Infection, Genetics and Evolution 32 (2015) 51–59

Contents lists available at ScienceDirect

Infection, Genetics and Evolution

journal homepage: www.elsevier.com/locate/meegid

Identifying the pattern of molecular evolution for *Zaire ebolavirus* in the 2014 outbreak in West Africa

Si-Qing Liu [a,b], Cheng-Lin Deng [a,b], Zhi-Ming Yuan [b], Simon Rayner [a], Bo Zhang [a,b,*]

[a] Key Laboratory of Etiology and Biosafety for Emerging and Highly Infectious Diseases, Wuhan Institute of Virology, Chinese Academy of Sciences, Wuhan 430071, China
[b] Key Laboratory of Agricultural and Environmental Microbiology, Wuhan Institute of Virology, Chinese Academy of Sciences, Wuhan 430071, China

ARTICLE INFO

Article history:
Received 6 December 2014
Received in revised form 16 February 2015
Accepted 24 February 2015
Available online 4 March 2015

Keywords:
Zaire ebolavirus
Positive selection
Evolutionary rate
West Africa

ABSTRACT

The current Ebola virus disease (EVD) epidemic has killed more than all previous Ebola outbreaks combined and, even as efforts appear to be bringing the outbreak under control, the threat of reemergence remains. The availability of new whole-genome sequences from West Africa in 2014 outbreak, together with those from the earlier outbreaks, provide an opportunity to investigate the genetic characteristics, the epidemiological dynamics and the evolutionary history for *Zaire ebolavirus* (ZEBOV). To investigate the evolutionary properties of ZEBOV in this outbreak, we examined amino acid mutations, positive selection, and evolutionary rates on the basis of 123 ZEBOV genome sequences. The estimated phylogenetic relationships within ZEBOV revealed that viral sequences from the same period or location formed a distinct cluster. The West Africa viruses probably derived from Middle Africa, consistent with results from previous studies. Analysis of the seven protein regions of ZEBOV revealed evidence of positive selection acting on the GP and L genes. Interestingly, all putatively positive-selected sites identified in the GP are located within the mucin-like domain of the solved structure of the protein, suggesting a possible role in the immune evasion properties of ZEBOV. Compared with earlier outbreaks, the evolutionary rate of GP gene was estimated to significantly accelerate in the 2014 outbreak, suggesting that more ZEBOV variants are generated for human to human transmission during this sweeping epidemic. However, a more balanced sample set and next generation sequencing datasets would help achieve a clearer understanding at the genetic level of how the virus is evolving and adapting to new conditions.

1. Introduction

Ebola viruses (EBOV) are non-segmented, single stranded, linear and negative sense RNA viruses, which are members of family *Filoviridae*. The complete EBOV genomes, which are about 18.9 kb long, contain seven protein-coding genes and various intergenic regions (Sanchez et al., 2007). The functions performed by these genes have been studied in the process of viral replication, virus–host interaction, and virion assembly (Feldmann et al., 1999; Hoenen et al., 2010b; Mühlberger et al., 1998; Mateo et al., 2011; Nanbo et al., 2010, 2013). The nucleoprotein (NP) is associated with virus replication and viral assembly, together with the RNA-dependent RNA polymerase cofactor (VP35), the transcriptional activator (VP30), the RNA-dependent RNA polymerase (L), and the minor matrix protein (VP24) (Becker et al., 1998; Mühlberger et al., 1998, 1999; Watanabe et al., 2006). The surface spike-like glycoprotein (GP) plays an important role in virus entry into cells by mediating receptor binding and fusion (Feldmann et al., 1999; Lee et al., 2008; Nanbo et al., 2010). Finally, the major matrix protein (VP40), along with VP24, plays a critical role in the regulation of viral replication and transcription (Hoenen et al., 2010a,b), as well as viral assembly (Hoenen et al., 2010b; Mateo et al., 2011).

Ebola viruses are known to be the etiological agents of hemorrhagic fever, which has a high mortality rate (Baron et al., 1983; Suzuki and Gojobori, 1997). Since 1976, the *Zaire ebolavirus* (ZEBOV) has been associated with outbreaks in the Democratic Republic of Congo (DRC), Gabon, and the Republic of Congo (Baize et al., 2014; Feldmann and Geisbert, 2011). The most recent outbreak in West Africa, which began in February of 2014, is the largest recorded outbreak, spreading through Guinea, Liberia and Sierra Leone, with incursions into other countries including Nigeria, the USA and Mali (Alexander et al., 2014; Gire et al.,

* Corresponding author at: Xiao Hong Shan Zhong Qu 44, Wuhan, Hubei 430071, China. Tel.: +86 27 87197822.
E-mail addresses: liusiqing310@gmail.com (S.-Q. Liu), dengchenglin29@126.com (C.-L. Deng), yzm@wh.iov.cn (Z.-M. Yuan), simon.rayner.cn@gmail.com (S. Rayner), zhangbo@wh.iov.cn (B. Zhang).

http://dx.doi.org/10.1016/j.meegid.2015.02.024
1567-1348/

2014). The role of human population density in the spread of many viruses is well known and can lead to shorter infection cycles and the evolution of higher virulence strains (Holmes, 2004). Social factors can also impact an epidemic. For example, a recent study highlighted the role of an improved transportation infrastructure, growing social activities and a lack of public awareness in the reemergence of rabies in China in the 1990s (Song et al., 2014). In the region of West Africa, population growth has been dramatic, with population densities increasing by nearly 200% in Guinea, Sierra Leone, and Liberia in the past five decades (Alexander et al., 2014). In addition, an overburdened public health network, delayed response and coordination challenges have also facilitated the spread of ZEBOV.

A previous study demonstrated that the ZEBOV outbreaks occurring after the Yambuku outbreak in DRC in 1976 resulted from either direct or closely related offsprings of a Yambuku-like virus (Wahl-Jensen et al., 2005). Many phylogenetic studies from the sequence of available ZEBOV strains have placed the viruses from this earliest recorded outbreak at the tree root (Calvignac-Spencer et al., 2014; Dudas and Rambaut, 2014; Wahl-Jensen et al., 2005). Furthermore, these derived phylogenetic relationships indicate that the viruses of later outbreaks evolved from those of earlier outbreaks, implying an epidemiological connection amongst these outbreaks (Wahl-Jensen et al., 2005). This ladder-like phylogenetic structure has been repeatedly reconstructed in different studies (Calvignac-Spencer et al., 2014; Dudas and Rambaut, 2014; Gire et al., 2014; Wahl-Jensen et al., 2005). This structure has also been observed in many other RNA viruses and has been attributed to the pressure of positive selection (Grenfell et al., 2004; Holmes, 2004), suggesting a similar mechanism is driving the evolution of ZEBOV. Furthermore, the restricted genetic diversity and rapidly generated variants observed in earlier ZEBOV outbreaks could also be the consequence of continuous positive selection (Wahl-Jensen et al., 2005).

Before the ZEBOV outbreak recorded in 2007, the nucleotide substitution rate of isolates was estimated to be constant over time (Wahl-Jensen et al., 2005). However, compared with retroviruses and influenza A virus, ZEBOV seems to be evolving relatively slowly (Cox et al., 1983; Suzuki and Gojobori, 1997). Three explanations have been proposed for this: (1) The RNA-dependent RNA polymerase of ZEBOV may not be as error-prone as other viruses (Suzuki and Gojobori, 1997); (2) the replication frequency was relatively low in the natural (reservoir) host(s) for the ~20 years between the 1976 and 1995 outbreaks (Suzuki and Gojobori, 1997); (3) given the limited ZEBOV sampling and overburdened public health resources in parts of West Africa, as well as limited RNA genome sequencing (Matranga et al., 2014), the degree of mutational differences in earlier outbreaks could be underestimated.

ZEBOV has invaded West Africa from Middle Africa in a wave like pattern within the last decade (Wahl-Jensen et al., 2005; Baize et al., 2014). While change of environment and specific conditions in communities in West Africa could facilitate the spread of ZEBOV disease from human to human, the distinct evolutionary properties of ZEBOV in the current outbreak have not been addressed. Our aim in the present study was to investigate the genetic variation and evolutionary pattern for ZEBOV amongst the different outbreaks based on evolutionary analyses, followed by the interpretation of ZEBOV adaptation to new conditions. Specifically, we estimated amino acid mutations, tested the hypothesis of positive selection, and compared the evolutionary rates of the GP gene among each outbreak in an attempt to identify the evolutionary properties of ZEBOV during the large epidemics which occur in West Africa.

2. Materials and methods

2.1. Sequences processing and phylogenetic analysis of ZEBOV genomes

We collected all available whole genome sequences (as of 16 September 2014) of *Z. ebolavirus* for the period of 1976–2014 from the GenBank database, of which 99 were nearly complete genomes derived from the Sierra Leone outbreak in West Africa in 2014. We then removed the sequences with one or more ambiguous nucleotide sequences within the protein-coding regions. Three datasets were used for inferences of phylogenetic trees: combined non-coding sequences only (5′UTR, 3′UTR, and six intergenic regions); combined coding sequences only (seven protein-coding genes, NP, VP35, VP40, GP, VP30, VP24, and L, corresponding to an alignment of 14,516 bp); and non-coding plus coding sequences. Nucleotide sequence variation and amino acid mutation were estimated using the MEGA 4.0 software package (Tamura et al., 2007). Nucleotide sequences were initially aligned using the online MUSCLE program through the NIAID Virus Pathogen Database and Analysis Resource (ViPR, http://www.viprbrc.org).

Phylogenetic analysis was performed for each of the three datasets using the neighbor-joining (NJ) and partitioned Bayesian Analysis (BA) methods. The NJ trees were implemented in MEGA 4.0 and bootstrap analysis with 1000 replicates was used to evaluate support values for phylogenetic relationships (Felsenstein, 1985). BA analyses were implemented in MrBayes 3.1.1 (Huelsenbeck and Ronquist, 2001). Prior to analyses, the most appropriate model of nucleotide substitution and parameter values for each dataset were estimated under a nested array of substitution models using the Akaike Information Criterion (AIC) as implemented in Modeltest 3.7 (Posada and Crandall, 1998). In the Bayesian analyses, two independent searches were conducted for each dataset. Four independent Markov Monte Carlo (MCMC) chains were run for 2,000,000 generations, with sampling of one tree per 100 replicates in each run. The first 1000 trees with non-stationary log likelihood values represented "burn-in" and were discarded. Posterior probabilities (PP) of phylogenetic inferences were determined from remaining trees. Trees shown herein represent 50% majority-rule consensus trees and BA PP for each node.

2.2. Investigation of positive selection

The CODEML program within the PAML software package was used to assess parameters in models of sequence evolution and to test relevant hypotheses (Yang, 2007). We examined three pairwise codon-based substitution models to assess non-synonymous vs. synonymous substitution rates (denoted as dN/dS ratio or ω ratio) for all ZEBOV codon sites and all branches of the phylogeny: M0 (one-ratio) vs. M3 (discrete ω), M1a (nearly neutral) vs. M2a (positive selection), and M7 (β distribution) vs. M8 (β distribution and a fraction of sites with $\omega > 1$). Likelihood ratio tests (LRTs) were performed to compare the fit of two pairwise models. It is assumed that twice the log likelihood difference between nested models ($2\Delta lnL$) follows a chi-squared distribution with a number of degrees of freedom equal to the difference in the number of free parameters (Whelan and Goldman, 1999). When LRTs indicated positive selection, we used the Bayesian empirical Bayesian (BEB) approach (Yang et al., 2005) to calculate posterior probabilities for identifying sites under positive selection.

In order to confirming the results of PAML analyses, the data sets were re-analyzed using the Datamonkey web server (Delport et al., 2010; Kosakovsky Pond and Frost, 2005), which implements