Small African Village or ‘Ugandan’ Epidemiological Model Sequence Simulation

An HIV epidemic was simulated in a population of ~8,000 individuals using an individual-based model. The simulations ran for 70 years. The population grows at an exponential rate of around 1% per year, and is about 8,000 individuals at the time of sampling. Partnerships and contact rate depend on the gender and risk group of the individuals, with HIV transmissibility varying through acute, chronic, and AIDS stages. Individuals may have contact with individuals from villages outside the focal population.

Thirteen similar HIV epidemics were simulated. In each simulation, samples were taken over a five year period, starting from year 40. The sampling density varies between simulations. In twelve of the simulations, treatment is introduced to the population at year 40. The speed of treatment ‘roll-out’ (how quickly individuals are put on treatment) may vary between simulations. In the thirteenth simulation, no treatment is introduced to the population (samples are still taken from year 40). Finally, the infectiousness of the individual during the acute phase, and the number of introductions from villages outside the focal population may vary between simulations.

For four of the simulations, sequences have been simulated and provided. Some ‘ancestral’ sequences from before the main sampling period are included for the benefit of participants. For the remaining nine simulations, viral phylogenies have been provided.

Some additional information has been also provided:

- Information on regency of diagnoses has been provided for all sampled individuals. If the infection was acquired no more than one year before sampling, the sample is regarded as ‘recent.’
- Rounded prevalence estimates (#infected/#alive) from before treatment is introduced have been provided. These are taken from the end of year 30, end of year 35, and end of year 39.
- Rounded numbers of infected individuals on treatment have been provided. These are taken from the end of year 40, end of year 41, end of year 42, end of year 43, and end of year 44.

(Continued on next page)

Information about Sampling in the Simulations:

Simulation Numbers	Sample Density (Range)	Format
Vill_00 (no treatment)	15-40%	Phylogeny
Vill_01, Vill_02, Vill_03, Vill_04	15-40%	Sequences
Vill_05, Vill_06, Vill_07, Vill_08	>=40%	Phylogeny
Vill_09, Vill_10, Vill_11, Vill_12	15-40%	Phylogeny

Data File Information

*.nex – Gives the viral phylogeny of the sampled sequences, in Nexus format

*.fasta – Gives simulated *gag*, *pol*, and *env* sequences for the sampled individuals, in fasta format

*_prevalence.csv – Gives the HIV prevalence (rounded) in the population at 10 years, 5 years, and just before treatment starts

*_recency.csv – Gives the recency of all sampled individuals. If their HIV infection was acquired no more than one year before sampling, they have a value of 'YES'

*_treatment.csv – Gives the number of individuals on treatment (rounded) at the end of each year of sampling.

Dating:

As in release one, participants should be aware that sample time may have been blinded. First, meaningless years were used to avoid preconceived bias about what was happening in the HIV epidemic in Africa at any given real date. Second, the sample dates for each time point may have been adjusted to avoid bias based on the relative timing of the samples between simulations. In all simulations, treatment begins at year 40, and samples are taken in years 40 through 44. As in real life, participants do not know beforehand the current dynamics of the epidemic.

Because of this, combining the data from any of the separate samples will give erroneous results.

Sequences:

Each sequence is a concatenated sequence of *gag*, *pol*, and *env*. *Gag* runs from 1-1479bp, *pol* from 1480-4479, and *env* from 4480-6987.

Each sequence is labelled with the user ID, gender, and sample date (in decimal-year format). User IDs are randomly assigned and meaningless.