal.Pcbi.1003505

Walker, P. R., Pybus, O. G., Rambaut, A., & Holmes, E. C. (2005). Comparative population dynamics of HIV-1 subtypes B and C: subtype-specific differences in patterns of epidemic growth. *Infect Genet Evol, 5*(3), 199-208. doi: 10.1016/j.meegid.2004.06.011