EBOV analysis example for BEAST 2.5

Louis du Plessis

October 31, 2018

Abstract

Phylodynamics example for BEAST 2.5 paper using data from the West African Ebola dataset. The example uses a sampled-ancestor birth-death skyline to infer population dynamics. This can also double as a bModelTest example.

Contents

1 Tree models for unstructured populations

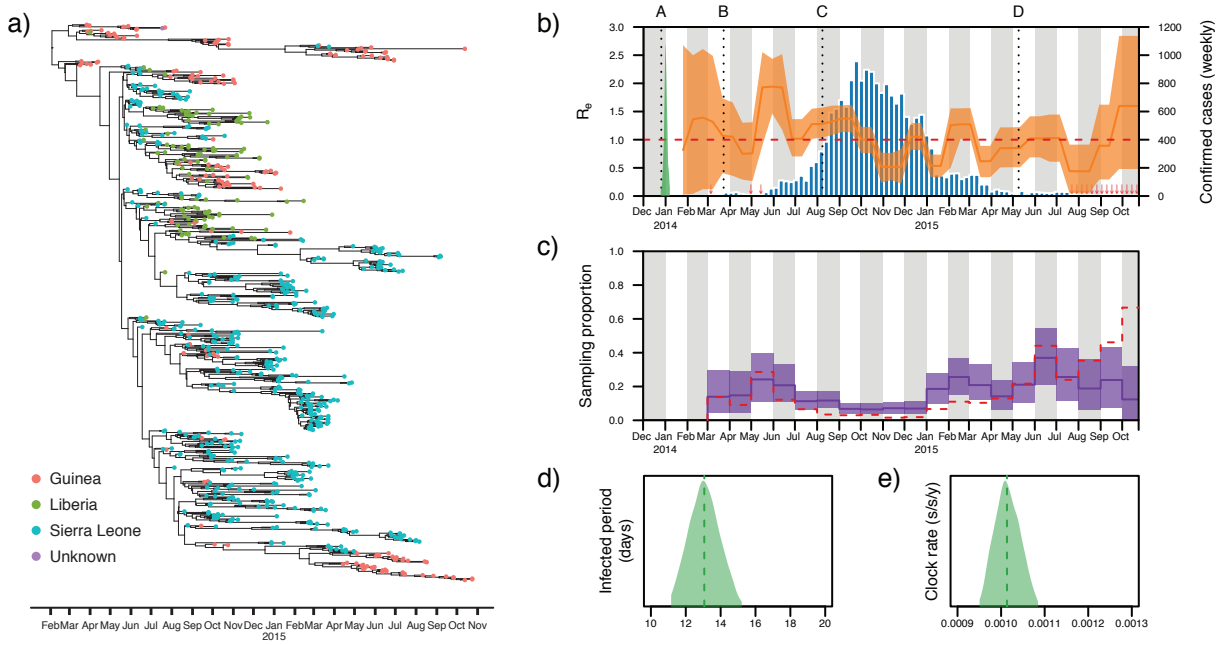


Figure 1: Birth-death skyline (bdsky) analysis of the 2013–2016 West African Ebola virus disease epidemic. **(a)** The maximum clade credibility tree of the 811 sequences used in the analysis. **(b)** The median posterior estimate of the estimated effective reproductive number (R_e) over time is shown in orange, with the 95% highest posterior density (HPD) interval in orange shading. The red dotted line indicates the epidemic threshold ($R_e = 1$). If R_e is below this threshold the epidemic has reached a turning point and is no longer spreading. The posterior distribution of the origin time of the epidemic (t_0) is shown in green. The number of laboratory-confirmed cases per epiweek is shown in blue. Red arrows indicate weeks with fewer than 10 confirmed cases. The dotted line at A indicates the onset of symptoms in the suspected index case (WHO Ebola Response Team, 2016). The dotted lines at B and C indicate the dates at which the WHO declared an Ebola virus disease outbreak in Guinea and a Public Health Emergency of International Concern (PHEIC), respectively. The dotted line at D indicates the first time any of the three countries with intense transmission (Liberia) was declared Ebola free following 42 days without any new infections being reported (new cases were subsequently detected in Liberia in June 2015). **(c)** The median posterior estimate of the monthly sampling proportion is shown in purple, with the 95% HPD interval in purple shading. The red dashed line indicates the number of sampled sequences in the dataset, divided by the number of laboratory-confirmed cases, for each month in the analysis. This serves as an empirical estimate of the true sampling proportion. The posterior distributions and medians (dashed lines) of the infected period and the mean clock rate (truncated at the 95% HPD limits) are shown in panels **(d)** and **(e)**.

In epidemiological investigations the birth-death model can be reparameterised by setting the rate of becoming noninfectious, $\delta = \mu + \psi r$ (the total rate at which lineages are removed), the effective reproductive number, $R_e = \lambda/\delta$, and the sampling proportion $p = \psi/\delta$ (the proportion of removed lineages that are sampled). Figure 1 shows the posterior estimates from a bdsky analysis of the 2013–2016 West African Ebola epidemic. Estimates are based on the coding regions of 811 sequences sampled through October 24, 2015, representing more than 2.5% of known cases. There is evidence that hospital-based transmission and unsafe burials contributed infections to the epidemic (Whitty et al., 2014), thus the sampled ancestor package was used to account for some percentage of patients continuing to transmit the virus after being sampled (by allowing r to be less than 1). R_e was allowed to change over 20 time intervals, equally-spaced between the origin of the epidemic (t_0) and the time of the most recent sample, while the the sampling proportion was estimated for every month from March 2014 onwards (when an Ebola virus disease outbreak was declared and the first samples collected). The estimated origin time of the epidemic coincides with the onset of symptoms in the suspected index case on December 26, 2013 (WHO Ebola Response Team, 2016). Estimates of R_e are consistent with WHO estimates (WHO Ebola Response Team, 2015), based on surveillance data alone, but with greater uncertainty. For the majority of the period between mid-May and October 2014 R_e is estimated to be above 1, consistent with the observation that September 2014 was the turning point of the epidemic and that case incidence stopped growing in October (WHO Ebola Response Team, 2015). After peak incidence was reached during the last week of September 2014, R_e estimates drop below 1 during October and November 2014 and then fluctuates around 1 during 2015 as transmissions persisted in some areas, due to a combination of unwillingness to seek medical care, unsafe burials and imperfect quarantine measures (WHO Ebola Response Team, 2016). R_e estimates before May 2014 and after August 2015 have a large amount of uncertainty attached to them, due to the small amount of sequences sampled during these time periods. Trends in sampling proportion estimates follow empirical estimates based on the number of confirmed cases, however the sampling proportion is overestimated during the period of intense transmission, which suggests the existence of transmission chains not represented in the sequence dataset. In the final two months of the study period the sampling proportion is underestimated, which may indicate ongoing cryptic transmission during this period, but may also be indicative of a model bias resulting from the remaining transmission chains at this time being highly isolated from each other, which is not taken into account by the model.

2 Substitution models

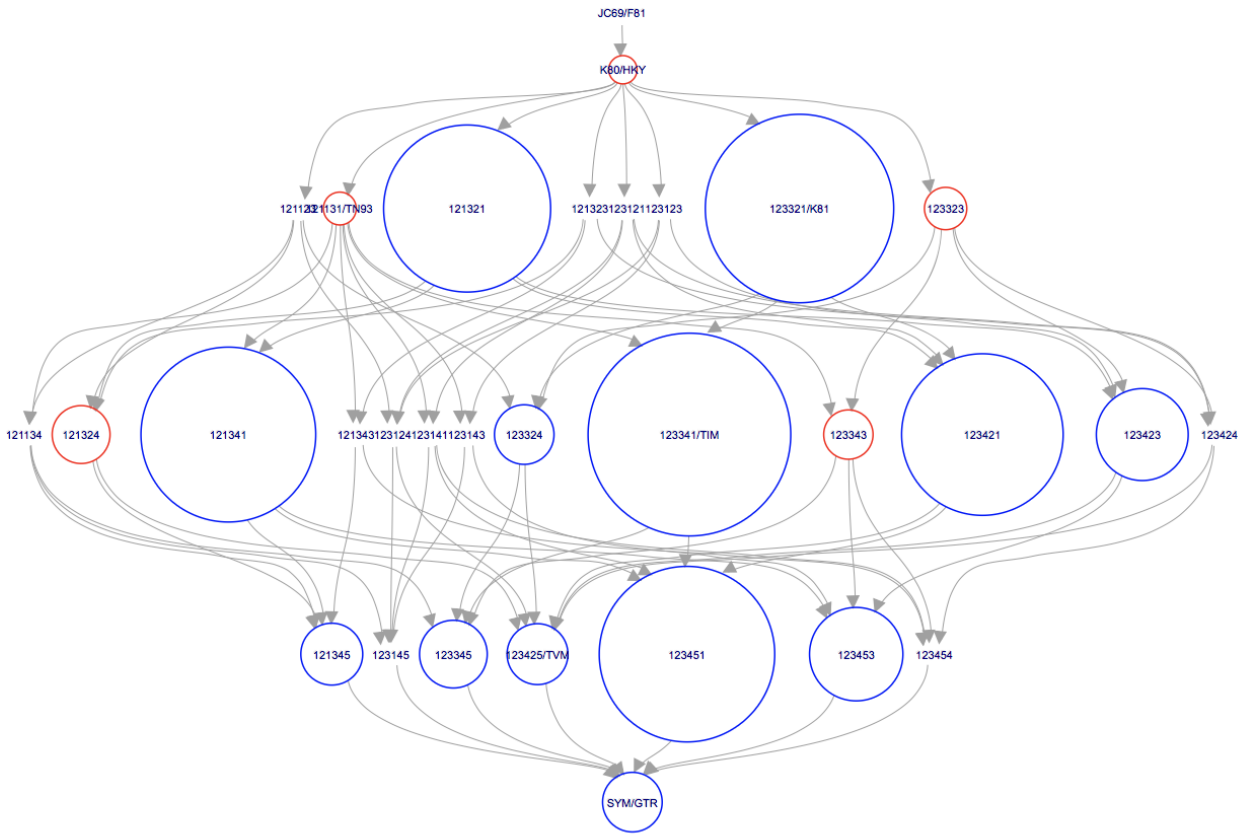


Figure 2: Posterior distribution of substitution models from an analysis of the 2013–2016 West African Ebola virus epidemic. Each circle represents a substitution model indicated by a six digit number corresponding to the six rates of reversible substitution models. In alphabetical order, these are A→C, A→G, A→T, C→G, C→T, and G→T, which can be shared in groups. The six digit numbers indicate these groupings, for example 121121 indicates the HKY model, which has shared rates for transitions and shared rates for transversions. Here, only models are considered that are reversible and do not share transition and transversion rates (with the exception of the Jukes Cantor model). Other substitution model sets are available. Links between substitution models indicate possible jumps during the MCMC chain from simpler (tail of arrow) to more complex (head of arrow) models and back. There is no single preferred substitution model for this dataset, as the posterior probably is spread over a number of alternative substitution models. Blue circles indicate the 13 models contained in the 95% credible set, red are outside, and models without circles have neglegible support. In addition, the analysis indicated 100% posterior probability for gamma-distributed rate heterogeneity across sites and unequal base frequencies.

Figure 2 shows the posterior distribution resulting from a bModelTest analysis of substitution models for 14,517 nucleotides from the coding regions of 811 EBOV sequences sampled during the 2013–2016 West African Ebola virus epidemic. Each circle represents a substitution model indicated by a six digit number corresponding to the six rates of reversible substitution models (see Figure 2 caption for more details).

A Materials and Methods

A.1 Sequencing and surveillance data

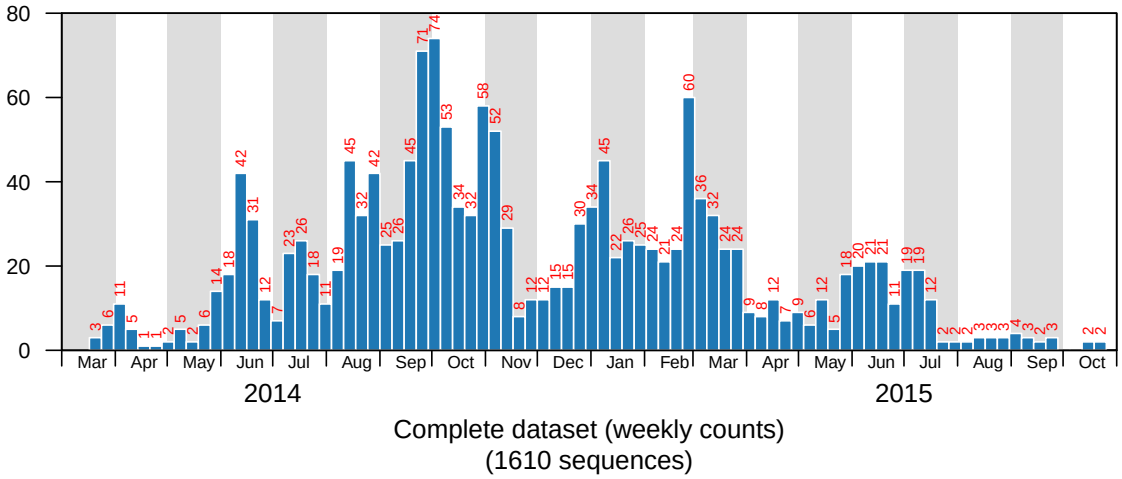
I downloaded 1,610 EBOV sequences, sampled between 17 March 2014 and 24 October 2015, used in the analyses presented in Dudas et al. (2017)[1]. I extracted the coding regions of the sequences, resulting in an alignment of 14,517 bp. Since no sites in the alignment contained more than 419 unknown bases (26%), no sites were excluded from the alignment. I further subsampled the dataset to 811 sequences (approximately 50%). Subsampling was stratified by month to maintain the approximate sampling proportion over time (i. e. within each month 50% of sampled sequences were removed at random). In order to prevent a loss of phylogenetic signal, all sequences from months where less than 5 sequences were sampled were included in the final dataset. The weekly and monthly numbers of sampled sequences in the complete and subsampled datasets are shown in Figure S1 and S2. The geographic distribution of sequences in the complete and subsampled datasets are shown in Figure S3 and S4.

I downloaded weekly probable and confirmed numbers of newly reported cases for Guinea, Liberia and Sierra Leone from the WHO website[2]. This included reported cases up to the week starting on 8 May 2016. In weeks where data are available from the patient database and WHO situation reports I use the maximum of the two counts. In all analyses I use only the number of laboratory confirmed cases.

A.2 Phylodynamic analylses

The serially sampled birth-death skyline model (Stadler et al., 2013) was used as a tree-prior in the analyses. A lognormal prior distribution with mean 0 and standard deviation 1.25 was placed on R_e, which was allowed to vary over 20 time intervals, equally spaced between the origin and the time of the most recent sample (October 24, 2015). The sampling proportion was estimated for independently for every month between March 2014 and October 2015. A Beta-distributed prior with $\alpha = 2$ and $\beta = 10$ was placed on the sampling proportion for each month. Before March 2014 the sampling proportion was set to 0, since no sampling effort was made before this time. A normally-distributed prior with mean at January 1, 2014 and standard deviation of 1.5 months was placed on the origin parameter, which was further bounded to be after

[1] `https://github.com/ebov/space-time.git`, downloaded on 8 August 2018.
[2] `http://apps.who.int/gho/data/node.ebola-sitrep`, downloaded on 10 June 2016.



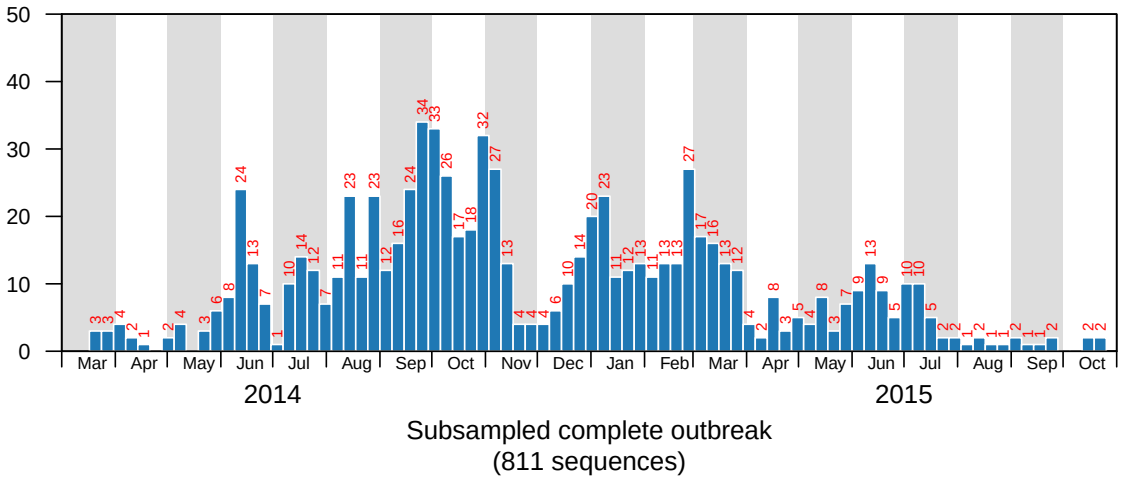
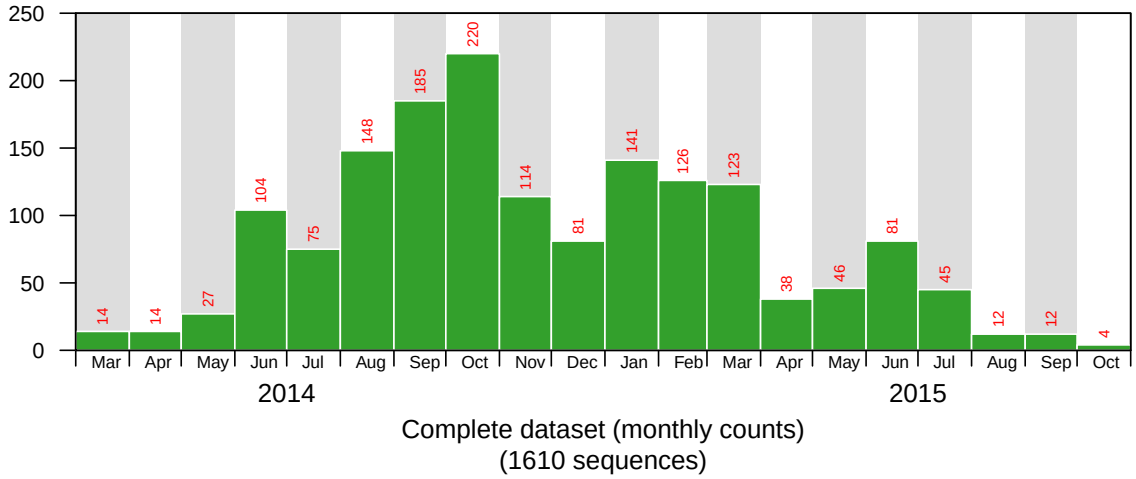


Figure S1: Weekly numbers of sampled sequences in the complete and subsampled datasets.



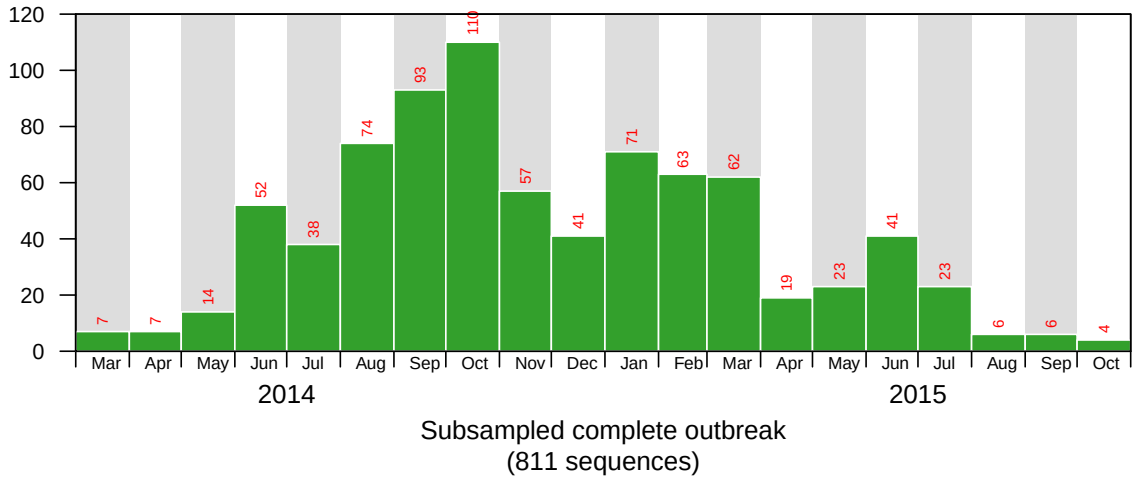
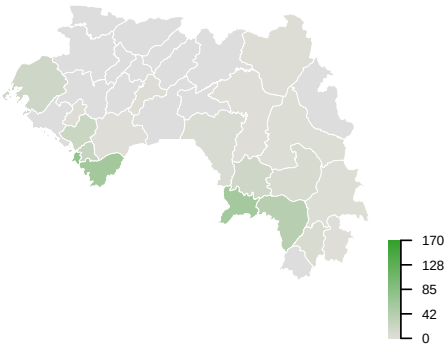
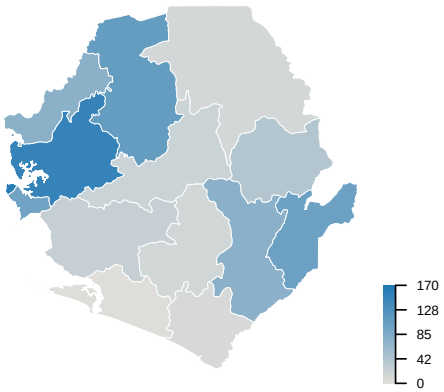


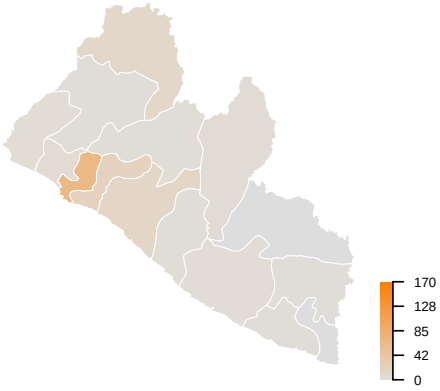
Figure S2: Monthly numbers of sampled sequences in the complete and subsampled datasets.

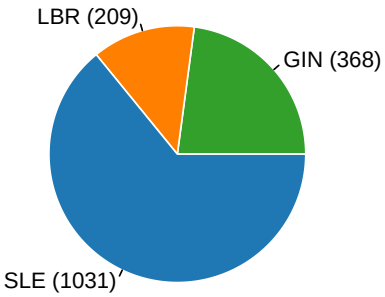


(a) Guinea (GIN)



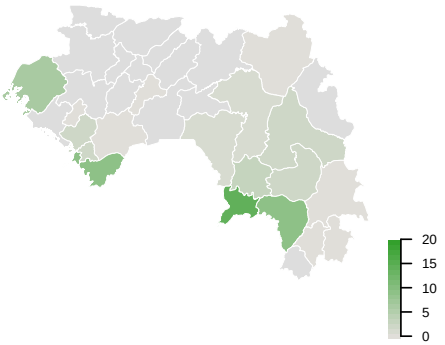
(b) Sierra Leone (SLE)



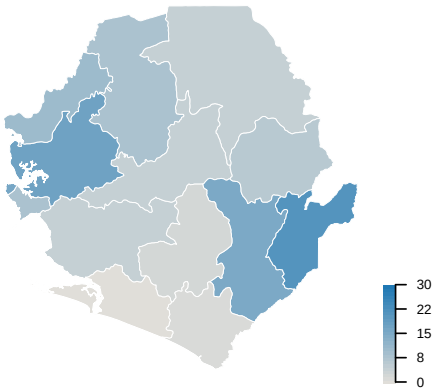


(c) Liberia (LBR)

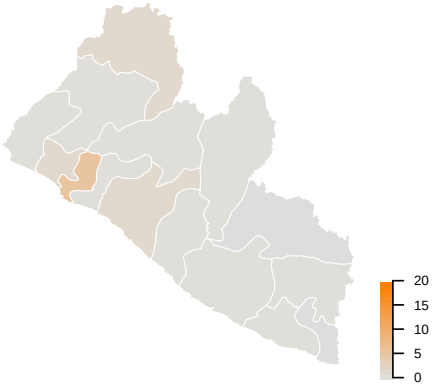
Figure S3: Geographic distribution of samples in the complete dataset.

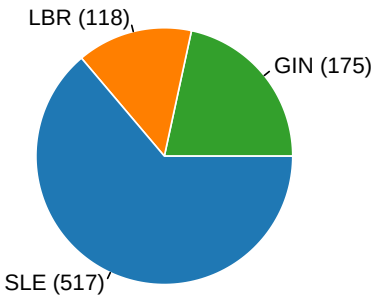


(a) Guinea (GIN)



(b) Sierra Leone (SLE)





(c) Liberia (LBR)

Figure S4: Geographic distribution of samples in the subsampled dataset.

December 1, 2013. A Beta-distributed prior with $\alpha = 5$ and $\beta = 2$ was placed on r (the removal probability) and the sampled ancestor package (Gavryushkina et al., 2014) was used to allow sampled ancestors in the estimated tree. The bModelTest (Bouckaert and Drummond, 2017) package was used to perform Bayesian model averaging across all time-reversible substitution models with a transition/transversion ratio split, while simultaneously estimating support for gamma-distributed rate heterogeneity across sites and unequal base frequencies. An uncorrelated lognormal relaxed clock model (Drummond et al., 2006) was used to estimate the substitution rate. A lognormal prior with mean 1.2×10^{-3} s/s/y (in real space) and standard deviation 0.05 s/s/y was placed on the mean molecular clock rate parameter.

All analyses were performed in BEAST v2.5.0 (Bouckaert et al., 2014). I computed 8 independent MCMC chains of 100 million steps and sampled parameters and trees every 20,000 steps. Tracer v1.7.1 ("Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7") was used to check convergence and Logcombiner v2.5.0 was used to combine and subsample chains, removing 25% as burn-in and resampling parameters every 100,000 steps. Treeannotator v2.5.0 was used to produce the maximum clade credibility tree, from the combined and subsampled posterior tree distribution, using median heights.

Figures were produced using the R software platform using in-house scripts and the R-package bdskytools (available at `https://github.com/laduplessis/bdskytools`) and ggtree (Yu et al., 2017). Further details are available at `https://github.com/laduplessis/Beast2Example`.

B Comparison to other studies

Reproductive number estimates

Estimates of R_e are always below 3 and fall between 1 and 2 for the majority of the growth phase of the epidemic. This is consistent with nearly all estimates of R_0 for the West African Ebola virus disease epidemic (Chretien, Riley, and George, 2015). Initial estimates of R_e between February and May 2014 are highly uncertain, but show a decreasing trend. This is consistent with observations that the number of cases appeared to be declining by May 2014, which prompted the initial epidemic response to be scaled back (Coltart et al., 2017). This also agrees with WHO estimats of R_e from surveillance data, showing a decline over March and April 2014, and a very low R_e estimate in Liberia before May 2014 (WHO Ebola Response Team, 2015) (also see Figure S5). There is no estimate for R_e from surveillance data in Sierra Leone before

25 May 2014, when the first cases were reported.

The observed increase in the estimated R_e in May coincides with the start of the epidemic in Eastern Sierra Leone, which resulted in a large increase in the number of reported cases (Coltart et al., 2017; WHO Ebola Response Team, 2015). R_e estimates are above 1 for most of the period between mid-May and end-September 2014, coinciding with the periods of exponential increases in the number of cases in all three heavily affected countries (WHO Ebola Response Team, 2016). Similarly, WHO estimates of R_e are above 1 for most of this period in all three countires (WHO Ebola Response Team, 2015).

After peak incidence in the last week of September 2014, R_e estimates begin to fall until the end of 2014. Aside from a small increase in February 2015, R_e estimates remain below 1 or fluctuate around 1 for the remainder of the study period. The increase in R_e in February 2015 coincides with a spike in incidence in Guinea (WHO Ebola Response Team, 2015). After August 2015 the estimates become very uncertain, likely due to the small number of lineages in the tree after this time.

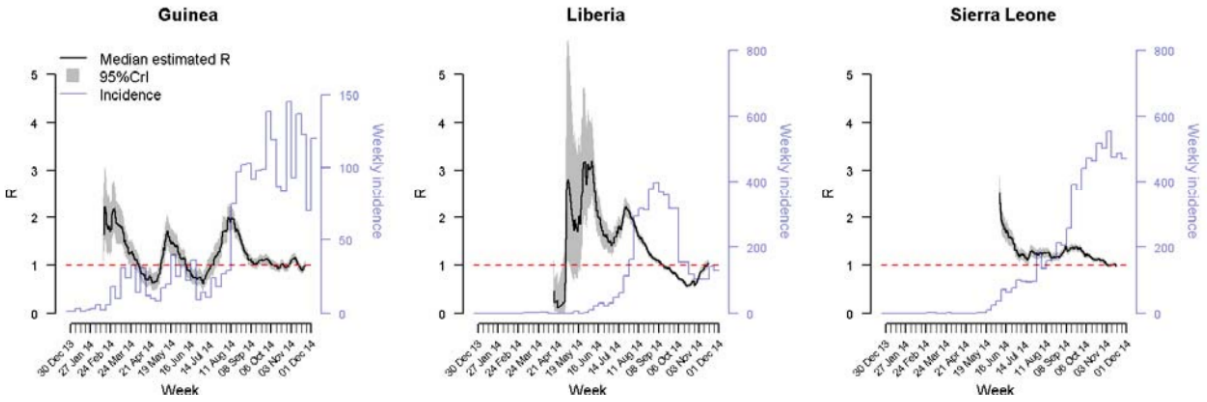


Figure S5: **[From (WHO Ebola Response Team, 2015) (supplement), reproduced *without* permission]** Shown are the estimated case reproduction numbers in Guinea, Liberia and Sierra Leone on the basis of data reported through December 7, 2014 for Guinea and November 30, 2014 in Liberia and Sierra Leone. R_e was estimated over sliding 4-week windows and plotted at the midpoint of these windows.

Other parameters

Trends in sampling proportion estimates follow empirical estimates based on the number of confirmed cases, however the sampling proportion is overestimated during the period of intense transmission, which suggests unsampled transmission chains not represented in the dataset. In the final two months of the study period

the sampling proportion is underestimated, which may indicate ongoing cryptic transmission during this period, but may also be indicative of a model bias resulting from the remaining transmission chains at this time being highly isolated from each other, which is not taken into account by the model.

The estimated origin time of the epidemic coincides with the onset of symptoms and the death of the suspected index case (WHO Ebola Response Team, 2016) on December 26 and 28, 2013, respectively (although earlier reports placed it at the start of December 2013 (Baize et al., 2014)).

The median tMRCA estimate is February 16, 2014 (95% HPD interval Jan 22–March 3, 2014), which overlaps with estimates reported in Dudas et al. (2017) using all 1,610 sequences (95% HPD interval December 16, 2013–February 20, 2014).

The median estimate for the infected period, time from being infected to losing infectiousness (i. e. incubation + infectious periods) is 13 days (95% HPD interval 11.15–15.21 days). This period includes the incubation and infectious periods. This time period should roughly correspond to the generation time (when assuming that most infected patients do not cause multiple secondary infections, which agrees with $R_e < 2$), the time from infection of an index case to infection in a secondary infection. The generation time is difficult to estimate, but follows the same distribution as the serial interval, the time interval between symptom onset in an index case and symptom onset in a secondary case (WHO Ebola Response Team, 2014). WHO estimates of the serial interval is 13 days (IQR 8–18 days) (WHO Ebola Response Team, 2015), which agrees with the bdsky estimates for the infected period.

The median posterior estimate of the mean clock rate was 1.01×10^{-3} s/s/y (95% HPD interval $0.9 - 1.08 \times 10^{-3}$ s/s/y). This is slightly slower than estimates using the complete dataset of 1,610 sequences and the coding and noncoding regions of the genome. Holmes et al. (2016) reported a median estimate of 1.2×10^{-3} s/s/y (95% HPD interval $1.13 - 1.27 \times 10^{-3}$). This is to be expected, since there are more substitutions in the noncoding part of the genome, resulting in a faster clock rate when using both coding and noncoding regions.

References

Baize, Sylvain et al. (2014). "Emergence of Zaire Ebola Virus Disease in Guinea". In: *New England Journal of Medicine* 371.15, pp. 1418–1425.

Bouckaert, Remco R and Alexei J Drummond (2017). "bModelTest: Bayesian phylogenetic site model averaging and model comparison". In: *BMC evolutionary biology* 17.1, p. 42.

Bouckaert, Remco, Joseph Heled, Denise Kühnert, Tim Vaughan, Chieh-Hsi Wu, Dong Xie, Marc A Suchard, Andrew Rambaut, and Alexei J Drummond (2014). "BEAST 2: a software platform for Bayesian evolutionary analysis". In: *PLoS computational biology* 10.4, e1003537.

Chretien, Jean-Paul, Steven Riley, and Dylan B George (2015). "Mathematical modeling of the West Africa Ebola epidemic". In: *eLife* 4. Ed. by Mark Jit, e09186. (Visited on 10/25/2018).

Coltart, Cordelia E. M., Benjamin Lindsey, Isaac Ghinai, Anne M. Johnson, and David L. Heymann (2017). "The Ebola outbreak, 2013–2016: old lessons for new epidemics". In: *Phil. Trans. R. Soc. B* 372.1721, p. 20160297.

Drummond, Alexei J, Simon YW Ho, Matthew J Phillips, and Andrew Rambaut (2006). "Relaxed phylogenetics and dating with confidence". In: *PLoS biology* 4.5, e88.

Dudas, Gytis et al. (2017). "Virus genomes reveal factors that spread and sustained the Ebola epidemic". In: *Nature* 544.7650, pp. 309–315. (Visited on 05/09/2017).

Gavryushkina, Alexandra, David Welch, Tanja Stadler, and Alexei J Drummond (2014). "Bayesian inference of sampled ancestor trees for epidemiology and fossil calibration". In: *PLoS computational biology* 10.12, e1003919.

Holmes, Edward C., Gytis Dudas, Andrew Rambaut, and Kristian G. Andersen (2016). "The evolution of Ebola virus: Insights from the 2013–2016 epidemic". In: *Nature* 538.7624, pp. 193–200.

Rambaut, Andrew, Alexei J. Drummond, Dong Xie, Guy Baele, and Marc A. Suchard. "Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7". In: *Systematic Biology*.

Stadler, Tanja, Denise Kühnert, Sebastian Bonhoeffer, and Alexei J Drummond (2013). "Birth–death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV)". In: *Proceedings of the National Academy of Sciences* 110.1, pp. 228–233.

WHO Ebola Response Team (2014). "Ebola Virus Disease in West Africa The First 9 Months of the Epidemic and Forward Projections". In: *New England Journal of Medicine* 371.16, pp. 1481–1495.

WHO Ebola Response Team (2015). "West African Ebola Epidemic after One Year Slowing but Not Yet under Control". In: *New England Journal of Medicine* 372.6, pp. 584–587.

WHO Ebola Response Team (2016). "After Ebola in West Africa — Unpredictable Risks, Preventable Epidemics". In: *New England Journal of Medicine* 375.6, pp. 587–596.

Whitty, Christopher J. M., Jeremy Farrar, Neil Ferguson, W. John Edmunds, Peter Piot, Melissa Leach, and Sally C. Davies (2014). "Infectious disease: Tough choices to reduce Ebola transmission". In: *Nature* 515.7526, pp. 192–194.

Yu, Guangchuang, David K. Smith, Huachen Zhu, Yi Guan, and Tommy Tsan-Yuk Lam (2017). "ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data". In: *Methods in Ecology and Evolution* 8.1, pp. 28–36.