Intro:

Ebolavirus is a negative sense single strand RNA virus, with a 19kb genome, that can cause severe infectious disease in humans. The first outbreak occurred in 1976, in the Democratic Republic of Congo (DRC) and was followed by 12 small outbreaks in central Africa. In 2013 an outbreak occurred that lasted until 2016 and was characterised by much higher duration and magnitude than previous events and it was the first reported in west Africa and places outside Africa. During this outbreak there was a concern that ebolavirus is mutating faster, which could lead to the virus becoming more virulent and harder to cure. This assumption was based on computational studies that were carried out early during the epidemic and estimated atypically high rates of mutation of the virus, that were almost two times higher than what had been observed until that time. However, later studies that included a larger number of viral sequences sampled over the whole epidemic, found lower mutation rates and disproved this belief. It has now been proved that those high rate estimations at the beginning of the outbreak were due to the small number of samples and the limitations of the computational methods used, and the fact that evolutionary rates sampled over short periods of time can lead to higher mutation rate estimations as many deleterious mutations have not been weeded out yet by natural selection.

In May of 2018 a new Ebola outbreak was reported in DRC. This epidemic continues to spread and because of this it is hard to assess its magnitude. Despite the fact that we do not yet have information about the entirety of this outbreak, it has been clear since its beginning that it is one of the largest outbreaks that have occurred. Further studies have shown that the epidemiological features as well as the case fatality ratio are similar to other outbreaks that had previously occurred. Sequencing analysis of early samples revealed that the ebolavirus Tumba, which evolves at a slower rate than other ebolavirus variants, was the cause of this outbreak. Although there were some differences found in the rate of evolution between ebolavirus Tuba and ebolavirus Makona, that was responsible for the 2013-2016 outbreak, their intra-outbreak rates were found to be similar.

Datasets:

We obtained nucleotide sequences of the Ebola virus, between the years 2014 and 2019, from different databases, such as GenBank and Nextstrain and from publications (reference). We collected 5 datasets, each of which included 72 to 2013 samples, with specified sampling times, in calendar years, and sampling periods ranging from 0.005 to 3 years (Table 1: dataset information). We used MAFFT to align the sequences and then studied them using Beast v1.10.4, TempEst and LSD to get rate estimates.

Three of the datasets we used had samples collected from the earlier, 2014-2016, epidemic. These included a small dataset, containing 72 samples, collected between May and June of 2014 in Sierra Leonne and a dataset with 245 samples from Liberia, collected between 2014 and 2015. We also obtained a large dataset with 2013 samples, from multiple areas, dated between 2014 and 2017. We divided this large dataset into smaller subsets, to make it easier to analyse and to observe any changes in the mutation rates over time and in different locations. The first set of subsets was created based on the dates and we divided the dataset in 4 groups with a time range of half a year each. We weren’t able to include subsets after 2016 as there was not a sufficient number of samples after that date. The next set of subsets was based on geographic locations and we separated 1302 samples from Sierra Leone, 409 from Guinea, 253 from Liberia and 22 from the Democratic Republic of Congo (DRC). We also analysed two more recent datasets, from the 2018-2019 outbreak, collected from DRC. One of the datasets had 83 samples collected in 2018 and the other contained 455 samples from 2018-2019.

BEAST:

For the dataset analysis in BEAST we selected a coalescent exponential growth tree prior, that implies that the taxa on the phylogenetic tree are paraphyletic. The tree prior parameters included the scaled population size, the prior of which was sampled from an exponential distribution, and the growth rate parameter, who’s prior was drawn from a Laplace distribution. We also specified an HKY substitution model with a gamma distribution accounting for among-site rate heterogeneity. For the determination of the marginal likelihood estimator (MLE), we chose a generalised stepping-stone (GSS) path, with a beta path step distribution.

The datasets were analysed under multiple settings with different clock models that we either calibrated with the use of tip dates (heterochronous) or for which we constrained the samples to be contemporaneous (isochronous). The clock models included a strict clock model, that assumes homogenous rates among branches with the rate of evolution (substitution rate/ site) being its only parameter, and an uncorrelated relaxed clock model that was sampled under a lognormal distribution and suggests that the branch specific rates vary. Thus, there were four models created for each set of sequences: i) strict cock, isochronous, ii) strict clock, heterochronous, iii) relaxed clock, isochronous and iv) relaxed clock heterochronous.

Once the different models were created, the best fitting one for each dataset was chosen, based on the MLE. The results were then analysed in Tracer, in order to draw inferences about the substitution rates within and across datasets.

TempEst:

We analysed the temporal signal of the datasets with a root to tip regression implemented in TempEst, using the sampling times for calibrations. This kind of analysis required the construction of a maximum likelihood tree.

LSD: