Supplemental Methods - Epidemiological and viral genomic sequence analysis of the 2014 Ebola outbreak reveals clustered transmission

Samuel V. Scarpino[1]*, Atila Iamarino[2,3]*, Chad Wells[4,5], Dan Yamin[4,5], Martial Ndeffo-Mbah[4,5], Natasha S. Wenzel[4], Spencer J. Fox[6], Tolbert Nyenswah[7], Frederick L. Altice[5,8], Alison P. Galvani[4,5,9,10], Lauren Ancel Meyers[1,6], Jeffrey P. Townsend[2,9,10]□

[1]Santa Fe Institute, Santa Fe, NM 87501, USA

[2]Department of Biostatistics, Yale School of Public Health, 135 College St, New Haven, CT 06510, USA

[3]Department of Microbiology, Biomedical Sciences Institute, University of São Paulo, São Paulo, Brazil

[4]Center for Infectious Disease Modeling, Yale School of Public Health, New Haven, CT 06510, USA

[5]Department of Epidemiology of Microbial Diseases, Yale School of Public Health, New Haven, CT, USA

[6]Department of Integrative Biology, University of Texas at Austin, Austin, TX 78712, USA
[7]Ministry of Health and Social Welfare, Monrovia, Liberia
[8]Section of Infectious Diseases, Yale University School of Medicine, New Haven, CT, USA
[9]Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT 06520
[10]Department of Ecology and Evolutionary Biology, Yale University, New Haven, CT 06520

*contributed equally to this manuscript

□Direct correspondence to Jeffrey.Townsend@yale.edu, Tel: (203) 737-7042

Methods:

Simple SEIR case-count estimation of R_0

Two methods were used to estimate the basic reproduction number from the case counts of EVD in Sierra Leone. For both methods, we used empirically estimated exponentially distributed durations of the latent and infectious periods from the EVD outbreak in DRC in 1995 [1, 2]. Case counts from the Sierra Leone Ministry of Health and Sanitation and from the WHO were used to separately estimate R_0, using only cases through June 20, 2014 because this was date of the last sample in the sequence data.

The first method estimates the exponential rate of growth of the cumulative number of cases using linear regression. The growth rate is then combined with the distributions of the lengths of the latent and infectious periods to estimate R_0 [3, 4]. We estimated confidence intervals for R_0 as the 2.5% and 97.5% quantiles of 1000 sample values of the growth rate drawn from a normal distribution with mean and standard deviation as estimated by the linear regression.

The second method uses a Poisson likelihood for the number of new cases in each reporting period, given the assumed distributions of the latent and infectious periods [5]. To enforce the condition that $R_0 > 0$, we reparameterized the likelihood in terms of log R_0. We then found the maximum-likelihood estimate of R_0 using the reparameterized likelihood.

Pair approximation estimation of R_0

Rather than assuming a population is well-mixed and equally likely to contract EVD from an infectious individual, a pair approximation model includes contact structure of a population [4], which can influence the spread of disease and affect the estimation of R_0[6]. A pair approximation model captures the average disease progression through a network by modeling the dynamics of the states of connections using differential equations [4]. Pair approximation models can be used to model the epidemiological dynamics in both regular and irregular networks [4]. The benefit of using pair approximations to model disease dynamics in a network is that for most cases you do not require knowledge about the shape of the degree distribution. However, pair approximation models require the specification of a hierarchy of equations. To close the system of equations, one must approximate the higher moments with a closure method [4].

We assume the network contains clustering. The amount of clustering that occurs in a population is measured by the clustering coefficient φ, which ranges between zero and one. A clustering coefficient of zero indicates no clustering in the population, whereas a clustering coefficient of one indicates many shared contacts between contacts. Therefore, the appropriate choice to close our system of equations is the triangular pair approximation, which accounts for the progression of a disease through a clustered network [4].

Other closure methods, such as the ordinary pair approximation, only account for the average number of contacts per individual (or average degree) in the network [4]. In the presence of no clustering in the network, $\varphi = 0$, the triangular pair approximation becomes the ordinary pair approximation [4].

For a given individual X, the clustering coefficient is calculated using

$$C_X = \frac{\sum_{i=1}^{k}\sum_{j=1}^{i} e_{i,j}}{\binom{k}{2}}.$$

where e_{ij} is an indicator variable such that

$$e_{ij} = \begin{cases} 1 & \text{if } i \text{ and } j \text{ are contacts} \\ 0 & \text{if } i \text{ and } j \text{ are not contacts} \end{cases},$$

and the $\mathrm{E}(e_{ij}) = p$.

We estimated the average number of contacts per individual (or average degree), k, and clustering coefficient, φ, from contact tracing data collected by the Liberian Ministry of Health and Social Welfare between August 7–August 26 of 2014.

We used 84 sub-networks, of traced individuals, to estimate the local clustering coefficient of traced infected individuals. To obtain an estimate of the global clustering coefficient, we calculated the average of the local clustering coefficients from the contact tracing data. Given we only have information only about two traced individuals, and not the entire network, we estimate the clustering coefficient for individual X by

$$\overline{C_X} = \frac{\sum_{i=1}^{k} e_{i,m}}{k-1}.$$

where the index *m* represents a traced individual whose contacts are known. The maximum number of mutual contacts between individual *X* and individual *m* is *k* − *1*. Our estimate represents the likelihood that *m* shares a contact with *X*. Our estimate is unbiased because

$$E(\overline{C_X}) = E\left(\frac{\sum_{i=1}^{k} e_{i,m}}{(k-1)}\right) = \frac{1}{k-1}E\left(\sum_{i=1}^{k} e_{i,m}\right) = \frac{1}{k-1}(k-1)p = p$$

$$E(C_X) = E\left(\frac{\sum_{i=1}^{k}\sum_{j=1}^{i} e_{i,j}}{\binom{k}{2}}\right) = \frac{1}{\binom{k}{2}}E\left(\sum_{i=1}^{k}\sum_{j=1}^{i} e_{i,j}\right) = \frac{1}{\binom{k}{2}}\binom{k}{2}p = p.$$

From the contact tracing data, we estimate the average number of contacts per individual to be *k* = 5.74 (95% CI: 4.89–6.60) and the clustering coefficient to be φ = 0.21 (0.196–0.223). The estimated household size of Liberia is 5.1 [7] and the average household size in Sierra Leone is 5.4 [8]. With the average household sizes being similar, we chose to use the average number of contacts obtained from the empirical contact tracing data.

We based our equations from previous *SEIR* pair approximation models [4, 9, 10], where infection in the population is driven by the number of susceptible-infected. Specifically, a susceptible individual (*S*) becomes exposed (*E*) after being infected by an infectious contact, $\beta[SI]$, where [*SI*] denotes the number of susceptible and infected contacts. An exposed individual becomes infectious after an average period of $1/\sigma$ days. Once an individual is symptomatic and infectious, the average time to death is $1/\delta$ days and the average time to recovery is $1/\rho$, where we assumed the case fatality ratio is δ / (δ + ρ). Once the infected individual has become deceased or recovered they enter the removed state. We assumed the initial number of exposed (*E*) individuals was randomly distributed throughout the population. This assumption corresponds to the following set of initial conditions

$$\begin{aligned}[S](0) &= N - [E](0)\\ [SE](0) &= k[E](0)\\ [SS](0) &= Nk - 2[SE](0)\\ [EE](0) &= 0,\end{aligned}$$

where all other states are equal to zero, where *N* is the population size, and where *k* is the average number of contacts per individual. The following terms denote possible pairs of individuals, with [*SS*] indicating twice the number of susceptible-susceptible connections, [*EE*] indicating twice the number of exposed-exposed connections, and the term [*SE*] indicating the number of susceptible-exposed connections in the population. The following set of differential equations were coded in MATLAB and solved using *ode15s* [11]:

$$\frac{\mathrm{d}[S]}{\mathrm{d}t} = -\beta[SI]$$

$$\frac{\mathrm{d}[E]}{\mathrm{d}t} = \beta[SI] - \sigma[E]$$

$$\frac{\mathrm{d}[I]}{\mathrm{d}t} = \sigma[E] - \left(\frac{\delta^2 + \rho^2}{\delta + \rho}\right)[I]$$

$$\frac{\mathrm{d}[R]}{\mathrm{d}t} = \left(\frac{\delta^2 + \rho^2}{\delta + \rho}\right)[I]$$

$$\frac{\mathrm{d}[SS]}{\mathrm{d}t} = -2\beta\left((1-\phi)\frac{k-1}{k}\frac{[SS][SI]}{[S]} + \phi\frac{N(k-1)}{k^2}\frac{[SS]^2[SI]}{[I][S]^2}\right)$$

$$\begin{aligned}\frac{\mathrm{d}[SE]}{\mathrm{d}t} = &-\beta\left((1-\phi)\frac{k-1}{k}\frac{[SE][SI]}{[S]} + \phi\frac{N(k-1)}{k^2}\frac{[SE][EI][SI]}{[S][I][E]}\right)\\ &+\beta\left((1-\phi)\frac{k-1}{k}\frac{[SS][SI]}{[S]} + \phi\frac{N(k-1)}{k^2}\frac{[SS]^2[SI]}{[I][S]^2}\right) - \sigma[SE]\end{aligned}$$

$$\frac{\mathrm{d}[SI]}{\mathrm{d}t} = -\beta\left([SI] + (1-\phi)\frac{k-1}{k}\frac{[SI]^2}{[S]} + \phi\frac{N(k-1)}{k^2}\frac{[SI]^2[II]}{[S][I]^2}\right) + \sigma[SE] - \left(\frac{\delta^2 + \rho^2}{\delta + \rho}\right)[SI]$$

$$\frac{\mathrm{d}[SR]}{\mathrm{d}t} = -\beta\left((1-\phi)\frac{k-1}{k}\frac{[SR][SI]}{[S]} + \phi\frac{N(k-1)}{k^2}\frac{[SR][IR][SI]}{[S][I][\mathrm{R}]}\right) + \left(\frac{\delta^2 + \rho^2}{\delta + \rho}\right)[SI]$$

$$\frac{\mathrm{d}[EE]}{\mathrm{d}t} = 2\beta\left((1-\phi)\frac{k-1}{k}\frac{[SE][SI]}{[S]} + \phi\frac{N(k-1)}{k^2}\frac{[SE][EI][SI]}{[S][I][E]}\right) - 2\sigma[EE]$$

$$\frac{\mathrm{d}[EI]}{\mathrm{d}t} = \beta\left([SI] + (1-\phi)\frac{k-1}{k}\frac{[SI]^2}{[S]} + \phi\frac{N(k-1)}{k^2}\frac{[SI]^2[II]}{[S][I]^2}\right) - \sigma[EI] + \sigma[EE] - \left(\frac{\delta^2 + \rho^2}{\delta + \rho}\right)[EI]$$

$$\frac{\mathrm{d}[ER]}{\mathrm{d}t} = \beta\left((1-\phi)\frac{k-1}{k}\frac{[SR][SI]}{[S]} + \phi\frac{N(k-1)}{k^2}\frac{[SR][IR][SI]}{[S][I][\mathrm{R}]}\right) - \sigma[ER] + \left(\frac{\delta^2 + \rho^2}{\delta + \rho}\right)[EI]$$

$$\frac{\mathrm{d}[II]}{\mathrm{d}t} = 2\sigma[EI] - 2\left(\frac{\delta^2 + \rho^2}{\delta + \rho}\right)[II]$$

$$\frac{\mathrm{d}[IR]}{\mathrm{d}t} = \sigma[ER] - \left(\frac{\delta^2 + \rho^2}{\delta + \rho}\right)[IR] + \left(\frac{\delta^2 + \rho^2}{\delta + \rho}\right)[II]$$

$$\frac{\mathrm{d}[RR]}{\mathrm{d}t} = 2\left(\frac{\delta^2 + \rho^2}{\delta + \rho}\right)[IR].$$

The first-order equations denote the absolute number of individuals in that epidemiological state. For example, $[S]$ denotes the number of susceptible individuals in the population and $[I]$ denotes the number of infected individuals in the population. The second-order equations denote the number of pairs between different states. For example, $[SI]$ denotes the number of susceptible-infected pairs in the network and $[SR]$ denotes the number of susceptible-removed pairs in the network. The terms $[SS]$, $[EE]$, $[II]$, and $[RR]$ represent twice the number of pairs. Thus, there is a factor of two in the respective differential equations. The assumptions of disease transmission of the first-order equations also pertain to the second order equations. For example, a connection between a susceptible and an infected individual ($[SI]$) transitions to a connection between an exposed and an infected individual ($[EI]$) if the susceptible is infected. In addition, a connection between a susceptible individual and infected individual ($[SI]$) transitions to a connection between a susceptible individual and a removed individual ($[SR]$) after the infectious individual recovers.

We parameterized our inference with the average time to from exposure to symptoms, $1/\sigma$, as 9 days [12], the average time from symptoms to death, $1/\delta$, as 8.6 days [12], a population size $N = 6{,}348{,}350$ [13], and an initial 14 exposed individuals [14]. We chose to use the entire population of Sierra Leone as the majority of the regions had at least one confirmed case of Ebola by August 31 [13].

To determine the presence of clustering, we used a likelihood approach in determining the model that best fits the known data. During the testing of the model we found that the estimated clustering coefficient was correlated with the origin time of the epidemic. Previous studies have indicated an origin time of approximately April 23 [15]. We assumed the origin time was gamma distributed with an average origin time of April 23 (95% CI: March 31 , May 10), based on findings from a previous study [14]. In addition to using a prior based on the results from previous studies [14], we separately examined the extent of clustering based on our phylodynamic results, which are from Sierra Leone only. We assumed the error in estimating the cumulative incidence was distributed normally for each time point. To estimate the variance of the normal distribution,

$$\sigma_\varepsilon^2 = \frac{\sum_{i=1}^{N_I}(T_i - \tau_i)^2}{N_I - 1},$$

we took an iterative approach such that the estimated variance is from the model that maximized the likelihood function

$$L = \Gamma(t(0))^{N_I} \prod_{i=1}^{N_I} \frac{\exp\left\{-(T_i - \tau_i)^2 / \left(2\sigma_\varepsilon^2\right)\right\}}{\sqrt{2\pi\sigma_\varepsilon^2}},$$

where $t(0)$ is the estimated origin time, T_i is the estimated cumulative incidence at time point i, τ_i is the reported cumulative incidence at time point i, and N_I is the total number of cumulative incidence points.

We then sampled 10,000 origin times from our gamma distribution and estimated the transmission rate, β, the time to recovery, $1/\rho$, and clustering coefficient φ. To estimate these parameters, we fit the model using the confirmed cumulative incidence τ, confirmed cumulative mortality μ, from Sierra Leone from the WHO reports. We fit the model using a MATLAB non-linear least squares fitting algorithm, *lsqnonlin* [16], that minimized the mean square error (*MSE*), where T_i is the estimated cumulative incidence at time point *i*, τ_i is the reported cumulative incidence at time point *i*, N_I is the total number of cumulative incidence points, M_i is the estimated cumulative mortality at time point *i*, μ_i is the reported cumulative mortality at time point *i*, N_M is the total number of cumulative mortality points. We chose to include the cumulative mortality in the fitting process as it helps fit the estimated time to recovery. We did not include the cumulative mortality in the likelihood function because the cumulative incidence and cumulative mortality were highly correlated ($R^2 = 0.992$).

$$MSE = \frac{\sum_{i=1}^{N_I}\left(T_i - \tau_i\right)^2 + \sum_{i=1}^{N_M}\left(M_i - \mu_i\right)^2}{N_I + N_M - 1}.$$

We then calculated the basic reproductive number and likelihood for the estimated parameter values. The basic reproductive number depends on the quasi-steady state of the susceptible and infectious interactions during the early stages of infection [4, 6]. The interaction between susceptible and infected individuals is measured by the susceptible-infected pair correlation [6]

$$C_{SI} = \frac{N}{k}\frac{[SI]}{[S][I]},$$

Once we obtained the parameters for the best fit, we introduced a single infectious individual into the population and numerically calculated

$$R_0 = \frac{\beta k(\delta + \rho)C_{SI}}{\delta^2 + \rho^2}.$$

Using our 10,000 samples, we determined which parameter set maximized our likelihood function to obtain our estimate of the clustering coefficient and R_0. To obtain the 95% CI for the clustering coefficient and R_0, we took the sets of parameters and R_0 values in the top 95% of the likelihood and calculated the minimum and maximum values for the lower and upper bounds of the confidence interval, respectively.

Parameter	**Description**	**Sierra Leone sequences**	**Gire et al origin time**	**Optimal non-clustered model ***
φ	Clustering coefficient	0.71	0.36	0
β	Transmission rate per contact per day	0.61	0.071	0.048
$1/\rho$	Average duration to recovery (days)	8.80	8.80	8.83
$t(0)$	Estimated time of origin	May 27	April 27	April 17
R_0	Basic reproductive number	1.29	1.41	1.47
Mean Square Error	Cumulative incidence only	363	984	1334
Mean Square Error	Cumulative incidence and cumulative mortality	329	661	897

Table S2. The maximum likelihood estimated parameters based on an average of 5.74 contacts per individual. The model was fitted to the WHO data set of confirmed cumulative incidence and confirmed cumulative mortality in Sierra Leone from May 27–August 31, assuming prior distributions of the origin time from our phylodynamic results and from a previous study [14]. *The optimal non-clustered model was fit to minimize the mean square error of cumulative incidence and mortality without any prior distribution of the origin time.

Phylodynamic estimation of R_0

Our sequence-based parameter estimates of the recent outbreak could in principle be subject to the bias due to a lack of appropriate samples for rooting this outbreak in the EBOV species tree, as described in [14, 17, 18]. We avoided these issues by focusing only on Sierra Leone samples and using a transmission-oriented evolutionary reconstruction appropriate for beginning epidemics [19, 20]. Thus, we sacrificed the ability to infer the number of epidemiological introductions in Sierra Leone, already estimated as two different events by Gire et al. [14] in favor of more reliable estimates of R_0. We used different methods to estimate R_0 from genome-sequence data of EBOV collected and dated from infected cases in Sierra Leone from late May to the middle of June 2014 [14]. The data set included 78 whole genomes that were trimmed to a 18,538 bp alignment. For strains with more than one sample per patient, we kept only the oldest sequence available.

We inferred a phylodynamic estimate using Bayesian Markov chain Monte Carlo (MCMC) methods implemented in BEAST 2.1 [21] with the HKY substitution model, as suggested by jModeltest 2.1 [22], and a proportion of invariable sites model under the *B*irth-*D*eath *S*usceptible *I*nfectious *R*emoved (BDSIR) tree model [19]. Through a transmission-oriented branching process [23], the BDSIR tree model used for the evolutionary reconstruction accommodates stochastic processes that are characteristic of the first stages of an epidemic [24], such as fluctuations of population size [19], and is especially suited to incorporate uncertainties that come from epidemics with low values of R_0 [19, 20]. We used a rate prior with a mean of 7×10^{-4} s/s/y (substitutions per site per year), a value compatible with previous independent estimates for whole genomes and the glycoprotein gene of *Zaire ebolavirus* [25, 26]. Convergence was obtained with four independent MCMC runs of 125 million generations and checked with effective sample size (ESS) values above 200 calculated with TRACER v1.6 [27]. The best-fit model, according to Bayes Factor model comparison, was the relaxed, lognormal molecular clock [28]. Consistent with findings that the BDSIR-based phylodynamics estimate reliable values of R_0 that are comparable to more complex stochastic-coalescent SIR models [20], the BDSIR relaxed exponential molecular clock, with different starting priors for R_0 ranging from 1–10, produced consistent results ($R_0 = 1.4$ for the lognormal clock and $R_0 = 1.38$ for the exponential clock).

Outbreaker estimation of R$_0$

The second method, implemented in the R package outbreaker, reconstructs the transmission tree(s) for the sampled sequences in a Bayesian framework using the pathogen's serial interval distribution and a simple mutational model to construct a likelihood equation for a specific transmission tree [29]. We assumed that the serial interval distribution was equal to a Gamma distribution with mean 15.3 days and coefficient of variation 0.66, which was empirically estimated by the WHO Ebola Response Team [12]. The width of the credible intervals around the Bayesian clustering estimates stems from uncertainty in the inter-case, or serial, interval distribution. We assumed a gamma distributed serial interval, and estimated lower R_0 values when the distribution was closer to a Gaussian and higher values as it approached an exponential; however, the R_0 estimates were comparatively robust to shifts in the mean of the serial interval distribution

Supplemental References

1. Chowell G, Hengartner NW, Castillo-Chavez C, Fenimore PW, Hyman JM. The basic reproductive number of Ebola and the effects of public health measures: the cases of Congo and Uganda. Journal of theoretical biology **2004**; 229(1): 119-26.

2. Lekone PE, Finkenstadt BF. Statistical inference in a stochastic epidemic SEIR model with control intervention: Ebola as a case study. Biometrics **2006**; 62(4): 1170-7.

3. Wallinga J, Lipsitch M. How generation intervals shape the relationship between growth rates and reproductive numbers. Proceedings Biological sciences / The Royal Society **2007**; 274(1609): 599-604.

4. Rand D. Correlation equations and pair approximations for spatial ecologies. Advanced ecological theory: principles and applications **1999**; 100.

5. White LF, Pagano M. A likelihood-based method for real-time estimation of the serial interval and reproductive number of an epidemic. Statistics in medicine **2008**; 27(16): 2999-3016.

6. Keeling M. The implications of network structure for epidemic dynamics. Theoretical population biology **2005**; 67(1): 1-8.

7. National population and housing census, 2008: preliminary results. Monrovia, Liberia: Liberia Institute of Statistics and Geo-Information Services (LISGIS), **2008**.

8. Sanitation MoHa. Sierra Leone LLIN Universal Access Campaign Post-Campaign Ownership and Use Survey. In: Sanitation MoHa. Sierra Leone: Government of Sierra Leone, **2011**.

9. Ringa N, Bauch CT. Dynamics and control of foot-and-mouth disease in endemic countries: A pair approximation model. Journal of theoretical biology **2014**; 357: 150-9.

10. Parham PE, Singh BK, Ferguson NM. Analytic approximation of spatial epidemic models of foot and mouth disease. Theoretical population biology **2008**; 73(3): 349-68.

11. MathWorks ode15s function. Available at: http://www.mathworks.com/help/matlab/ref/ode15s.html. Accessed Nov 27th 2014.

12. Team WHOER. Ebola Virus Disease in West Africa - The First 9 Months of the Epidemic and Forward Projections. The New England journal of medicine **2014**.

13. Sanitation MoHa. EBOLA VIRUS DISEASE - SITUATION REPORT (Sit-Rep) – 01September, 2014 In: Sanitation MoHa: Government of Sierra Leone, **2014**.

14. Gire SK, Goba A, Andersen KG, et al. Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. Science **2014**.

15. Althaus CL. Estimating the Reproduction Number of Ebola Virus (EBOV) During the 2014 Outbreak in West Africa. PLoS currents **2014**.

16. MathWorks lsqnonlin function. Available at: http://www.mathworks.com/help/optim/ug/lsqnonlin.html. Accessed Nov 27th 2014.

17. Calvignac-Spencer S, Schulze JM, Zickmann F, Renard BY. Clock Rooting Further Demonstrates that Guinea 2014 EBOV is a Member of the Zaire Lineage. PLoS currents **2014**; 6.

18. Dudas G, Rambaut A. Phylogenetic Analysis of Guinea 2014 EBOV Ebolavirus Outbreak. PLoS currents **2014**; 6.

19. Kuhnert D, Stadler T, Vaughan TG, Drummond AJ. Simultaneous reconstruction of evolutionary history and epidemiological dynamics from viral sequences with the birth-death SIR model. Journal of the Royal Society, Interface / the Royal Society **2014**; 11(94): 20131106.

20. Popinga A, Vaughan T, Stadler T, Drummond A. Bayesian Coalescent Epidemic Inference: Comparison of Stochastic and Deterministic SIR Population Dynamics. arXiv preprint arXiv:14071792 **2014**.

21. Bouckaert R, Heled J, Kuhnert D, et al. BEAST 2: a software platform for Bayesian evolutionary analysis. PLoS computational biology **2014**; 10(4): e1003537.

22. Darriba D, Taboada GL, Doallo R, Posada D. jModelTest 2: more models, new heuristics and parallel computing. Nature methods **2012**; 9(8): 772.

23. Stadler T, Kouyos R, von Wyl V, et al. Estimating the basic reproductive number from viral sequence data. Molecular biology and evolution **2012**; 29(1): 347-57.

24. Rouzine IM, Coffin JM. Linkage disequilibrium test implies a large effective population number for HIV in vivo. Proceedings of the National Academy of Sciences of the United States of America **1999**; 96(19): 10758-63.

25. Carroll SA, Towner JS, Sealy TK, et al. Molecular evolution of viruses of the family Filoviridae based on 97 whole-genome sequences. Journal of virology **2013**; 87(5): 2608-16.

26. Li YH, Chen SP. Evolutionary history of Ebola virus. Epidemiology and infection **2014**; 142(6): 1138-45.

27. Rambaut A, Drummond A, Suchard M. Tracer v1. 6. **2013**.

28. Drummond AJ, Ho SY, Phillips MJ, Rambaut A. Relaxed phylogenetics and dating with confidence. PLoS biology **2006**; 4(5): e88.

29. Jombart T, Cori A, Didelot X, Cauchemez S, Fraser C, Ferguson N. Bayesian Reconstruction of Disease Outbreaks by Combining Epidemiologic and Genomic Data. PLoS computational biology **2014**; 10(1): e1003457.