A little phylogenetic study on Ebola virus in the 2014 Sierra Leone outbreak

Shofi Andari

Introduction

Although it was first identified in 1976, when Ebola virus striked near Ebola River and infected over 300 people, the first hideous outbreak of Ebola virus disease (EVD) happened in 2014 in West Africa [3, 6]. It at least began spreading in December 2013. Moreover, one of the huge hits at that time occured in Sierra Leone.

The species that was responsible to overtake its host during the outbreaks was Ebola virus (EBOV), formerly *Zaire ebolavirus*. The fatality rate on average is 78 % (I) [6]. This species and the family of filoviruses to which it belongs have an incredible mechanism to disable the immune response and destroy the vascular system. The virus can cause massive inflammation and leaky blood vessels [10] which may damage tissues and lead to hemorrahaging inside or outside the body. These damages might lead to death due to shock and multiple organ failures [11]. There are many studies conducted to understand the evolution of EBOV. It is crucial for gathering how the virus is maintained from one outbreak to another, how it creates such devastation, and how the outbreaks can be lessened in the future [1].

The first case of EBOV in Sierra Leone was found on May 25th. The epidemic happened since then until June the 18th. From this 2014 outbreak, [13] collected and sequenced data from 72 patients in Sierra Leone. They also used phylogenetic trees to study how the virus population structure affected the epidemic. In the analysis, [13] included lengths of incubation and infectious periods estimation. The data were first introduced in [6] from 78 individuals contracted with Ebova virus. [13] classified the outbreak in Sierra Leone as a larger outbreak (with 72 patients) and a smaller outbreak (6 patients) then decided to focus on the larger one.

The sizes of population in particular for RNA viruses may change in complex fashions due to a changing host population, seasonal factor, or publuc health interventions [12]. The coalescent skyline plot [14] then introduced to extend the classical coalescent models to accomodate the arbritrary changing in population sizes. Bayesian skyline plot was applied in BEAST [4] and became the standard models used to reconstruct ancestral dynamics of evolving population [12].

Birth-death skyline model was first introduced in [12] to overcome the limitiation of Bayesian skyline plot: (1) models cannot accomodate incident and prevalence which affect coalescent rates,

and (2) models assume the sample to be small, while in cohort studies in epidemic outbreak the infections sampled may be quite large.

My main goal in this final project is to construct a phylogenetic tree with birth-death process Bayesian model (birth death model, or BDM) [13], apply molecular clock analysis to it (using strict method) and build a skyline plot [12] based on the model. The analysis is done using BEAST2 (`http://beast2.cs.auckland.ac.nz`) [4, 5], `Tracer v1.7.1` and `R 3.5.3`.

Amruta's review: 1. What does your molecular clock reveal? 2. Is there a way to know most virulent/infectious virus sub types through your analysis. 3. You can explain in more detail as why you chose the approach?

Methods

Birth-death models and skyline plot. To model the spread of the outbreak, we assume several parameters: a transmission rate, a becoming-noninfectious rate, and a sampling probability. These parameters could change in a piece-wise constant fashion. By using the BDM, it allows us to assume that transmission and death rates are estimated independently and therefore enables for the first time the estimation of the basic reproduction ratio (R_0) of the pathogen using only sequence data, i.e. there is no use to incorporate the average duration of infection [9]. BDM was developed based on birth-death process which is commonly use in epidemiology modeling, e.g. it is used to study the number of people infected EBOV in a population.

Christian's review: fix the reference and add more results

The following summarizes notations used in this analysis (mostly taken from [8]) :

- R_0: **basic reproduction ratio**. In order to determine whether a contagious disease, such as EVD, can penetrate a population which is in a steady demographic state with susceptible individuals, we define basic reproduction ration as the expected number of secondary cases produced [2]. It is the ratio of transmission rate over becoming-noninfectious rate. The cut-off is $R_0 > 1$ for the disease would be able to invade the population.
- R: **effective reproductive number**. The idea is pretty similar to R_0. The quantities are equal at the start of an apidemic outbreak.
- δ: **rate of becoming a non-infectious**. Individuals are non-infectious if they were treated (or cured) or die.
- s: **probability of sampling** an individual upon becoming non-infectious.
- λ: **rate of transmission** (birth rate)
- μ: viral lineage **death rate**
- ψ: rate of each individual being sampled

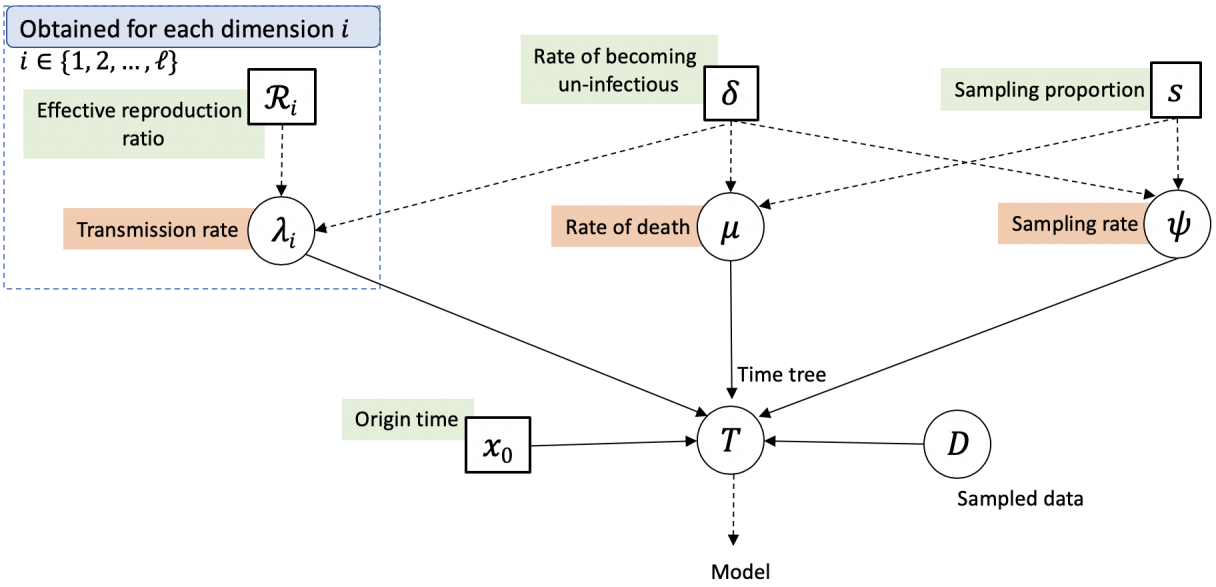


Fig. 1: Probabilistic graphical model of the skyline birth-death process [8, 12].

Birth-death skyline plot can be used as a model of transmission with rate transmission of λ and rate of becoming non-infectious δ. The relationships among parameters is shown in the following figure (Fig.). The estimable parameters are R, δ and s, as the remaining related closely to the other three parameters:

$$\delta = \mu + \psi \tag{1}$$

$$\begin{aligned} R &= \frac{\lambda}{\mu + \psi} = \frac{\lambda}{\delta} \Rightarrow \lambda = R\delta \\ s &= \frac{\psi}{\mu + \psi} = \frac{\psi}{\delta} \Rightarrow \psi = s\delta \end{aligned} \tag{2}$$

Thus, we can modify Eq. 1 to be

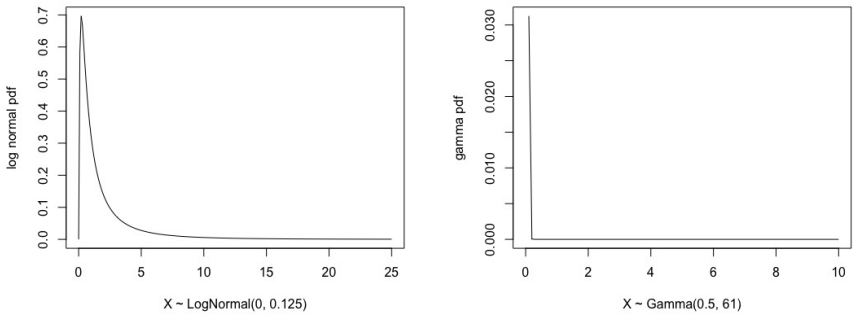
$$\mu = \delta - \psi = \delta - s\delta = \delta(1 - s) \tag{3}$$

It is appropriate to use `Birth Death Skyline Serial` since the samples were taken thorugh times. The skyline visualization was carried out using `bdskytools` package in `R` (it is not available in CRAN, so visit `https://rdrr.io/github/laduplessis/bdskytools/` to obtain the package). Skyline plot is a smooth line along reproductive numbers' medians and based on their highest posterior densities (HPDs).

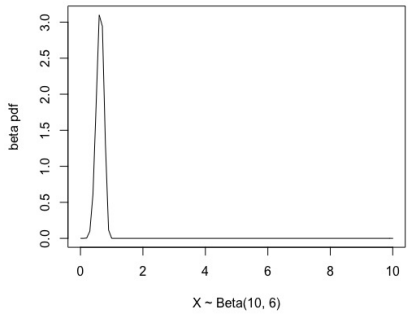
The priors and initialization. The prior of effective reproduction number R is LogNormal(0, 0.125) with assuming dimension ℓ is equal to 3, i.e. the effective reproductive number changed two times after the start of the epidemic. Therefore we would get 3 sets of R_i. Lognormal is a good prior distribution since it will always be positive valued, as a rate cannot be negative (Fig. 2(a)). The mean value is set to be 0, meaning the median is 1. The variance is set to 1.25, therefore most of the weight were placed below 7.815, that is under 95% quantile (`R> qlnorm(0.95, 0, 1.25)`).

For the rate of becoming uninfectious, δ, we use a gamma prior with $\alpha = 0.5$ and $\beta = 61$. It is another way to restrict nonnegative values of rate.

Sampling probability s has the prior under Beta distribution with $\alpha = 10$ and $\beta = 6$. Beta distribution is used for the prior on s it is one flexible class of distributions that are only defined between 0 and 1. That way it is easier to be used for proportions.



(a) (b)



(c)

Fig. 2: Probability function under prior: (a) lognormal(0, 1.25) for reproduction ratio R; (b) gamma(0.5, 61) for the rate of becoming uninfectious δ; and (c) beta(10, 6) for sampling probability s.

HKY [7] is used for the evolution model with both transition and transversion parameters, κ, are under LogNormal distribution and initialized with the value of 2. The sites substitution rate is set to follow gamma distribution under four categories.

Since the data were sampled from an EBOV epidemic in a single location, Sierra Leone, we can assume that there is no different rates of substitution for different lineages. Hence the default option for molecular clock analysis is chosen (strict method). Prior for clock rate or subtitution rate is normal with mean 0.001984 and standard deviation 0.000459 [6, 13]. The origin inferred by the BDM skyline is the time the first person in the outbreak was first infected [13]. Thus, it is not the same with the time of the most recent common ancestor of the tree (TMRCA).

Results

We set normal distribution as the prior for origin time (x_0). The posterior distributed as in Fig. with mean 0.0824.

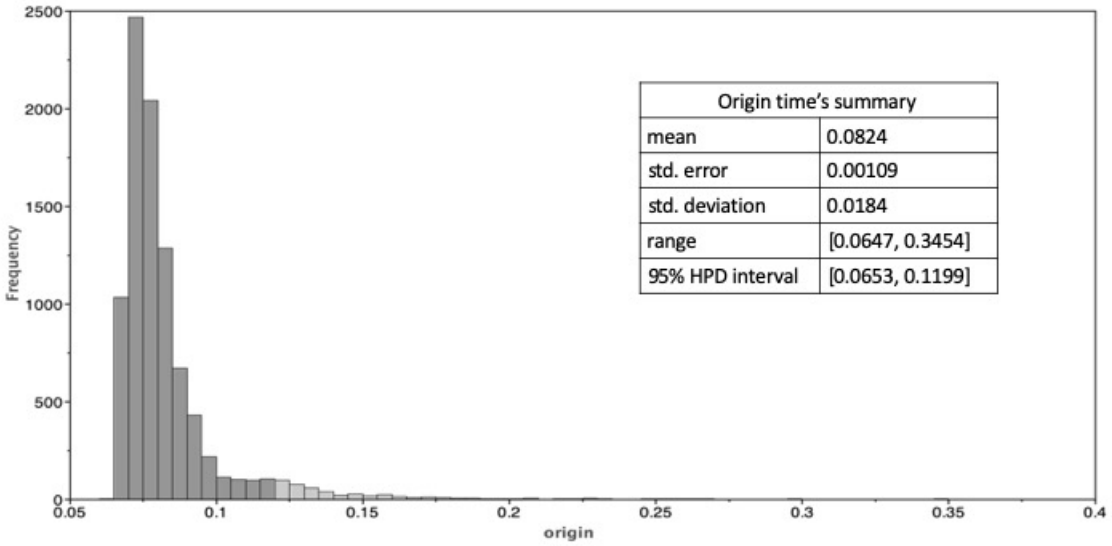


Fig. 3: Distribution of origin time

According to BDM we have designed in Fig. , the tree is dependent to all parameters and the data. The estimation of its height is shown in Fig. .

In Fig. 5(a), we can see the distribution of initial value of reproductive ratio and the other two dimensions. While the skyline plot is depicted in Fig. 5(b).

The annotated tree was drawn from 8001 trees (Fig.).

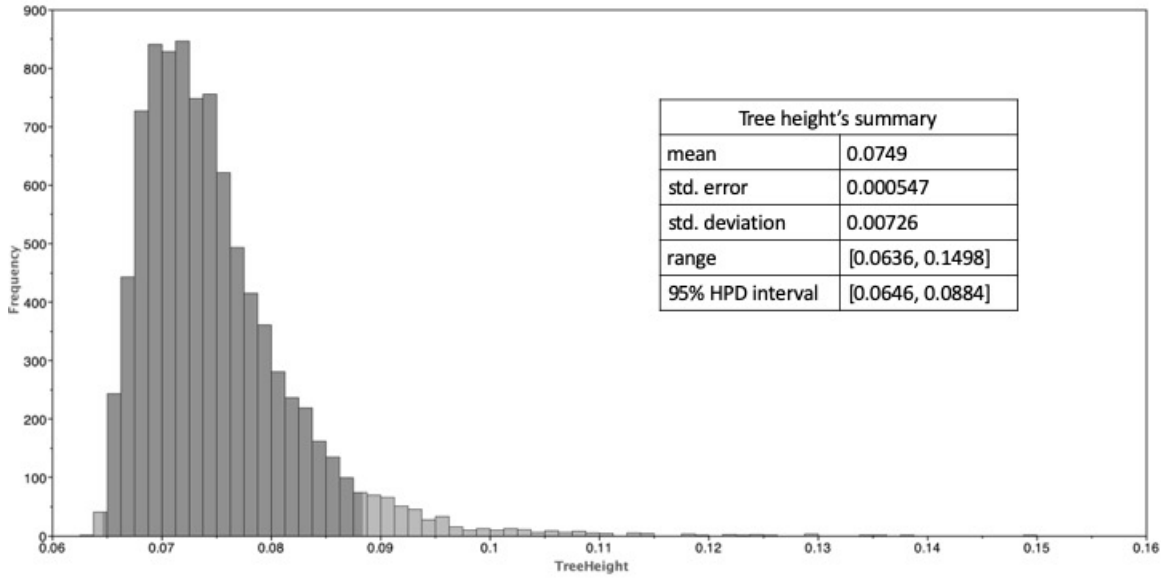
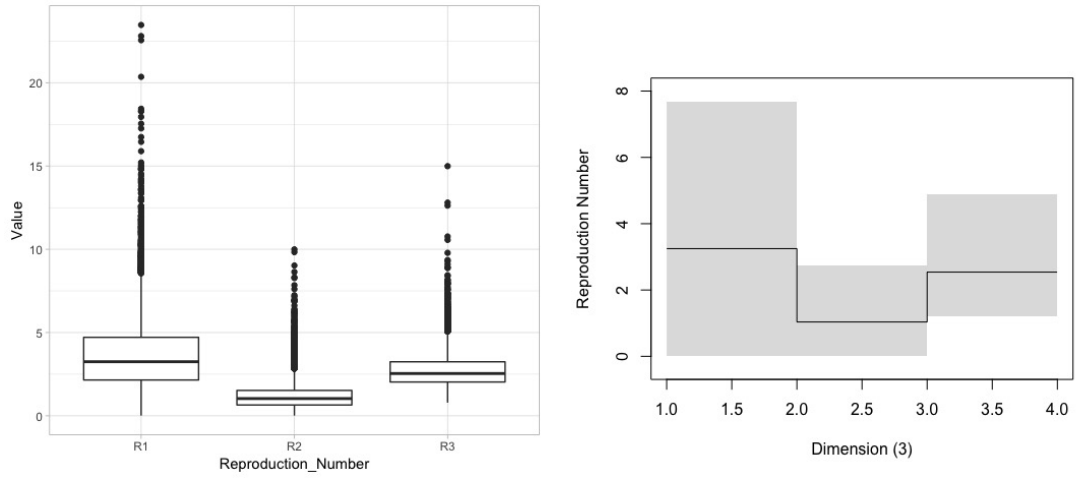


Fig. 4: Distribution of tree height



(a) (b)

Fig. 5: Reproduction numbers: (a) Box-Whisker plots of reproduction ratios, (b) Plot of effective reproduction number HPD across all dimensions

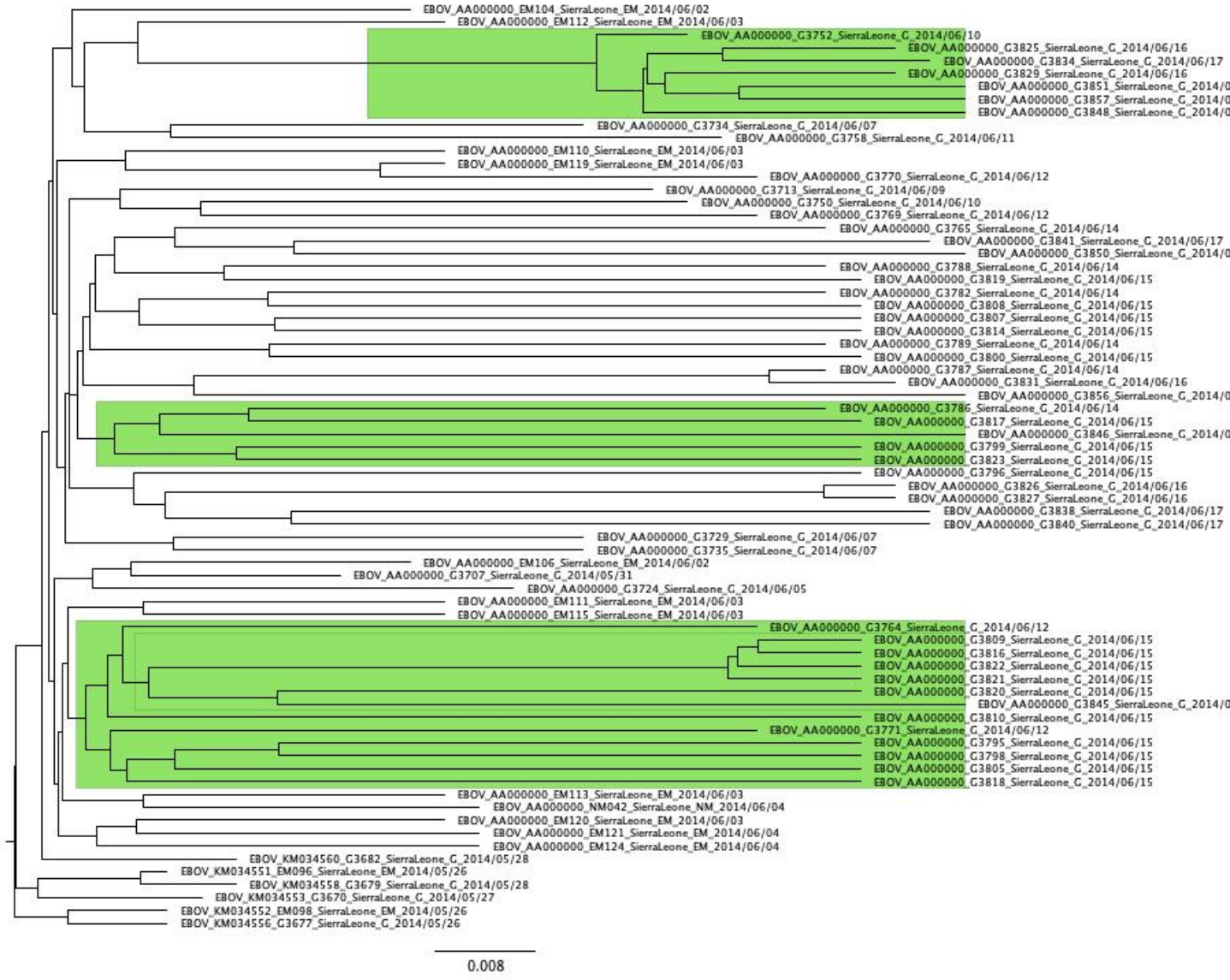


Fig. 6: Our tree!

Discussion

We can infer the start of epidemic to be $\frac{0.0824}{1/365} = 30.076 \approx 30$ days before the most recent sample date. Since the latest sample was taken on June 18th, then the first incident of the outbreak is estimated to be on May 19th. The interval of 95% highest posterial density of this parameter is between April 26th and May 26th.

The *TreeHeight* parameter represents the date of the root or MRCA for the 72 sampled sequences. The mean of the height is $\frac{0.0749}{1/365} = 27.338 \approx 27$ days.

In Fig. 5(a), we can see that the medians of posteriors for reproduction ratios are not extremely different among three values. This is also supported by the skyline we obtained using `bdskytools` in `R 3.5.3`. In the analysis, it was difficult to get a smooth skyline plot since the changes among the R_i's are not noticeable.

Through `Tree Annotator`, from `BEAST`, we got the phylogenetic tree from 72 sequences of EBOV (Fig.). The highlighted clades are the most recent samples.

References

[1] C. J. Brown et al. "New Perspectives on Ebola Virus Evolution". In: *PLoS ONE* 11 (8): e0160410 (2016). DOI: 10.1371/journal.pone.0160410.

[2] O. Diekmann, J. A.. P. Heesterbeek, and J. A. J. Metz. "On the definition and the computation of the basic reproduction ratio R_0 in models for infectious diseases in heterogeneous populations". In: *Journal of Mathematical Biology* 28 (1990).

[3] T. S. Do and Y. S Lee. "Modeling the spread of Ebola". In: *Osong Public Health Res Perspect* 7(1) (2016). DOI: http://dx.doi.org/10.1016/j.phrp.2015.12.012.

[4] A. J. Drummond and A. Rambaut. "BEAST: Bayesian evolutionary analysis by sampling trees." In: *Molecular Biology and Evolution* 7.214 (2007).

[5] A. J. Drummond et al. "Bayesian phylogenetics with BEAUti and the BEAST 1.7". In: *Molecular biology and evolution* 29(8) (2012). DOI: doi:10.1093/molbev/mss075.

[6] S. K. Gire et al. "Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak". In: *Science* 345 (2014).

[7] M. Hasegawa, H. Kishino, and T. Yano. "Dating of the human-ape splitting by a molecular clock of mitochondrial DNA". In: *J Mol Evol* 22.2 (1985), pp. 160–74.

[8] T. Heath and T. Stadler. "Estimating Epidemiological Parameters of an Ebola Outbreak using BEAST2". In: (2014). URL: http://phyloworks.org/workshops/Ebola_BEAST2_Exercise.pdf.

[9] S. Kouyos R.and Bonhoeffer et al. "Estimating the Basic Reproductive Number from Viral Sequence Data". In: *Molecular Biology and Evolution* 29.1 (Sept. 2011), pp. 347–357. ISSN: 0737-4038. DOI: 10.1093/molbev/msr217. eprint: http://oup.prod.sis.lan/mbe/article-pdf/29/1/347/24854347/msr217.pdf. URL: https://doi.org/10.1093/molbev/msr217.

[10] PLOS. "An Ebola virus protein can cause massive inflammation and leaky blood vessels". In: *ScienceDaily* (2014). URL: www.sciencedaily.com/releases/2014/11/141120141654.htm.

[11] K. Servick. "What does Ebola actually?" In: *Science* (2014). URL: https://www.sciencemag.org/news/2014/08/what-does-ebola-actually-do.

[12] T. Stadler et al. "Birth-death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV)". In: *PNAS* 110 (2013). DOI: 10.1073/pnas.1207965110.

[13] T. Stadler et al. "Insights into the Early Epidemic Spread of Ebola in Sierra Leone Provided by Viral Sequence Data". In: *PLOS Currents Outbreaks* Edition 1 (2014). DOI: 10.1371/currents.outbreaks.02bc6d927ecee7bbd33532ec8ba6a25f.

[14] K. Strimmer and O. G. Pybus. "Exploring the Demographic History of DNA Sequences Using the Generalized Skyline Plot". In: *Molecular Biology and Evolution* 18.12 (Dec. 2001), pp. 2298–2305. ISSN: 0737-4038. DOI: 10.1093/oxfordjournals.molbev.a003776. eprint: http://oup.prod.sis.lan/mbe/article-pdf/18/12/2298/23449256/mbev_18_12_2298.pdf. URL: https://doi.org/10.1093/oxfordjournals.molbev.a003776.