Phylogenetics and Networks for Generalised HIV Epidemics in Africa

## Phylodynamic Methods Comparison Exercise – Appendix Training data for the Regional Simulation

The Feb-15 PANGEA-HIV data sets contain either simulated sequences or the true simulated phylogeny but not both. We were asked if we could provide a training data set that contains both simulated sequences and the true simulated phylogeny in order to fine-tune existing methods in their ability to reconstruct phylogenetic trees from full genome HIV sequence data.

We simulated a training data set, ** <u>omitting some of the complications that we added to the Feb-15 PANGEA-HIV data sets</u> **. One apparent difference to the Feb-15 PANGEA-HIV data sets is that sequences are sampled more uniformly since 2000.

This document describes phylogenetic analyses that we conducted on the Feb-15 training data set.

Root-to-tip distance

We added HXB2 as outgroup, estimated a tree with RaXML, rooted the tree at the HXB2 taxon, removed HXB2, and calculated the root-to-tip distance on among the simulated pol genes:
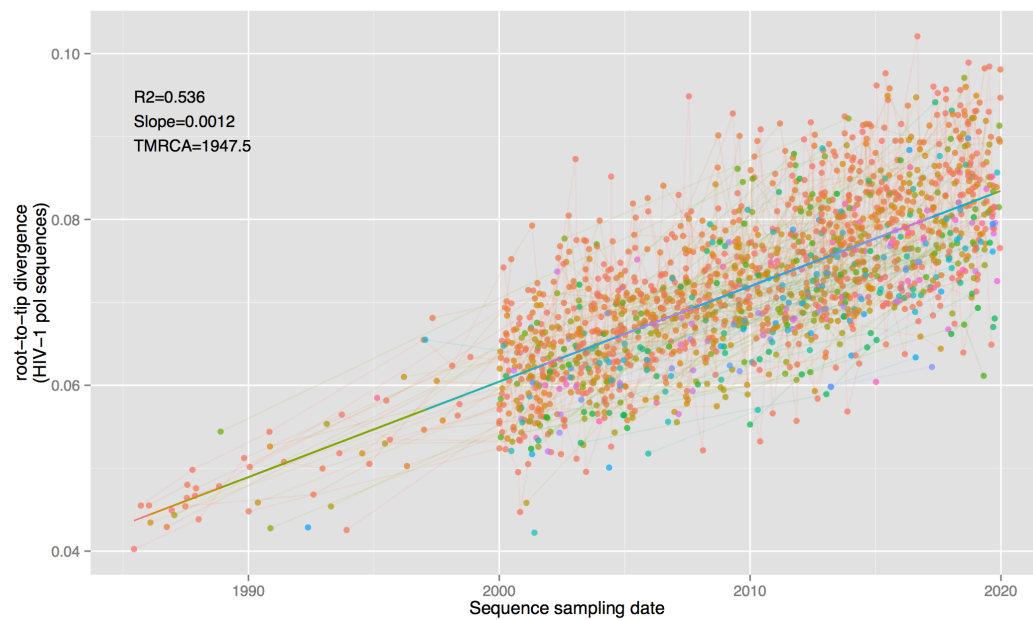
BEAST 1.8 skygrid analysis

We conducted a BEAST analysis using sequences from all simulated transmission chains that have more than 2 viral sequences sampled. The tree was kept fixed at the true simulated tree, and the substitution model used to simulate the sequences was specified.

This is the trace plot of the log likelihood over 10 million iterations:
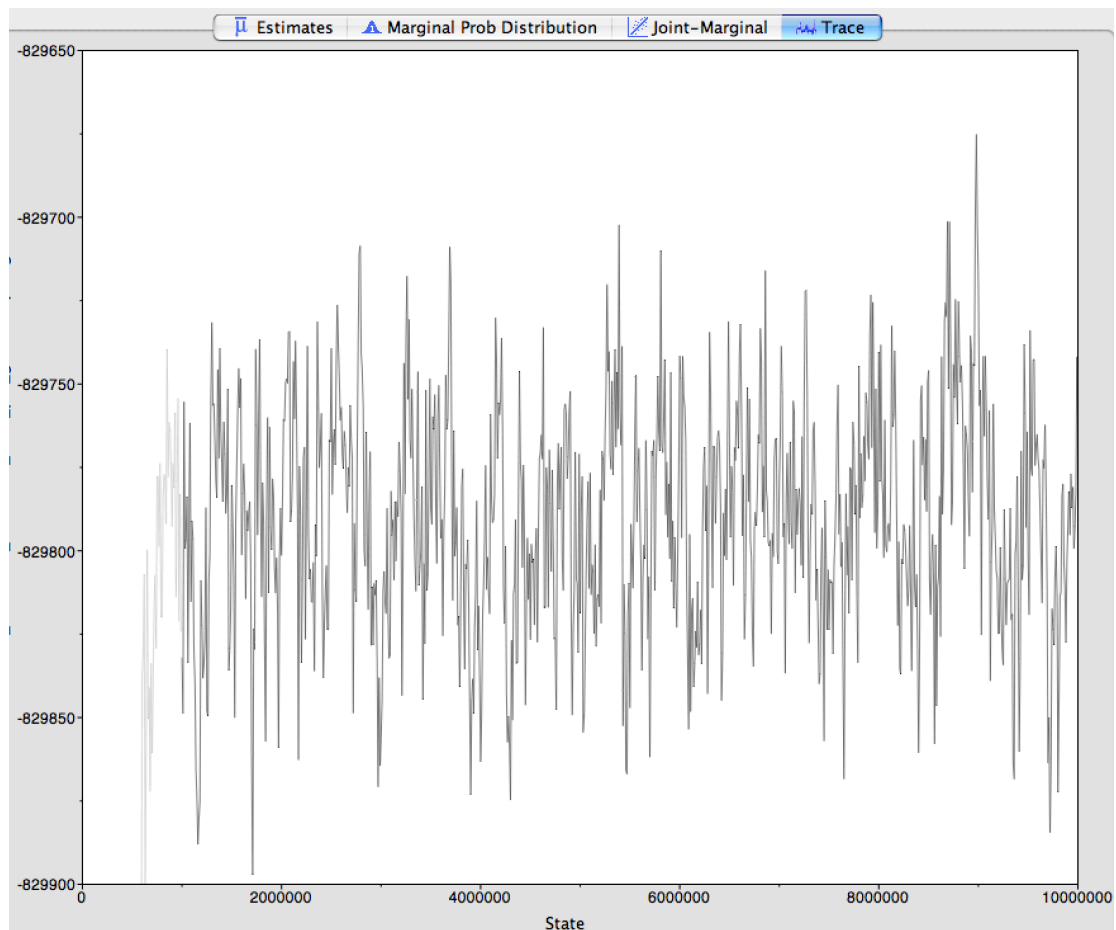
This is the estimated effective population size: