|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Dataset details | | | | | |
| dataset | #samples | earliest sample | latest sample | number of variable sites | where did the data come from? |
| early 2014 | 72 | 2014.455 | 2014.510 | 25 | Gir et al 2014  (Stadler 2014) |
| 2018 public | 83 | 2017.387 | 2018.705 | 1342 | GenBank |
| 2014 Liberia | 245 | 2014.310 | 2015.906 | 2014 | GenBank |
| 2018-2019 | 455 | 2018.436 | 2019.887 | 3658 | Nexstrain |
| large 2014 | 2013 | 2014.258 | 2017.403 | 6749 |  |
| Ebola 2018-2019 dataset🡪subsets | | | | | |
| 1 | 224 | 2018.81 | 2019.88 | 1675 | Nexstrain |
| 2 | 98 | 2018.64 | 2019.67 | 1829 |
| 3 | 133 | 2018.61 | 2019.71 | 316 |

Models:

🡪Strict clock, Isochronous

🡪Strict clock, Heterochronous

🡪Relaxed clock, Isochronous

🡪Relaxed clock, Heterochronous

Beast set up:

🡪Tip dates: (use of sampling times for molecular clock calibration)

-In heterochronous trees, information about the tip dates was included. This was used to calibrate the molecular clock, enabling the estimation of substitution rates over a specific time scale.

-Isochronous trees don’t have sufficient temporal signal, so the calibration of the molecular clock using sampling times is not favoured. In this case the molecular clock provides estimates of the divergence times.

🡪Sites (substitution rate model selection)

HKY model:

-This model predicts different substitution rates between the four nucleotides, that are expressed by the parameter kappa (transitions/transversions). The base frequencies were calculated using empirical data, derived from the prior. A discrete gamma distribution, with 4 rate categories, was used to describe site heterogeneity.

🡪Clocks: (assumes a constant rate of evolution among lineages that can be used to estimate the timing of divergence events)

-The strict clock is the simplest form of molecular clock. It assumes homogenous rates among branches and its only parameter is the rate of evolution (substitution rate/ site). This can be a good model for sequence data for which we expect low rate variation among branches (conspecific individuals).

-The relaxed molecular clock assumes variation among branch specific rates. In this study, an uncorrelated relaxed clock was selected. In this model the rate between neighbouring branches is independent. The branch specific rates were sampled from a lognormal distribution that suggests that the branch specific rates are clustered over a mean value

-Heterochronous trees: If the temporal signal of these trees is sufficient (temporal span of samples is long enough to allow accumulation of genetic variation) the ages of the samples are used for calibration.

🡪Trees:

A coalescent-exponential growth model was chosen for the tree prior. This model suggests that gene variants sampled from a population have originated from a common ancestor and that the population expands overtime.

🡪Priors:

-Kappa (HKY transition/transversion parameter): lognormal distribution

-Alpha: The shape parameter of the gamma distribution. High alpha means that most sites evolve with the same rate and low alpha means that substitution rates among sites are different. An exponential distribution was chosen for this parameter.

-Tree model root height: The root height of the tree is estimated using the tree prior

-Exponential population size: Uniform distribution was chosen with bounds between 0 and 0.5. We specify these bounds because the rate of substitution for viruses falls within 10-2 to 10-5 range, so we don’t expect it to go above 0.5.

- Exponential growth rate: Laplace distribution

-Clock rate (substitution rate): Prior for strict clock models that has a fixed value.

-Ucld mean: Prior for relaxed clock model that has a fixed value.

-Ucld standard deviation: Prior for relaxed clock model with an exponential distribution

🡪Marginal likelihood estimation (MLE)

The posterior and prior are very divergent, with the posterior being more peaked due to the effect of the likelihood.

Generalised steppingstone sampling (GSS): Estimates marginal likelihood by taking many small steps from a reference distribution that covers the same parameter space as the posterior. The closer the reference is to the posterior the, the easier it is to estimate the marginal likelihood. GSS uses samples from a series of power-posterior distributions in between the reference and the posterior, that makes the estimation of the marginal likelihood more accurate. It is considered to be the most accurate estimator of the marginal likelihood.

2018 Public dataset: Best model🡪 Relaxed clock + Isochronous

Early 2014 dataset: Best model🡪Strict clock + Heterochronous

2014 Liberia dataset: Best model🡪Relaxed clock + Heterochronous