THE UNIVERSITY *of* EDINBURGH
School of Biological Sciences

[Comparative and Evolutionary Genomics]

[ICA1]

Student Exam Number: …B196466…………

The file name of your uploaded document **must** include your exam number

EBOV Phylodynamic Analysis

Introduction

In West Africa, the Ebola virus (EBOV) outbreak spreads very quickly in the area. [1][2][3]. Nevertheless, the virus's origin in each country and time of transmission is still unknown. A recently published sequence analysis of the outbreak suggests that the spread of EBOV in Guinea was resulted from a different spectrum of Zaire Ebola viruses[2].

The Ebola virus genome consists of a single strand of negative-sense RNA and contains seven protein-coding genes[4]. Because it took time for scientists to complete virus genome sequencing, Ebola, which outbreaks in Guinea, has spread to other countries and caused severe hazards, unfortunately. Control measures have failed to halt the rapid increase in the number of infections. Mastering the changing location and timing dynamics of emerging epidemics is very significant for designing real-time disease interventions, so for analyzing the time-scale and spatial-spread of the pandemic, the sequenced samples from Guinea, Sierra Leone and Liberia are recorded, and the dates each sample was taken and the location of each sample are also recorded.

By this research, people could understand the timing, geography and growth of the epidemic. According to the research result, it shows when and where this virus outbreak occurred and whether the epidemic was caused by one or multiple zoonotic events. In order to interpret the timing and spread of the Ebola virus, it calculates its evolutionary rate. Observing the phylogenetic tree creates a discrete phylogeographic model, and it reveals that in different countries, the virus is spread by imports or local transmission and its number of incoming transfers. Also, it calculates its adequate population size and the growth rate of the pandemic and epidemiological parameters like R_0. Mathematical models play an important role in estimating epidemic growth rates and the basic reproduction number (R_0) from recently molecular sequence data and newly recorded date[5]. At the same time, it describes the demographics of the outbreak.

Methods

Firstly, to build a phylogenetic tree, IQTree2 is used because it is a new version of IQTree and is suitable to analyze virus sequences. The model of molecular evolution which is used in this case is HKY. Some output files will be produced, and for example,"Makona_HKY.bionj" is a Neighbor-Joining tree created at the beginning as a starting tree for the tree search, "Makona_HKY.midist" is a matrix of pairwise genetic distances between all the sequences and "Makona_HKY.treefile" is the final tree file in a machine-readable format. And all the results will be put into the file whose suffix is "iqtree".

Secondly, to explore epidemiological processes, after converting the branch lengths in the tree to years, the timing and spread of the Ebola virus can be analyzed. The reason is that its distance to the root positively correlates with its sampling time. "Treetime" is used to search for the root position that gives the best temporal signal and then scales the branch lengths in a tree to years using maximum likelihood optimization. To tell tree time how long the sequences are, "sequence length" is applied, and the "reroot" option defines how to determine the best root position. Here it is based on least-squares while accounting for the covariance between tips in the tree based on their shared ancestry.

Thirdly, “Treetime” also estimates the location of the root and patterns of spatial spread with. It uses mugration models because they are similar to the substitution models used to estimate genetic distance. It also contains a discrete phylogeographic model for reconstructing ancestral traits on a tree.

And then, “Treetime” can estimate the effective population size of the dataset throughout time. These figures can be used to estimate the growth rate of the pandemic by using the “coalescent skyline” option. Furthermore, the demographic could be used to estimate the growth rate of the pandemic and epidemiological parameters like R_0. All of these estimations depend on global sampling strategies because all statistical results come from these figures, and these figures are made by just a few isolates.

At last, FigTree is used to view trees and show figures by adjustment of various parameters.

Results

Timing estimation

These date has a good temporal signal. The outbreak of Ebola virus appears in Guinea in January 2014 and its estimated evolutionary rate is 8.73e-04+/- 1e-04.(Figure 1)

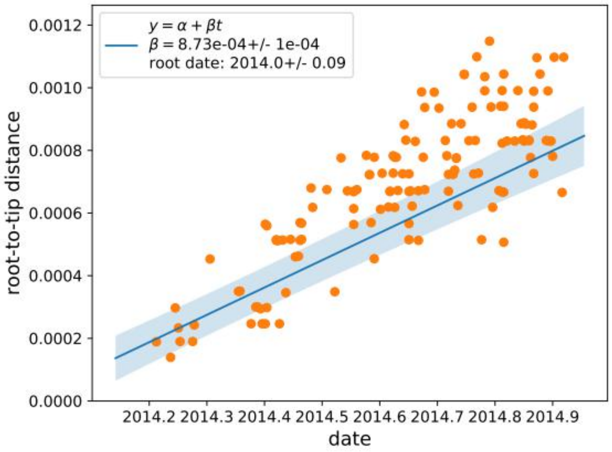


Figure 1 evolutionary rate equation

The epidemic caused by one zoonotic event because phylogenetic tree has a common root and an reference shows a zoonotic transmission which transfer from a bat to a young boy which was only two year old in December 2013 is the origin of Ebola virus[2].(Figure 2)

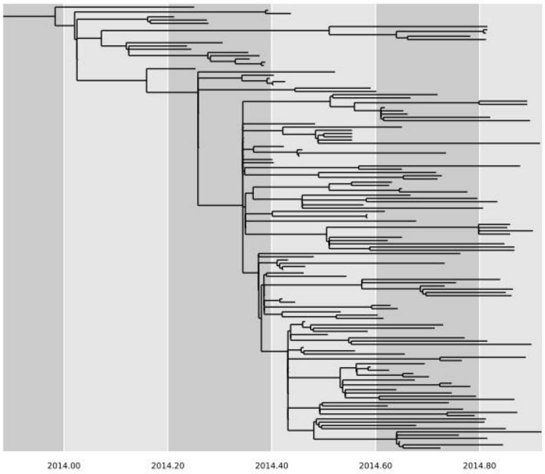


Figure 2 phylogenetic tree which are made by Guinea, Sierra Leone and Liberia

Population estimation

About the demographics of this outbreak, from 2014.1 to 2014.9, the size of the infected population has a upward trend but there was a small decrease in March. The number of the population gradually rises to 10^3 afterwards.(Figure 3)

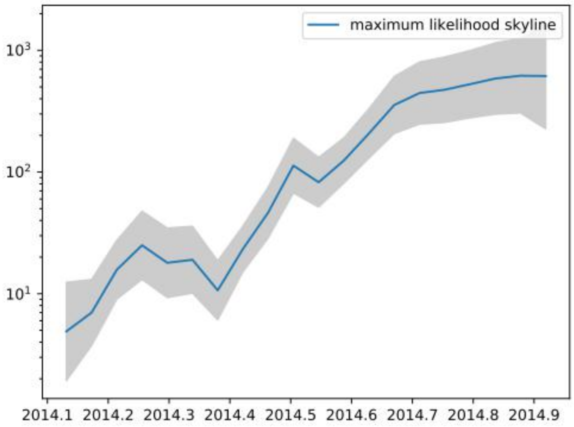
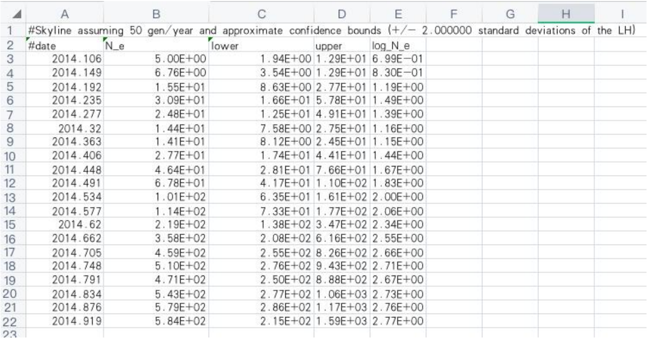


Figure 3 raw demographic estimation which is calculated by maximum likelihood

R_0 is the average number of individuals who would transmit the disease from one initial case of an infectious disease to another without any vaccination and with no immunity. And an estimate for the mean serial interval of Ebola virus is 11.6 days.

To calculate R_0, opening the result file in Excel,(Figure 4) setting a new column which is equal to the natural log of the N_e column, producing a plot, trimming the plot to the first period of exponential grow and adding a linear trendline with equation to the plot is its main procedure.(Figure 5)



#Skyline assuming 50 gen/year and approximate confidence bounds (+/− 2.000000 standard deviations of the LH)

#date	N_e	lower	upper	log_N_e
2014.106	5.00E+00	1.94E+00	1.29E+01	6.99E−01
2014.149	6.76E+00	3.54E+00	1.29E+01	8.30E−01
2014.192	1.55E+01	8.63E+00	2.77E+01	1.19E+00
2014.235	3.09E+01	1.66E+01	5.78E+01	1.49E+00
2014.277	2.48E+01	1.25E+01	4.91E+01	1.39E+00
2014.32	1.44E+01	7.58E+00	2.75E+01	1.16E+00
2014.363	1.41E+01	8.12E+00	2.45E+01	1.15E+00
2014.406	2.77E+01	1.74E+01	4.41E+01	1.44E+00
2014.448	4.64E+01	2.81E+01	7.66E+01	1.67E+00
2014.491	6.78E+01	4.17E+01	1.10E+02	1.83E+00
2014.534	1.01E+02	6.35E+01	1.61E+02	2.00E+00
2014.577	1.14E+02	7.33E+01	1.77E+02	2.06E+00
2014.62	2.19E+02	1.38E+02	3.47E+02	2.34E+00
2014.662	3.58E+02	2.08E+02	6.16E+02	2.55E+00
2014.705	4.59E+02	2.55E+02	8.26E+02	2.66E+00
2014.748	5.10E+02	2.76E+02	9.43E+02	2.71E+00
2014.791	4.71E+02	2.50E+02	8.88E+02	2.67E+00
2014.834	5.43E+02	2.77E+02	1.06E+03	2.73E+00
2014.876	5.79E+02	2.86E+02	1.17E+03	2.76E+00
2014.919	5.84E+02	2.15E+02	1.59E+03	2.77E+00

Figure 4 Skyline assuming 50 gen/year and approximate confidence bounds

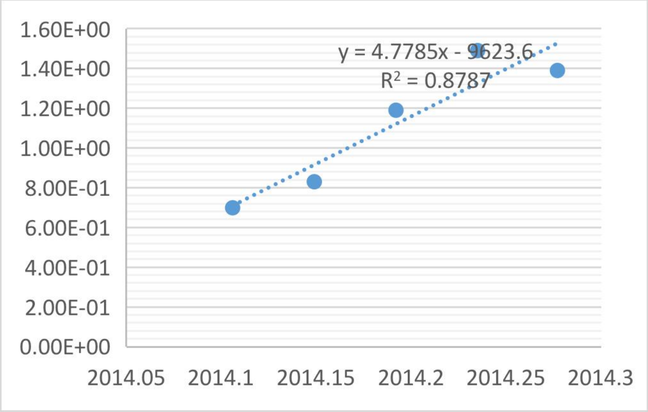


Figure 5 plot which is the first period of exponential grow

The slope of this line represents the yearly exponential growth rate of the pandemic and by equation and calculating, R_0 is 1.16. And R_0 is changed by overtime because it is calculated by the daily growth rate. And the daily growth rate is the equation's slope. When the time change, the date will change and the slope also will change. Regarding to location, R_0 does not have relationship with location so it will not change when choose different countries. At the same time, growth rate of infected population is its equation's slope which is 8.80e-04 +/- 1e-04.

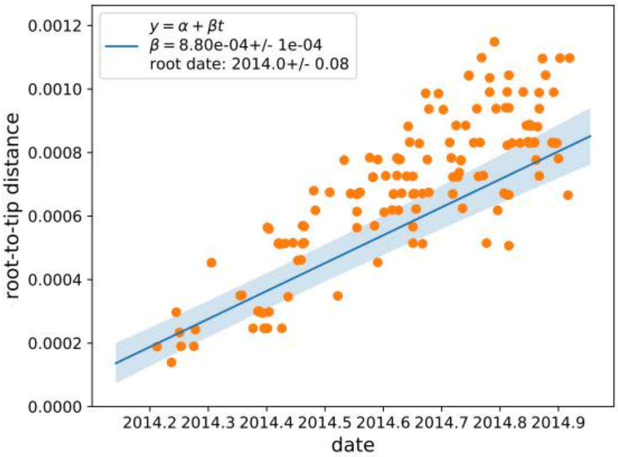


Figure 6 growth rate of infected population equation

Location estimation

In 2014.1, it is the first time that the virus was introduced to Sierra Leone and only appear once.

After observing the figure, it can be inferred that the cases in Sierra Leone and Liberia are caused by imports transmission and the cases in Guinea are from local transmission.(Figure 7)

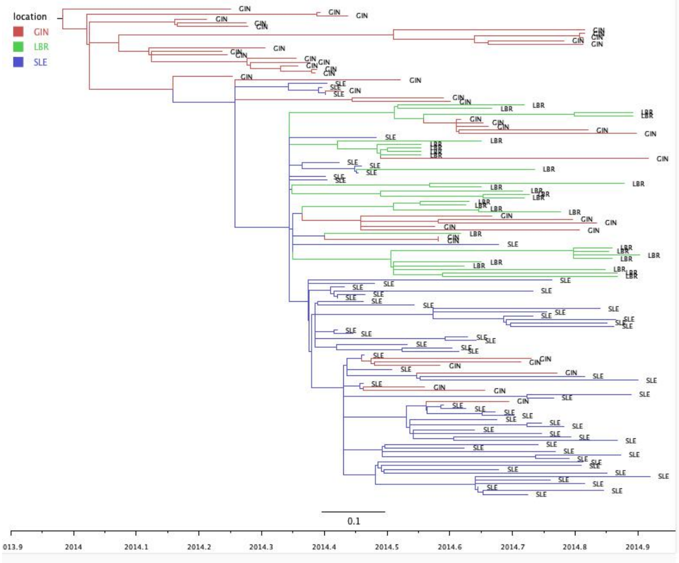


Figure 7 phylogenetic tree which are made by Guinea, Sierra Leone and Liberia

Discussion

In conclusion, the outbreak of the EBOV virus appears in Guinea in the beginning, and gradually it spread to other countries like Sierra Leone and Liberia. Afterwards, it causes inter-country dissemination. What's more, its evolutionary rate and its infected population are rising, and its R_0 is also changing following the date.

However, other information shows that "Though it looks like it is unlikely that the virus will continue to spread Sierra Leone borders throughout the epidemic, these recorded date suggest that cross-border importation was not a significant factor in the process of spreading the virus, at least at the end of 2014."[7] It proves that domestic dissemination is the main reason that causes virus outbreaks in Sierra Leone. The imports transmission from Guinea is just a causal factor, and the main problem that causes such a severe impact is domestic dissemination. Regarding Liberia, multiple Introductions of the Ebola Virus to Liberia is what shows in the phylogenetic tree and There are reports that EBOV entered Liberia on several occasions, but one of these resulted in the majority of cases and spread. At the same time, another reference said that "In Guinea, the sudden outbreak of the virus is likely caused by a Zaire Ebola virus lineage that has spread from Central Africa into Guinea and West Africa which are happening more frequently in recent decades" [4], which reveal that Guinea is not where the virus originated. The endemic strain in Guinea is derived from the EBOV spectrum previously found in the Democratic Republic of the Congo, the Republic of Congo and Gabon[8]. Also, the article notes that the virus spreads throughout Guinea, Sierra Leone and Liberia through person-to-person contact. EBOV from Guinea is most likely to enter Sierra Leone in April or May. Viruses appear in the Guinea or Sierra Leone spectrum mix around June and July in 2014[9].

If $R_0 < 1$, the epidemic will gradually disappear. If $R_0 > 1$, the infection spreads exponentially and becomes an epidemic. However, it will not last forever, as the number of people who may be infected slowly decreases. Some of the population may die from the epidemic, while others may recover and become immune. EBOV has a short infection period and high transmission.For example, R_0 for the Sierra Leone outbreak ranges between 1.4 and 1.82, for the Liberia outbreak ranges between 1.93 and 2.27 and for the Guinea ranges between 1.04 and 1.42[10]. R_0 for EBOV is between 1.5 to 2.5 but for HIV is 4.5 which only counts Africa. The higher is $R_{0,}$ the harder for the epidemic to disappear and it will have wider dissemination.

Countries where people do not have access to health services and have inadequate health care systems will lead to both Ebola and AIDS spreading.And in Liberia, Sierra Leone or Guinea, health system almost do not exist and that making monitoring the disease difficult and increasing the likelihood of future outbreaks[11].

References

[1]Schieffelin, John S., et al. "Clinical illness and outcomes in patients with Ebola in Sierra Leone." New England journal of medicine 371.22 (2014): 2092-2100.

[2]Baize, Sylvain, et al. "Emergence of Zaire Ebola virus disease in Guinea." New England Journal of Medicine 371.15 (2014): 1418-1425.

[3]Gatherer, Derek. "The unprecedented scale of the West African Ebola virus disease outbreak is due to environmental and sociological factors, not special attributes of the currently circulating strain of the virus." BMJ Evidence-Based Medicine 20.1 (2015): 28-28.

[4]Dudas, Gytis, and Andrew Rambaut. "Phylogenetic analysis of Guinea 2014 EBOV Ebolavirus outbreak." PLoS currents 6 (2014).

[5]Volz, Erik, and Sergei Pond. "Phylodynamic analysis of Ebola virus in the 2014 Sierra Leone epidemic." PLoS currents 6 (2014).

[7]Park, Daniel J., et al. "Ebola virus epidemiology, transmission, and evolution during seven months in Sierra Leone." Cell 161.7 (2015): 1516-1526.

[8]Azarian, Taj, et al. "Impact of spatial dispersion, evolution and selection on Ebola Zaire Virus epidemic waves." Scientific reports 5.1 (2015): 1-9.

[9]Carroll, Miles W., et al. "Temporal and spatial analysis of the 2014–2015 Ebola virus outbreak in West Africa." Nature 524.7563 (2015): 97-101.

[10]Taylor, Bradford P., Jonathan Dushoff, and Joshua S. Weitz. "Stochasticity and the limits to confidence when estimating R0 of Ebola and other emerging infectious diseases." Journal of theoretical biology 408 (2016): 145-154.

[11]Whiteside, Alan, and Nicholas Zebryk. "Ebola and AIDS in Africa." Canadian Journal of African Studies/Revue canadienne des études africaines 49.2 (2015): 409-419.