Small African Village Epidemiological Model: Training Data Set

At the request of participants, a 'training data set' has been prepared for the Village simulations, in line with the training data set provided for the Regional simulations.

The training data set can be found on the PANGEA_HIV '201502' folder on Dropbox, in the 'Village' folder. It is labelled as "Vill_99."

The training data set contains the 'true' phylogenetic tree generated from the transmission information, *and* sequences simulated down this tree. Both of these were generated in the same manner as for the original released datasets. The training data set may be useful to groups to evaluate their phylogenetic reconstruction from the simulated sequences.

The simulation used to generate the training data set was run under the same variables and settings as the originally released data sets, though it is a completely independent run and so should not be assumed to exactly mimic or represent any of the original runs. However, all the details that apply to the original runs hold true with this data set:

- The population grows exponentially at 1% per year, and is contains about 8,000 individuals at the time of sampling.
- Sampling begins at year 40 and continues for 5 years, and intensive treatment also begins at year 40.
- Individuals may contact other villages outside the focal population
- A small number of 'ancestral' sequences from before year 40 are provided both in the phylogeny and in the sequences
- The number of transmissions from other villages, speed of treatment roll out, and infectiousness of the individual during the acute phase may vary, as with the original runs.

For this training data set, between 15-40% of the infected population was sampled.

Please contact Emma Hodcroft (emmahodcroft@gmail.com) if you have any questions about the training data set.