PLOS | CURRENTS OUTBREAKS

Phylogenetic Analysis of Guinea 2014 EBOV Ebolavirus Outbreak

May 2, 2014 · Research Article

Citation

Dudas G, Rambaut A. Phylogenetic Analysis of Guinea 2014 EBOV Ebolavirus Outbreak. PLOS Currents Outbreaks. 2014 May 2 . Edition 1. doi: 10.1371/currents.outbreaks.84eefe5ce43ec9dc0bf0670f7b8b417d.

Authors

Gytis Dudas

Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, UK.

Andrew Rambaut

University of Edinburgh, Edinburgh, UK.

Abstract

Members of the genus Ebolavirus have caused outbreaks of haemorrhagic fever in humans in Africa. The most recent outbreak in Guinea, which began in February of 2014, is still ongoing. Recently published analyses of sequences from this outbreak suggest that the outbreak in Guinea is caused by a divergent lineage of Zaire ebolavirus. We report evidence that points to the same Zaire ebolavirus lineage that has previously caused outbreaks in the Democratic Republic of Congo, the Republic of Congo and Gabon as the culprit behind the outbreak in Guinea.

Funding Statement

GD was supported by a NERC studentship D76739X. The research leading to these results has received funding from the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013) under Grant Agreement no. 278433-PREDEMICS and ERC Grant agreement no. 260864.

Introduction

A recent article[1] suggests that the currently ongoing outbreak in Guinea is caused by a divergent variant of the Zaire ebola (EBOV) lineage. The EBOV strain has previously caused ebola outbreaks in the Democratic Republic of Congo (DRC), the Republic of Congo (RC) and Gabon. The authors publish three complete genome sequences from the Guinea outbreak and perform a phylogenetic analysis using 24 sequences of the Zaire and other representative lineages. One finding is that the 2014 sequences fall as a divergent lineage outside the Zaire lineage suggesting that this may be a pre-existing endemic virus in West Africa rather than the result of spread of the EBOV lineage from the Central African countries that have had previous human outbreaks.

Previously, a dynamic re-interpretation of EBOV emergence in Central Africa has been suggested, citing correlations between time, geographic distance and genetic distance of Ebola haemorrhagic fever outbreaks[2] and the recent ancestry of related EBOV lineages in fruit bats[3].

Materials and Methods

All complete genome sequences from the genus Ebolavirus (which includes Bundibugyo BDBV, Reston RESTV, Sudan SUDV, Tai Forest TAFV and Zaire ebolavirus EBOV species) were collated from genbank including the sequences from the Guinea outbreak. Genbank accessions and sources for the sequences can be found at http://epidemic.bio.ed.ac.uk/ebolavirus_sequences.

The Ebolavirus genome consists of a single strand of negative sense RNA and contains 7 protein coding genes (in order 3′-NP-VP35-VP40-GP-VP30-VP24-L, separated by various intergenic regions)[4]. We collated the protein coding regions of each gene (alignment length 14647 nucleotides) and, in a separate alignment, the non-coding intergenic regions. Phylogenetic trees were inferred in PhyML[5] or MrBayes[6] using the GTR[7]+Γ substitution model. We were able to replicate the analysis presented in Baize *et al*.[1] only when omitting the accommodation of rate heterogeneity modelled as a discretized Γ distribution. We suspect the difficulty in replicating the analysis is due to a combination of using different sequences, a different alignment and the inherently unreliable rooting of the EBOV clade using highly divergent sequences from other ebolavirus clades. We have uploaded the alignments we used (whole genome, coding and non-coding) to a GitHub repository at https://github.com/evogytis/ebolaGuinea2014.

We also compiled a dataset containing only the glycoprotein (GP) sequences, for which more sequences are available. Many of the extra sequences come from wild ape carcasses[8] in Gabon and RC.

These sequences were analyzed in BEAST[9] to establish a time frame for the split of the Guinea viruses from other EBOV lineages. The data were analyzed using the GTR+Γ nucleotide substitution model, an uncorrelated relaxed molecular clock (following a lognormal distribution)[10] and under different demographic models (constant population size, exponential growth or the non-parametric Bayesian skyride[11]).

GP sequence results were recovered from a relaxed molecular clock analysis, under an exponential growth tree prior (as it can accommodate a constant population size scenario when the growth rate is 0) but the analysis was found to be quite robust to different demographic models.

Analysis

An alignment of complete genomes and a maximum likelihood tree (PhyML) appears to confirm the phylogenetic position shown in the recent paper[1] (Figure 1), albeit the position of the Guinea outbreak sequences is not very well supported.

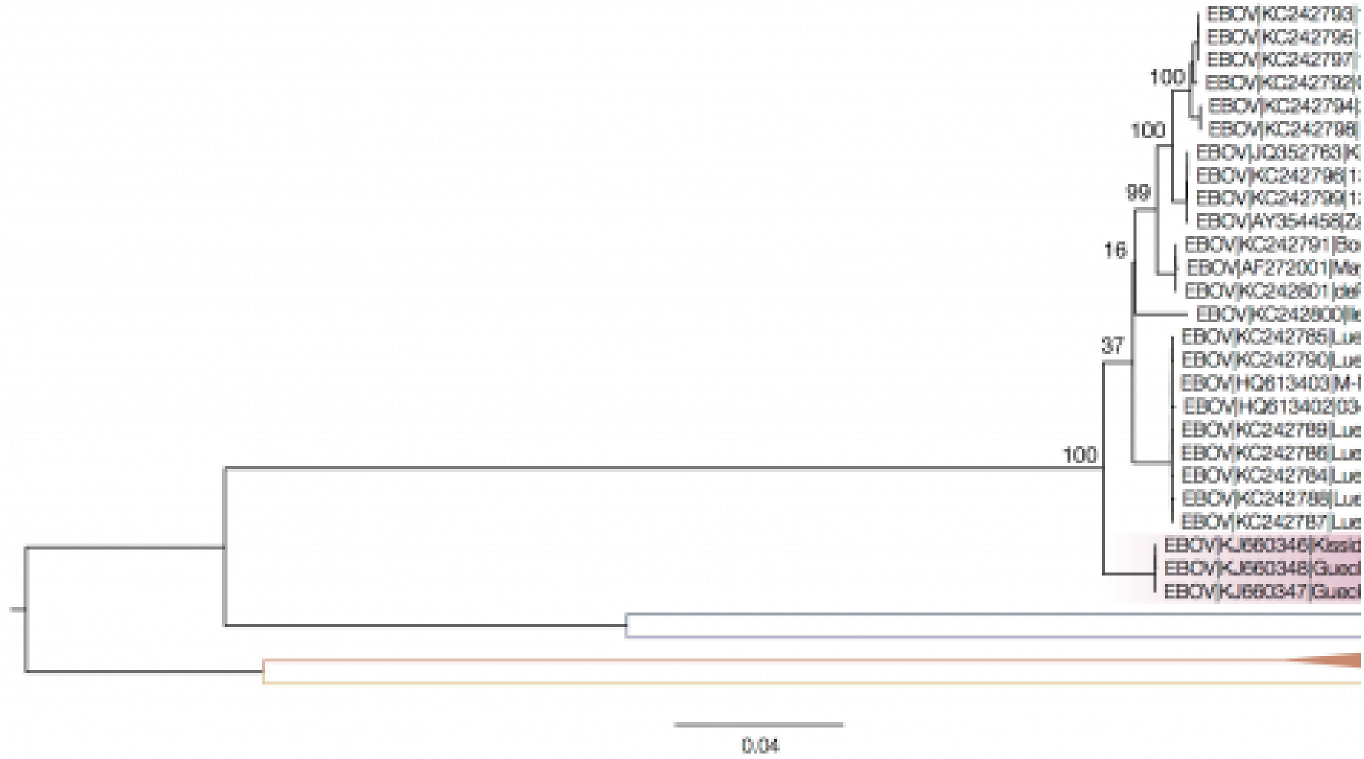

Fig. 1: ML tree of complete genomes.

ML tree of complete genomes without accommodating for rate heterogeneity shows the Guinea outbreak sequences (highlighted) as belonging to a divergent EBOV lineage. Tips belonging to the EBOV lineage are not collapsed. Numbers above key nodes in the EBOV clade are bootstrap values (100 replicates).

When the intergenic sequences are removed, however, the Guinea outbreak sequences fall within the diversity of Zaire ebolavirus (Figure 2).

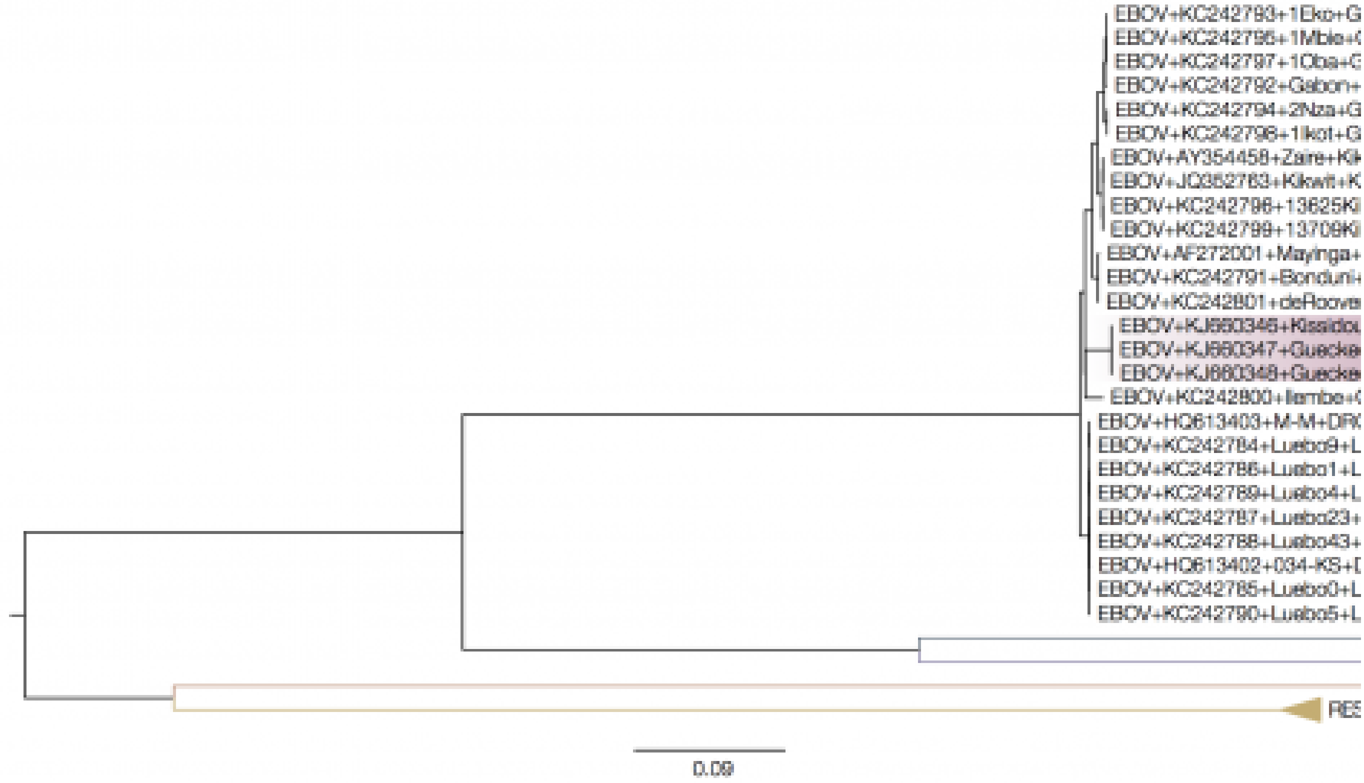

Fig. 2: MrBayes tree of concatenated coding sequences.

When only the coding sequences are used, the Guinea outbreak sequences appear to be derived from within the diversity of Gabon/DRC EBOV lineages.

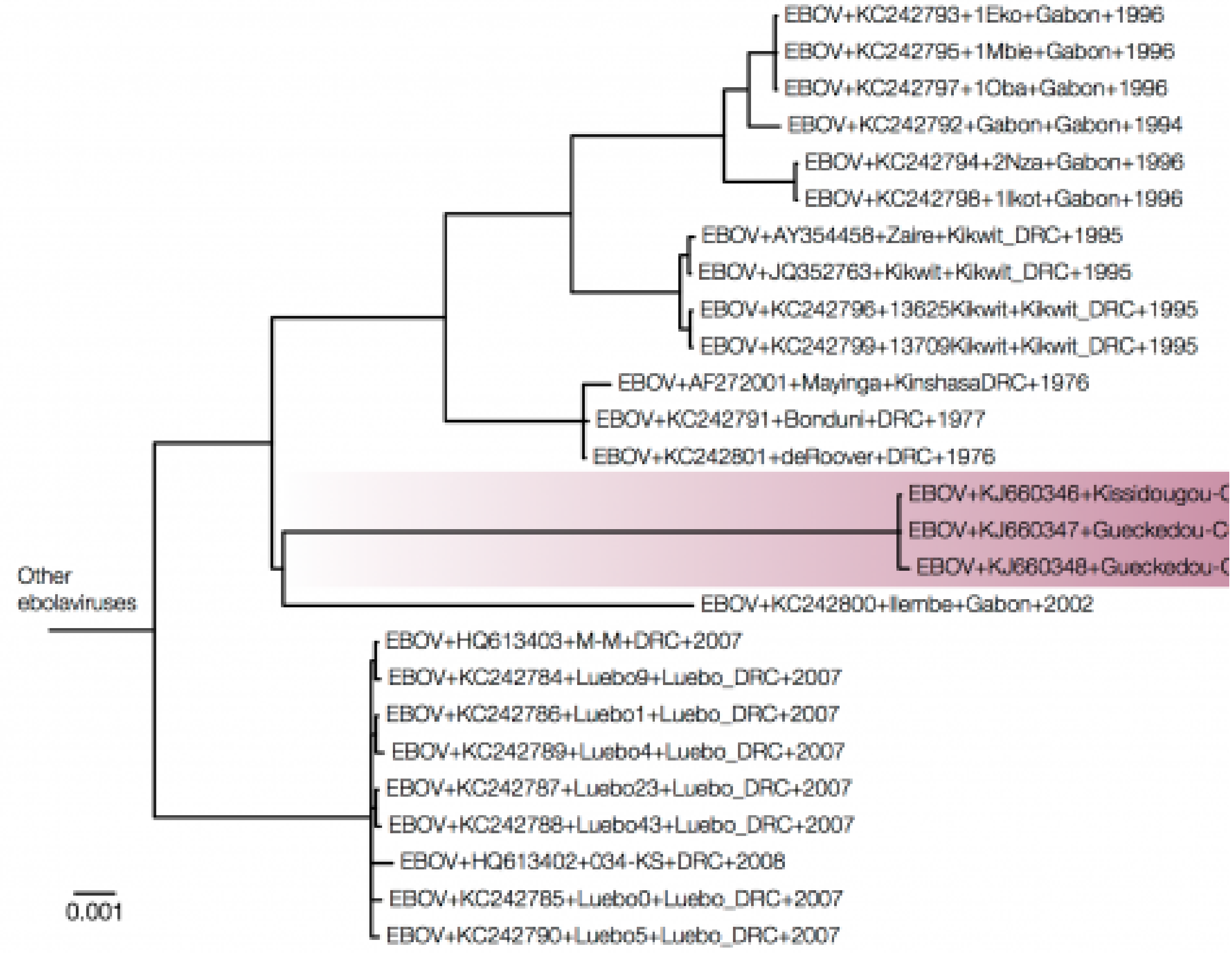
EBOV+KC242793+1Eko+Gabon+1996
EBOV+KC242795+1Mbie+Gabon+1996
EBOV+KC242797+1Oba+Gabon+1996
EBOV+KC242792+Gabon+Gabon+1994
EBOV+KC242794+2Nza+Gabon+1996
EBOV+KC242798+1Ikot+Gabon+1996
EBOV+AY354458+Zaire+Kikwit_DRC+1995
EBOV+JQ352763+Kikwit+Kikwit_DRC+1995
EBOV+KC242796+13625Kikwit+Kikwit_DRC+1995
EBOV+KC242799+13709Kikwit+Kikwit_DRC+1995
EBOV+AF272001+Mayinga+KinshasaDRC+1976
EBOV+KC242791+Bonduni+DRC+1977
EBOV+KC242801+deRoover+DRC+1976
EBOV+KJ660346+Kissidougou-C
EBOV+KJ660347+Gueckedou-C
EBOV+KJ660348+Gueckedou-C
Other ebolaviruses
EBOV+KC242800+Ilembe+Gabon+2002
EBOV+HQ613403+M-M+DRC+2007
EBOV+KC242784+Luebo9+Luebo_DRC+2007
EBOV+KC242786+Luebo1+Luebo_DRC+2007
EBOV+KC242789+Luebo4+Luebo_DRC+2007
EBOV+KC242787+Luebo23+Luebo_DRC+2007
EBOV+KC242788+Luebo43+Luebo_DRC+2007
EBOV+HQ613402+034-KS+DRC+2008
EBOV+KC242785+Luebo0+Luebo_DRC+2007
EBOV+KC242790+Luebo5+Luebo_DRC+2007
0.001

Fig. 3: Expanded view of a MrBayes tree of concatenated coding sequences.

Expanding the EBOV region of the tree (same tree as Figure 2, but with the divergent ebolavirus species cropped out) we see that the Guinea outbreak sequences are nested within the EBOV clade.

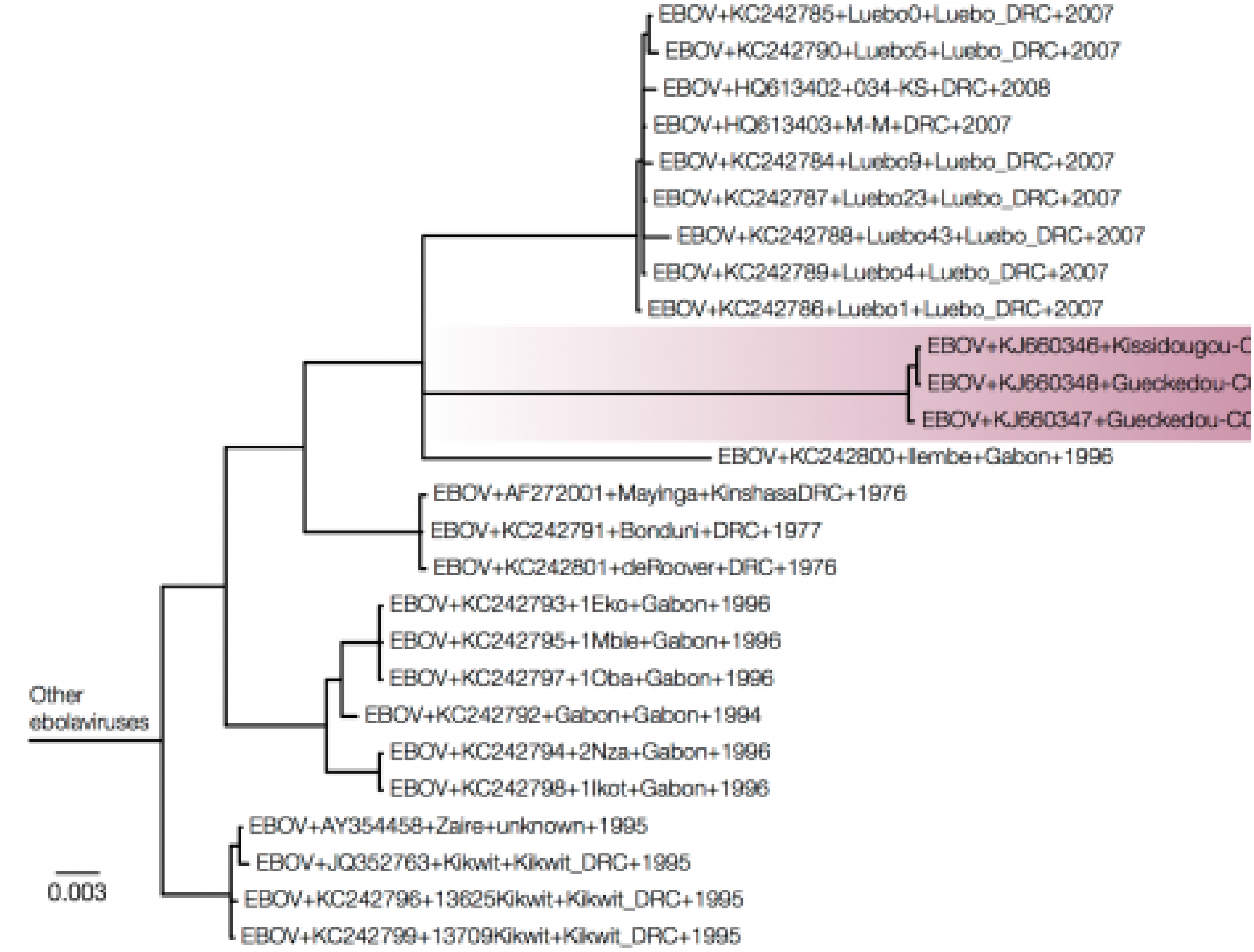
EBOV+KC242785+Luebo0+Luebo_DRC+2007
EBOV+KC242790+Luebo5+Luebo_DRC+2007
EBOV+HQ613402+034-KS+DRC+2008
EBOV+HQ613403+M-M+DRC+2007
EBOV+KC242784+Luebo9+Luebo_DRC+2007
EBOV+KC242787+Luebo23+Luebo_DRC+2007
EBOV+KC242788+Luebo43+Luebo_DRC+2007
EBOV+KC242789+Luebo4+Luebo_DRC+2007
EBOV+KC242786+Luebo1+Luebo_DRC+2007
EBOV+KJ660346+Kissidougou-C
EBOV+KJ660348+Gueckedou-C
EBOV+KJ660347+Gueckedou-C
EBOV+KC242800+Ilembe+Gabon+1996
EBOV+AF272001+Mayinga+KinshasaDRC+1976
EBOV+KC242791+Bonduni+DRC+1977
EBOV+KC242801+deRoover+DRC+1976
EBOV+KC242793+1Eko+Gabon+1996
EBOV+KC242795+1Mbie+Gabon+1996
EBOV+KC242797+1Oba+Gabon+1996
EBOV+KC242792+Gabon+Gabon+1994
EBOV+KC242794+2Nza+Gabon+1996
EBOV+KC242798+1Ikot+Gabon+1996
EBOV+AY354458+Zaire+unknown+1995
EBOV+JQ352763+Kikwit+Kikwit_DRC+1995
EBOV+KC242796+13625Kikwit+Kikwit_DRC+1995
EBOV+KC242799+13709Kikwit+Kikwit_DRC+1995
Other ebolaviruses
0.003

Fig. 4: MrBayes tree of intergenic sequences.

Intergenic regions show a similar picture with the Guinea sequences nested within EBOV.

EBOV lineages are rather poorly sampled and sequences from most outbreaks, because of the nature of the outbreaks, have nearly identical sequences. The branch leading to the Guinea outbreak is long, not because it is a divergent lineage but because it is the most recently sampled so has had the most time to evolve. Combined with a very divergent outgroup this leads to a situation where the root position of the EBOV clade is unreliably estimated.

Figures 3 and 4 show MrBayes trees from protein coding and intergenic regions of the EBOV genome, respectively, with more divergent ebolavirus strains cropped out. Note that trees in Figures 3 and 4 are essentially identical but differ by where the other ebolavirus species root the EBOV clade (on the 2007 Gabon outbreak for the coding regions in Figure 3 and on the 1995 Kikwit outbreak for the intergenic regions in Figure 4). This shows that the rooting of this clade using the highly divergent other ebolavirus species is very problematic.

However, EBOV is estimated to evolve at about 7×10^{-4} substitutions per site per year[12] which means that the virus will accumulate significant amounts of substitutions over the nearly 40 years since the first recorded outbreak in 1976. We can use this to root the EBOV tree and look at where the Guinea outbreak lies. Path-O-Gen (available at http://tree.bio.ed.ac.uk/software/pathogen/) was used to find the root that gave the best association between genetic divergence and time.

The relationship between genetic divergence and time after rooting the tree using least squares regression is shown in Figure 5.

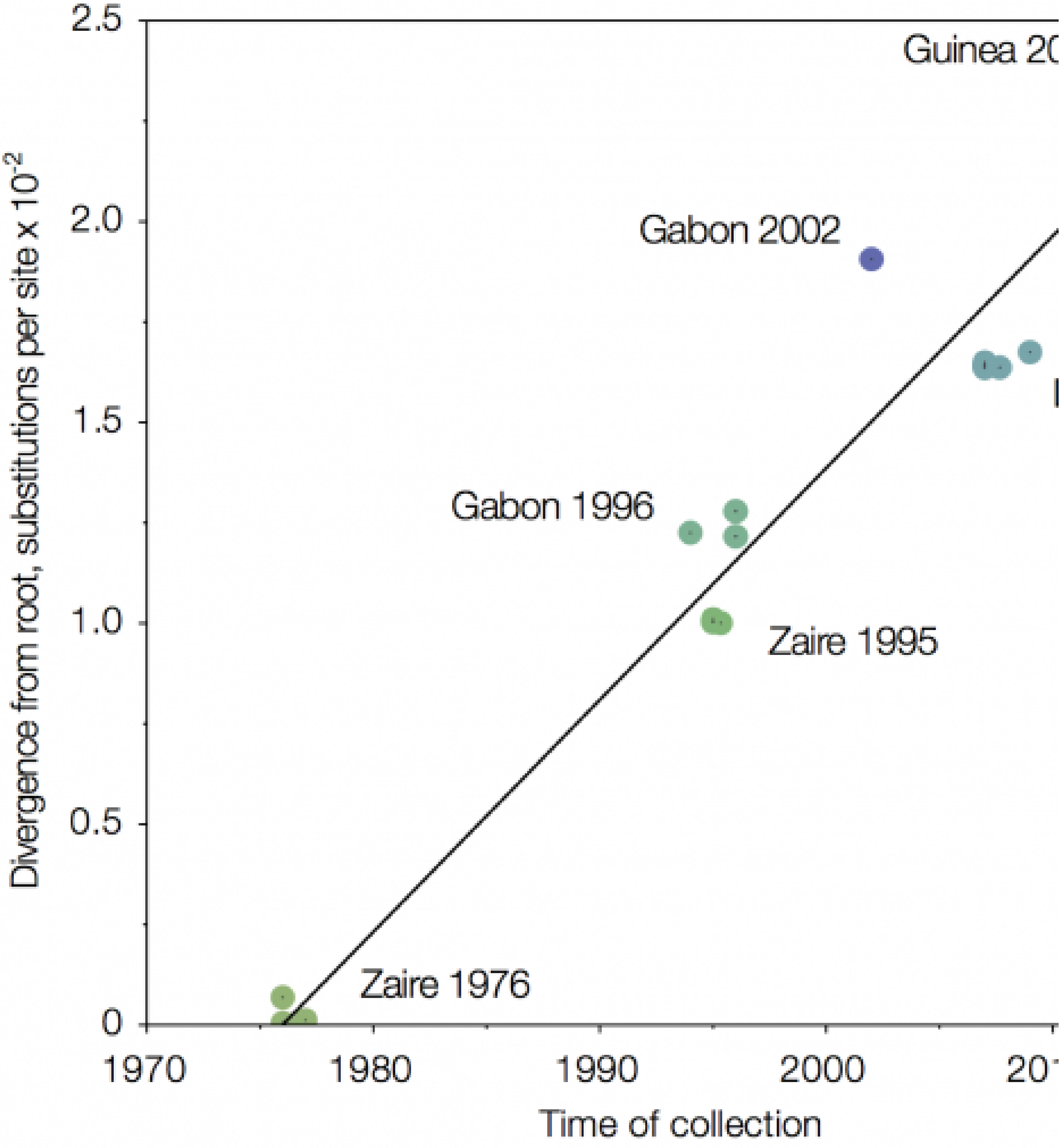

Fig. 5: Root-to-tip regression of a MrBayes tree of concatenated coding regions.

Sequences from the 1976 Zaire outbreak are very close to the root.

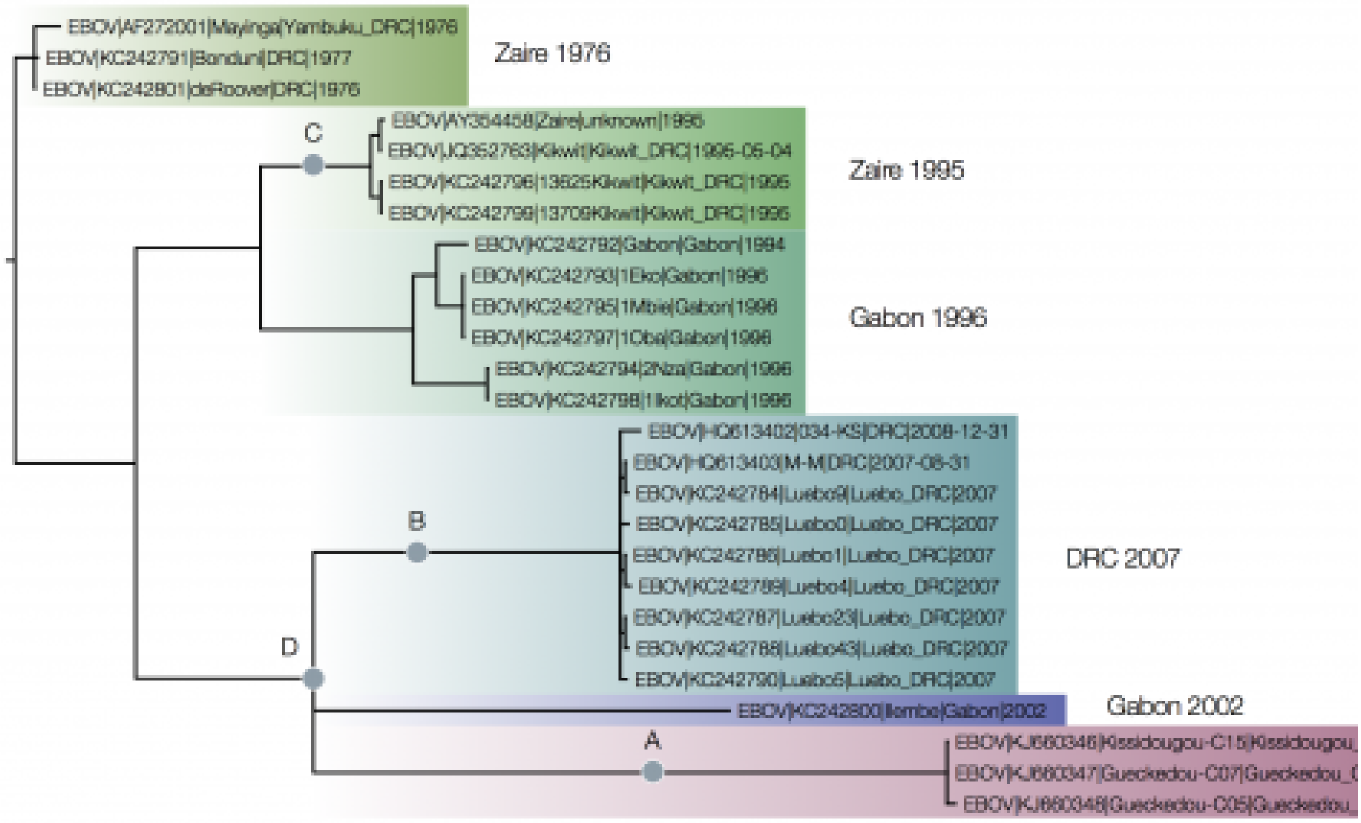
EBOV|AF272001|Mayinga|Yambuku_DRC|1976
EBOV|KC242791|Bonduni|DRC|1977
EBOV|KC242801|deRoover|DRC|1976
Zaire 1976
C
EBOV|AY354458|Zaire|unknown|1995
EBOV|JQ352763|Kikwit|Kikwit_DRC|1995-05-04
EBOV|KC242796|13625Kikwit|Kikwit_DRC|1995
EBOV|KC242799|13709Kikwit|Kikwit_DRC|1995
Zaire 1995
EBOV|KC242792|Gabon|Gabon|1994
EBOV|KC242793|1Eko|Gabon|1996
EBOV|KC242795|1Mbie|Gabon|1996
EBOV|KC242797|1Oba|Gabon|1996
EBOV|KC242794|2Nza|Gabon|1996
EBOV|KC242798|1Ikot|Gabon|1996
Gabon 1996
EBOV|HQ613402|034-KS|DRC|2008-12-31
EBOV|HQ613403|M-M|DRC|2007-08-31
EBOV|KC242784|Luebo9|Luebo_DRC|2007
EBOV|KC242785|Luebo0|Luebo_DRC|2007
EBOV|KC242786|Luebo1|Luebo_DRC|2007
EBOV|KC242789|Luebo4|Luebo_DRC|2007
EBOV|KC242787|Luebo23|Luebo_DRC|2007
EBOV|KC242788|Luebo43|Luebo_DRC|2007
EBOV|KC242790|Luebo5|Luebo_DRC|2007
B
DRC 2007
D
EBOV|KC242800|Ilembe|Gabon|2002
Gabon 2002
A
EBOV|KJ660346|Kissidougou-C15|Kissidougou_
EBOV|KJ660347|Gueckedou-C07|Gueckedou_
EBOV|KJ660348|Gueckedou-C05|Gueckedou_

Fig. 6: MrBayes tree of concatenated coding sequences rooted by least squares regression.

The Bayesian posterior support for all the groupings between the outbreaks are 1.0 including for the grouping of Guinea 2014 with DRC 2007 and Gabon 2002. This demonstrates that the uncertainty about the position of the Guinea 2014 lineage in the complete ebolavirus trees was down to the rooting of the EBOV clade (*i.e.*, where the divergent outgroups connect to the EBOV tree). The relationships of the EBOV outbreaks is completely consistent for the simple whole genome alignment, the coding regions only and the intergenic regions only but the position of the root changes. In the figure A) denotes the position of the root for the full genome maximum likelihood tree, B) for the Bayesian coding-sequence only tree, C) the Bayesian intergenic regions only tree and D) the combined coding-sequence and intergenic region accommodating different rates of evolution.

Figure 6 shows the phylogeny of the coding sequences recovered by MrBayes (a maximum likelihood tree using PhyML gave an almost identical tree) rooted by least squares regression. The root of this tree is very close to the earliest sequences from the 1976 Zaire outbreak.

Estimating the date of introduction of EBOV into Guinea

The analysis of GP sequences in BEAST revealed rooting consistent with that found in Figure 6 as well as a nucleotide substitution rate (mean of lognormal distribution from which the rates were drawn is 1.07×10^{-3} substitutions per site per year, 95% HPD interval 5.99×10^{-4} – 1.75×10^{-3}) on a scale expected, given previously published rates[12] and the fact that GP codes for a surface glycoprotein.

In Figure 7 the estimate of the split between the lineage now causing an outbreak in Guinea and the Central African lineage that had caused outbreaks in DRC and Gabon is late 2002 (95% HPD interval 2000 – 2006). This gives us a lower boundary on the introduction of Central African lineage of EBOV into Guinea, although these estimates should be interpreted with caution. We also find very good support for the common ancestry of Guinea and DRC/Gabon lineages (posterior probability = 1.0).

Figure 7 also highlights the importance of environmental sampling – many sequences in the tree come from ape carcasses and are more diverse (not shown) than sequences from human outbreaks, giving this dataset much better resolution.

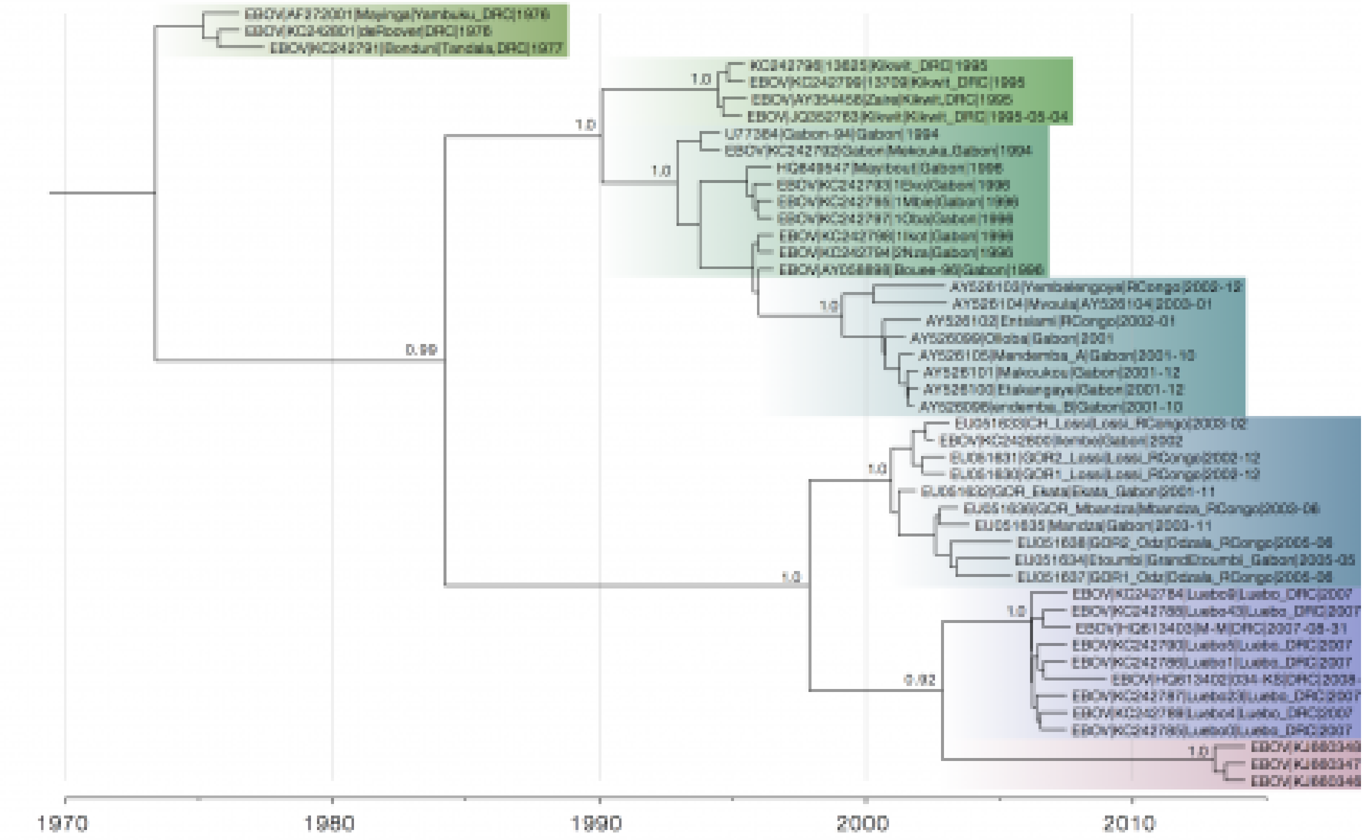
EBOV|AF272001|Mayinga|Yambuku_DRC|1976
EBOV|KC242801|deRoover|DRC|1976
EBOV|KC242791|Bonduni|Tandala,DRC|1977
KC242796|13625|Kikwit_DRC|1995
EBOV|KC242799|13709|Kikwit_DRC|1995
EBOV|AY354458|Zaire|Kikwit,DRC|1995
EBOV|JQ352763|Kikwit|Kikwit_DRC|1995-05-04
U77384|Gabon-94|Gabon|1994
EBOV|KC242792|Gabon|Makouka_Gabon|1994
HQ848547|Mayibout|Gabon|1996
EBOV|KC242793|1Eko|Gabon|1996
EBOV|KC242795|1Mbie|Gabon|1996
EBOV|KC242797|1Oba|Gabon|1996
EBOV|KC242798|1Ikot|Gabon|1996
EBOV|KC242794|2Nza|Gabon|1996
EBOV|AY058898|Boue-96|Gabon|1996
AY526103|Yembelengoye|RCongo|2002-12
AY526104|Mvoula|AY526104|2003-01
AY526102|Entsiami|RCongo|2002-01
AY526099|Olloba|Gabon|2001
AY526105|Mendemba_A|Gabon|2001-10
AY526101|Makoukou|Gabon|2001-12
AY526100|Etakangaye|Gabon|2001-12
AY526098|endemba_B|Gabon|2001-10
EU051633|CH_Lossi|Lossi_RCongo|2003-02
EBOV|KC242800|Ilembe|Gabon|2002
EU051631|GOR2_Lossi|Lossi_RCongo|2002-12
EU051630|GOR1_Lossi|Lossi_RCongo|2002-12
EU051632|GOR_Ekata|Ekata_Gabon|2001-11
EU051636|GOR_Mbandza|Mbandza_RCongo|2003-06
EU051635|Mandza|Gabon|2003-11
EU051638|GOR2_Odz|Odzala_RCongo|2005-06
EU051634|Etoumbi|GrandEtoumbi_Gabon|2005-05
EU051637|GOR1_Odz|Odzala_RCongo|2005-06
EBOV|KC242784|Luebo9|Luebo_DRC|2007
EBOV|KC242785|Luebo43|Luebo_DRC|2007
EBOV|HQ613403|M-M|DRC|2007-08-31
EBOV|KC242790|Luebo5|Luebo_DRC|2007
EBOV|KC242786|Luebo1|Luebo_DRC|2007
EBOV|HQ613402|034-KS|DRC|2008
EBOV|KC242787|Luebo23|Luebo_DRC|2007
EBOV|KC242789|Luebo4|Luebo_DRC|2007
EBOV|KC242788|Luebo0|Luebo_DRC|2007
EBOV|KJ660348
EBOV|KJ660347
EBOV|KJ660346
1.0
1.0
1.0
0.99
1.0
1.0
1.0
1.0
0.92
1.0
1970
1980
1990
2000
2010

Fig. 7: Maximum clade credibility tree of GP sequences.

Although the closest relatives of the Guinea lineage are not entirely certain (posterior probability 0.92), its relationship with Central African EBOV lineages is well-supported (posterior probability 1.0).

Conclusion

The phylogenetic analysis of the five ebolavirus species here does not substantially improve on that presented by Baize *et al*.[1] in that even when partitioning the alignment into coding and non-coding regions we get inconsistent rooting positions for the EBOV clade. We believe that at present no suitable outgroup sequences to root the EBOV phylogeny exist and that a temporal rooting gives the most consistent results.

This approach indicates that the outbreak in Guinea is likely caused by a Zaire ebolavirus lineage that has spread from Central Africa into Guinea and West Africa in recent decades, and does not represent the emergence of a divergent and endemic virus.

As the GP sequences show, without more diverse sequences, especially those from the animal reservoir, it is difficult to narrow down the estimates of when and through what means the Central African EBOV lineage has been introduced into West Africa.

Competing Interests

The authors have declared that no competing interests exist.

Acknowledgements

We would like to thank Stephan Günther and his co-authors for making the sequences from the Guinea outbreak available.

References

1. Baize S, Pannetier D, Oestereich L, Rieger T, Koivogui L, Magassouba N, Soropogui B, Sow MS, Keïta S, De Clerck H, Tiffany A, Dominguez G, Loua M, Traoré A, Kolié M, Malano ER, Heleze E, Bocquin A, Mély S, Raoul H, Caro V, Cadar D, Gabriel M, Pahlmann M, Tappe D, Schmidt-Chanasit J, Impouma B, Diallo AK, Formenty P, Van Herp M, Günther S. Emergence of Zaire Ebola Virus Disease in Guinea - Preliminary Report. N Engl J Med. 2014 Apr 16. PubMed PMID:24738640.

2. Walsh PD, Biek R, Real LA. Wave-like spread of Ebola Zaire. PLoS Biol. 2005 Nov;3(11):e371. PubMed PMID:16231972.

3. Biek R, Walsh PD, Leroy EM, Real LA. Recent common ancestry of Ebola Zaire virus found in a bat reservoir. PLoS Pathog. 2006 Oct;2(10):e90. PubMed PMID:17069458.

4. Sanchez A, Kiley MP, Holloway BP, Auperin DD. Sequence analysis of the Ebola virus genome: organization, genetic elements, and comparison with the genome of Marburg virus. Virus Res. 1993 Sep;29(3):215-40. PubMed PMID:8237108.

5. Guindon S, Gascuel O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol. 2003 Oct;52(5):696-704. PubMed PMID:14530136.

6. Huelsenbeck JP, Ronquist F. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics. 2001 Aug;17(8):754-5. PubMed PMID:11524383.

7. Tavare S. Some probabilistic and statistical problems in the analysis of DNA sequences. American Mathematical Society: Lectures on Mathematics in the Life Sciences. 1986;17: 57-86.

8. Wittmann TJ, Biek R, Hassanin A, Rouquet P, Reed P, Yaba P, Pourrut X, Real LA, Gonzalez JP, Leroy EM. Isolates of Zaire ebolavirus from wild apes reveal genetic lineage and recombinants. Proc Natl Acad Sci U S A. 2007 Oct 23;104(43):17123-7. PubMed PMID:17942693.

9. Drummond AJ, Suchard MA, Xie D, Rambaut A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol. 2012 Aug;29(8):1969-73. PubMed PMID:22367748.

10. Drummond AJ, Ho SY, Phillips MJ, Rambaut A. Relaxed phylogenetics and dating with confidence. PLoS Biol. 2006 May;4(5):e88. PubMed PMID:16683862.

11. Minin VN, Bloomquist EW, Suchard MA. Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. Mol Biol Evol. 2008 Jul;25(7):1459-71. PubMed PMID:18408232.

12. Carroll SA, Towner JS, Sealy TK, McMullan LK, Khristova ML, Burt FJ, Swanepoel R, Rollin PE, Nichol ST. Molecular evolution of viruses of the family Filoviridae based on 97 whole-genome sequences. J Virol. 2013 Mar;87(5):2608-16. PubMed PMID:23255795.