Current data show no signal of Ebola virus adapting to humans

Stephanie J. Spielman[1*], Austin G. Meyer[1,2*], and Claus O. Wilke[1]

November 13, 2014

Address:
[1]Department of Integrative Biology, Center for Computational Biology and Bioinformatics, and Institute of Cellular and Molecular Biology. The University of Texas at Austin, Austin, TX 78712, USA.
[2] School of Medicine, Texas Tech University Health Sciences Center, Lubbock, TX 79430, USA
*Authors contributed equally to this work.

Abstract

Gire et al. (*Science* 345:1369–1372, 2014) analyzed 81 complete genomes sampled from the 2014 Zaire ebolavirus (EBOV) outbreak and claimed that the virus is evolving far more rapidly in the current outbreak than it has been between previous outbreaks. This assertion has received widespread attention, and many have perceived Gire et al. (2014)'s results as implying rapid adaptation of EBOV to humans during the current outbreak. Here, we show that, on the contrary, sequence divergence in EBOV is rather limited, and that the currently available data contain no signal of rapid evolution or adaptation to humans. Gire et al.'s findings resulted from an incorrect application of a molecular-clock model to a population of sequences with minimal divergence and segregating polymorphisms. Our results highlight how indiscriminate use of off-the-shelf analysis techniques may result in highly-publicized, misleading statements about an ongoing public health crisis.

Zaire ebolavirus (EBOV) is currently devastating West African populations in an unprecedented epidemic that has begun to spill over into many parts of the world. Since its discovery in the 1970s, EBOV has, at regular intervals, caused zoonotic outbreaks in human populations. Unlike past outbreaks, however, the current EBOV outbreak shows sustained transmission among humans, prompting concerns that as the outbreak escalates, EBOV may evolve to become endemic in humans. Recently, Gire et al. (2014) published 78 genomes from the current outbreak, sampled during May and June of 2014 in Sierra Leone. They analyzed these new data, in combination with three EBOV genomes collected in Guinea during March 2014 (Baize et al., 2014), and reported "a rapid accumulation of...genetic variation." They additionally stated that "[t]he observed substitution rate is roughly twice as high within the 2014 outbreak as between outbreaks." The conclusions ultimately left readers, and indeed the scientific community at large, with the impression that EBOV is fast-evolving and possibly adapting to humans (see also Check Hayden 2014; Łuksza et al. 2014).

By contrast, we do not find any robust evidence in the available 2014-outbreak EBOV genomes supporting these conclusions. While it is clear that mutations are certainly occurring in EBOV, the available genomic data do not show concrete evidence that EBOV is evolving rapidly, much less accumulating mutations. In fact, among the 81 published genomes from the current outbreak, there are only 29 unique sequences (two from Guinea and 27 from Sierra Leone), and no genome contains more than 2 nonsynonymous mutations.

To put EBOV's evolutionary dynamics into context, we compared the extent of genetic diversity within 2014-outbreak EBOV genes to the genetic diversity accumulated by influenza H3N2. This

virus, known as the "Hong-Kong flu," has been the dominant seasonal influenza strain since its emergence in 1968, and it is the archetypal rapidly-adapting human virus. Like EBOV, influenza virus is a negative-sense single-stranded RNA virus. We constructed gene trees for EBOV nucleoprotein (*np*) and polymerase (*l*), selecting only genes from the 2014-outbreak EBOV sequences (Figure 1A). *np* and *l* have accumulated the most sequence diversity of all seven EBOV genes during this outbreak, and thus they likely represent the most rapidly-evolving EBOV genes. We constrasted these phylogenies with gene trees for H3N2 hemagglutinin (*HA*) and nucleoprotein (*NP*) (Figure 1B) for sequences collected within a single month (December 2012) in a small geographic area (Boston, MA). *HA*, the influenza surface protein responsible for host receptor-binding, is under intense selection pressure to evade host immunity, and at least 5% of *HA* sites experience positive selection (Bush et al., 1999; Meyer and Wilke, 2013). *NP* is not exposed on the viral envelope and evolves relatively slowly, with only 0.6% sites under positive selection (Lin et al., 2011).

As the juxtaposed gene trees in Figures 1A and 1B clearly demonstrate, even the fastest-evolving EBOV sequences from the 2014 outbreak are evolving much more slowly than H3N2 sequences do at similar temporal and geographic scales. Indeed, even the slowly-evolving influenza *NP* far outpaces EBOV genes in terms of accumulated genetic diversity. Furthermore, the average root-to-tip and pairwise distances among the EBOV genes *np* and *l* from the current outbreak are approximately an order magnitude lower than the distances observed for the H3N2 genes *HA* and *NP* (Table 1). Thus, EBOV is simply not accumulating mutations in a manner we would expect from a fast-evolving virus.

In fact, even considering all EBOV outbreaks since 1976, we find very limited evidence for evolutionary divergence, on par with the divergence observed in influenza virus in a single month (Table 1). Figure 1C shows a phylogeny of all EBOV sequences considered in Gire et al. (2014), as well as related ebolavirus species. The extent of sequence divergence across EBOV sequences collected from 1976 to present pales in comparison to divergence among ebolavirus species. The minimal sequence divergence among, and within, EBOV outbreaks indicates that the current EBOV outbreak should be considered a single, polymorphic population. Indeed, the mean pairwise sequence similarity among unique 2014-outbreak EBOV genomes is 99.84% (standard deviation of 0.39%). Moreover, according to Gire et al. (2014), there are least 55 segregating mutations in the 2014-outbreak EBOV sequences, yet these mutations generally do not co-occur in any particular genome. For example, among these EBOV genome sequences, only three (two from Sierra Leone and one from Guinea) contain two nonsynonymous mutations, no genome contains three or more nonsynonymous changes, and there is no evidence that any given site has experienced multiple mutation events in the current outbreak.

Taken together, these results reveal that EBOV sequence data must be analyzed in a population genetics, rather than a purely phylogenetic, context, which Gire et al. (2014) failed to recognize. For example, Gire et al. (2014) found that the 2014-outbreak EBOV substitution rate is elevated relative to a baseline EBOV substitution rate, measured from pooling all sequence data collected since 1976 (Figure 4F in Gire et al. 2014), by inferring the rate of a strict molecular clock. Given the minimal divergence of the 2014 sequences, this approach is fundamentally flawed. The molecular clock assumes that sequences have sufficiently diverged such that all differences are fixed substitutions rather than segregating polymorphisms. As a consequence, the rate of the molecular clock is highly time-dependent, such that the substitution rate is substantially elevated at short time-scales due to the confounding presence of segregating polymorphisms (Ho et al., 2005, 2007; Peterson and Masel, 2009; Ho et al., 2011). Thus, that Gire et al. (2014) observed an elevated substitution rate within the 2014-outbreak sequences is actually an expected outcome and does not provide any evidence for accelerated evolution. Instead, it is merely a signal of the minuscule divergence among these sequences. Until more time has passed and mutations have either fixed or been removed from the

population, results concerning EBOV substitution rate are unreliable and inconclusive.

In sum, we do not find any convincing evidence in Gire et al. (2014)'s study supporting their conclusion that EBOV 2014 sequences are rapidly evolving. As more sequence data from the current outbreak are collected and made available for analysis, evidence may emerge to support these claims. However, until such data are released, we cannot conclude that EBOV 2014 sequences show signatures of increased evolutionary rate, much less of adaptation to humans.

Acknowledgements

This work was supported in part by DTRA grant HDTRA1-12-C-0007 and NSF Cooperative Agreement No. DBI-0939454 (BEACON Center). Computational resources were provided by the University of Texas at Austin's Center for Computational Biology and Bioinformatics (CCBB).

References

Baize S, Pannetier D, Oestereich L, Rieger T, Koivogui L, Magassouba N, Soropogui B, Sow M S, Keta S, De Clerck H, Tiffany A, Dominguez G, Loua M, Traor A, Koli M, Malano E R, Heleze E, Bocquin A, Mly S, Raoul H, Caro V, Cadar D, Gabriel M, Pahlmann M, Tappe D, Schmidt-Chanasit J, Impouma B, Diallo A K, Formenty P, Van Herp M, Gnther S. 2014. Emergence of zaire ebola virus disease in guinea. NEJM 371:1418–1425.

Bush R M, Bender C A, Subbarao K, Cox N J, Fitch W M. 1999. Predicting the evolution of human influenza A. Science 286:1921–1925.

Check Hayden E. 2014. Ebola virus mutating rapidly as it spreads. *Nature* News, 28 August 2014. doi:10.1038/nature.2014.15777.

Gire S, Goba A, Andersen K, Sealfon R S, Park D, Kanneh L, Jalloh S, Momoh M, Fullah M, Dudas G, Wohl S, Moses L, Yozwiak N, Winnicki S, Matranga C, Malboeuf C, Qu J, Gladden A, Schaffner S, Yang X, Jiang P, Nekoui M, Colubri A, Coomber M, Fonnie M, Moigboi A, Gbakie M, Kamara F, Tucker V, Konuwa E, Saffa S, Sellu J, Jalloh A, Kovoma A, Koninga J, Mustapha I, Kargbo K, Foday M, Yillah M, Kanneh F, Robert W, Massally J L, Chapman S, Bochicchio J, Murphy C, Nusbaum C, Young S, Birren B, Grant D, Scheiffelin J, Lander E, Happi C, Gevao S, Gnirke A, Rambaut A, Garry R, Khan S, Sabeti P C. 2014. Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. Science 345:1369–1372.

Ho S Y, Phillips M J, Cooper A, Drummond A J. 2005. Time Dependency of Molecular Rate Estimates and Systematic Overestimation of Recent Divergence Times. Mol Biol Evol 22:1561–1568.

Ho S Y W, Lanfear R, Bromham L, Phillips M J, Soubrier J, Rodrigo A G, Cooper A. 2011. Time-dependent rates of molecular evolution. Mol Ecol 20:3087–3101.

Ho S Y W, Shapiro B, Phillips M J, Cooper A, Drummond A J. 2007. Evidence for time dependency of molecular rate estimates. Syst Biol 56:515–522.

Lin J H, Chiu S C, Cheng J C, Chang H W, Hsiao K L, Lin Y C, Wu H S, Salemi M, Liu H F. 2011. Phylodynamics and molecular evolution of influenza A virus nucleoprotein genes in Taiwan between 1979 and 2009. PLoS ONE 6:e23454.

Łuksza M, Bedford T, Lässig M. 2014. Epidemiological and evolutionary analysis of the 2014 Ebola virus outbreak. http://arxiv.org/abs/1411.1722.

Meyer A G, Wilke C O. 2013. Integrating sequence variation and protein structure to identify sites under selection. Mol Biol Evol 30:36–44.

Peterson G, Masel J. 2009. Quantitative prediction of molecular clock and Ka/Ks at short timescales. Mol Biol Evol 26:2595–2603.

Squires R, et al. 2012. Influenza research database: an integrated bioinformatics resource for influenza research and surveillance. Influenza Other Respir Viruses 6:404–416.

Stamatakis A. 2014. RAxML Version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313.

Table 1: Mean root-to-tip and pairwise distances for EBOV and influenza sequences. Duplicate sequences were excluded from all distance calculations.

virus	**outbreak**	**gene**	**mean root-to-tip**	**mean pairwise**
EBOV	2014	*np*	0.00102	0.00136
		l	0.00072	0.00088
	all outbreaks (1976-2014)	*np*	0.01326	0.01701
		l	0.01396	0.01607
H3N2	Boston, MA (Dec. 2012)	*HA*	0.01461	0.01083
		NP	0.00515	0.00598

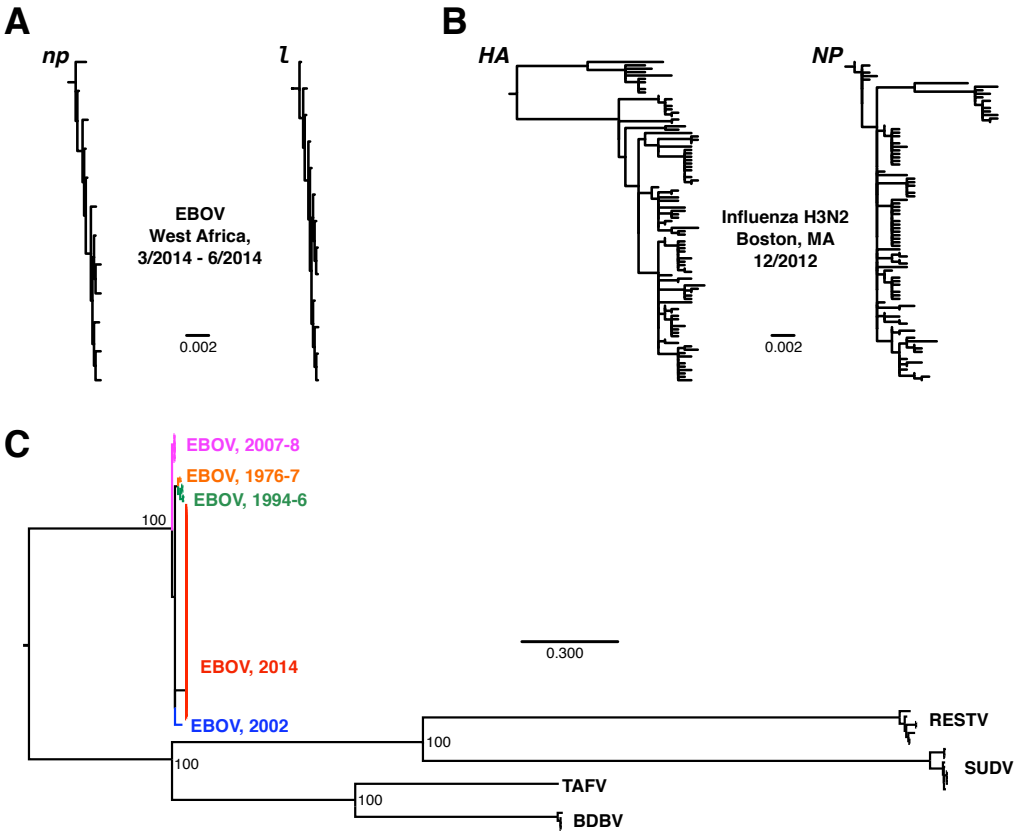

Figure 1: Limited divergence in EBOV 2014. Phylogenies in A and B were constructed from nucleotide data in RAxMLv8.1.1 (Stamatakis, 2014) under the GTR+GAMMA model. Sequences in A were restricted to 2014-outbreak EBOV sequences, and the phylogenies were rooted on *np* and *l* sequences sampled in 2002. Sequences in B were collected from the Influenza Research Database (Squires et al., 2012), with the search restricted to complete genomes sampled during December, 2012 in Boston, Massachusetts, USA. H3N2 phylogenies were rooted on the *HA* and *NP* genes from H3N2 strain A/Aichi/2/1968. Duplicate sequences were excluded from the phylogenies in parts A and B. The phylogeny in C was constructed from the genomic data of Gire et al. (2014) for all ebolavirus species, using RAxMLv.8.1.1 (Stamatakis, 2014) with the GTR+GAMMA model and a different partition for each gene. Numbers at nodes indicate bootstrap support. Abbreviations shown in the figure stand for Reston virus (RESTV), Bundibugyo virus (BDVD), Taï forest virus (TAFV), and Sudan virus (SUDV).