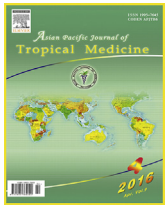
HOSTED BY
ELSEVIER

Contents lists available at ScienceDirect

Asian Pacific Journal of Tropical Medicine

journal homepage: http://ees.elsevier.com/apjtm



Original research http://dx.doi.org/10.1016/j.apjtm.2016.03.016

Genetic diversity in Ebola virus: Phylogenetic and *in silico* structural studies of Ebola viral proteins

Alba Grifoni[1], Alessandra Lo Presti[2], Marta Giovanetti[2,3], Carla Montesano[3], Massimo Amicosante[1,4], Vittorio Colizzi[3], Alessia Lai[5], Gianguglielmo Zehender[5], Eleonora Cella[2,6], Silvia Angeletti[7], Massimo Ciccozzi[2,8*]

[1]*ProxAgen Ltd, Sofia, Bulgaria*

[2]*Department of Infectious Parasitic and Immunomediated Diseases, National Institute of Health, Rome, Italy*

[3]*Department of Biology, University of Rome 'Tor Vergata', Rome, Italy*

[4]*Department of Biomedicine and Prevention, University of Rome 'Tor Vergata', Rome, Italy*

[5]*Laboratory of Infectious Diseases and Tropical Medicine, University of Milan, Italy*

[6]*Department of Public Health and Infectious Diseases, Sapienza University of Rome, Rome, Italy*

[7]*Clinical Pathology and Microbiology Laboratory, University Hospital Campus Bio-Medico of Rome, Rome, Italy*

[8]*University Campus Bio-Medico, Rome, Italy*

ARTICLE INFO

Article history:
Received 15 Jan 2016
Received in revised form 20 Feb 2016
Accepted 1 Mar 2016
Available online 9 Mar 2016

Keywords:
Ebola virus
Proteins
Evolutionary analysis

ABSTRACT

Objective: To explore the genetic diversity and the modification of antibody response in the recent outbreak of Ebola Virus.
Methods: Sequences retrieved from public databases, the selective pressure analysis and the homology modeling based on the all protein (nucleoprotein, VP35, VP40, soluble glycoprotein, small soluble glycoprotein, VP30, VP24 and polymerase) were used.
Results: Structural proteins VP24, VP30, VP35 and VP40 showed relative conserved sequences making them suitable target candidates for antiviral treatment. On the contrary, nucleoprotein, polymerase and soluble glycoprotein have high mutation frequency.
Conclusions: Data from this study point out important aspects of Ebola virus sequence variability that for epitope and vaccine design should be considered for appropriate targeting of conserved protein regions.

1. Introduction

Zaire Ebola virus (EBOV), a member of the Filoviridae family, is a virulent Category A pathogen that causes considerable morbidity and mortality. The EBOV genome is a linear, non-segmented, single-stranded RNA approximately of 19 kb. The virus is filamentous and pleomorphic with a mean unit length of 1 200 nm [1]. The viral genome encodes for a nucleoprotein (NP), a glycoprotein (GP), a RNA dependent RNA polymerase (L), and four structural proteins termed VP24, VP30, VP35 and VP40 [2–4]. The structural proteins, VP40 and VP24, represent viral matrix proteins connecting the nucleocapsid to the viral envelope. NP, VP30 and L proteins are of fundamental importance in the replication and transcription of the Ebola genome [2,5]. The envelope GP is an integral membrane protein, which forms spike-like protrusions on the surface of the virion. Recently, surface GP level regulated by RNA editing mechanism has shown a fundamental role in EBOV pathogenicity and viral immune escape [6].

EBOV causes Ebola viral disease, characterized by fever, malaise, and other nonspecific symptoms such as myalgia, headache, vomiting, and diarrhea. About 30%–50% patients manifest hemorrhagic symptoms. Moreover, in some severe cases multi-organ dysfunction, including hepatic damage, renal failure, and central nervous system involvement occur, leading to shock and death [7]. 'Cytokine storm' with immune suppression of CD4 and CD8 lymphocytes is a candidate mechanism for production of the terminal hemorrhagic fever [8].

EBOV was first identified in 1976 during the epidemic of hemorrhagic fever in Zaire, now Democratic Republic of Congo, with the epicenter of the outbreak in Yambuku. Zaire EBOV appeared again in Democratic Republic of Congo in 1977 near Yambuku and subsequent outbreaks among humans have occurred in west-central Africa in distinct waves during 1994–1997 and 2001–2005 [9]. The recent and ongoing outbreak of EBOV Disease began in December 2013 in forested areas of Southeastern Guinea

*Corresponding author: Massimo Ciccozzi, Department of Infectious Parasitic and Immunomediated Diseases, National Institute of Health, Rome, Italy.
Tel: +39 0 649903187
E-mail: ciccozzi@iss.it
Peer review under responsibility of Hainan Medical College.

affecting additionally the West African countries of Liberia, Nigeria, and Sierra Leone. In Sierra Leone, a total of 8 698 confirmed cases with 3 587 confirmed deaths were reported in the Ebola Situation Report of 2 September 2015, of which 302 (221 deaths) among health care workers [10,11]. A significant decline in both Ebola cases and deaths was observed until April 2015, although, sporadic outbreaks and deaths continue to occur, including infection among health care workers [10]. In Africa, EBOV disease infection have been documented through the handling of infected chimpanzees, gorillas, fruit bats, monkeys, forest antelope and porcupines found ill or dead or in the rainforest. Ebola spreads to the community through person-to-person transmission, with infection resulting from direct contact with the blood, secretions, organs or other bodily fluids, and indirect contact with environments contaminated [12].

Although the knowledge of clinical and pathogenic aspects of Ebola viral disease has recently improved the role of antibody response in viral clearance and protection against EBOV in humans is not fully understood. Fatal EBOV infection is characterized by a defective innate immune response, leading to uncontrolled release of inflammatory mediators and chemokines in the late stage of the disease, and correlates with the collapse of adaptive immunity with massive T and B lymphocyte apoptosis. Immune protection seems to be associated with the development of both cellular and humoral immunity [13–17].

Several amino acid differences have been characterized in the recent Ebola outbreak. A better knowledge of the viral protein structure modifications represent the key point for drug design and vaccination [18].

In this study, the selective pressure analysis was carried out to detect the presence of sites under positive selective pressure that could represent candidate 'hot spot' with a crucial rule in the viral escape and evolution. Homology modeling analysis has been performed to evaluate the virus evolution consequences in the protein recognition by host immune response. We previously performed these analyses considering only the contribution of GP protein [18].

In this paper, the study is extended on all EBOV genome transcripts to evaluate new targets for therapeutic and vaccine strategies. Bioinformatics and immune-informatic approaches can provide new insights into the pathogen's evolution, genetic diversity and heterogeneity and the related protective immune response against the virus to evaluate new targets for therapeutic and vaccine strategies.

2. Materials and methods

2.1. Sequence data set and phylogenetic analysis

Seven different data set were built, one for each protein [NP, VP35, VP40, soluble glycoprotein (sGP), VP30, VP24, L] downloading a number of sequences that ranged from 91 for NP and VP40 proteins to 101 to sGP protein. The small soluble glycoprotein (ssGP) was not investigated due to the fact that it is a portion with the same reading frame of the GP, already described [19]. All the sequences with known sampling date and geographical location were obtained from the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/).

The sampling dates for the sequences in the data set ranged from 1976 to 2014. All data sets were used to perform the selective pressure and the homology modeling analysis. All the sequences were aligned using ClustalX software and edited by using Bio-Edit software v. 7.0 [19], The best-fitting nucleotide substitution models were chosen with the hierarchical LRT strategy described by Swofford & Sullivan [20], as implemented in the MODELTEST v. 3.7 program [21].

2.2. Selective pressure analysis

Comparison of relative fixation rates of synonymous (silent) and non-synonymous (amino acid-altering) mutations provide a means for understanding the mechanisms of molecular sequence evolution. The non-synonymous/synonymous rate ratio ($\omega = d_N/d_S$) is an important indicator of selective pressure at the protein level, with $\omega = 1$ meaning neutral mutations, $\omega < 1$ purifying selection, and $\omega > 1$ diversifying positive selection.

The CODEML program implemented in the PAML 3.14 software package (http://abacus.gene.ucl.ac.uk/software/paml.html) [22] was used to investigate the adaptive evolution of the different data set of EBOV.

Six models of codon substitution: M_0 (one-ratio), M_{1a} (nearly neutral), M_{2a} (positive selection), M_3 (discrete), M_7 (beta), and M_8 (beta and omega) were used in this analysis [23]. Since these models are nested, we used codon-substitution models to fit the model to the data using the likelihood ratio test (LRT) [24]. The M_3 model, with three d_N/d_S (ω) classes, allows ω to vary among sites by defining a set number of discrete site categories, each with its own ω value. Through maximum-likelihood optimization, it is possible to estimate the ω and P values and the fraction of sites in the aligned data set that falls into a given category. Finally, the algorithm calculates the a posteriori probability of each codon belonging to a particular site category. Using the M_3 model, sites with a posterior probability exceeding 90% and a ω value > 1.0 were designated as being 'positive selection sites' [23]. The site rate variation was evaluated comparing M_0 with M_3, while positive selection was evaluated comparing M_1 with M_2. The Bayes empirical Bayes approach implemented in M_{2a} and M_8 was used instead to determine the positively selected sites by calculating the posterior probabilities of ω classes for each site [25]. It is worth noting that PAML LRTs have been reported to be conservative for short sequences (*e.g.* positive selection could be underestimated), although the Bayesian prediction of sites under positive selection is largely unaffected by sequence length [25,26]. The d_N/d_S rate (ω) was also estimated by the ML approach implemented in the program HyPhy to enforce the previous analysis [27]. Two different algorithms estimated site-specific positive and negative selection: the fixed effect likelihood and random effect likelihood. The fixed effect likelihood fits a ω rate to every site and uses the likelihood ratio to test if $d_N = d_S$. The random effect likelihood is a variant of the Nielsen–Yang approach which assumes that a discrete distribution of rates exists across sites and allows both d_S and d_N to vary independently site by site. The three methods have been described in more detail elsewhere [28]. In order to select sites under selective pressure and keep our test conservative, a P value of ≤ 0.1 or a posterior probability of ≥ 0.9 as relaxed critical values was assumed.

For evolutionary analysis, the reference sequence Accession Number: NC_002549 was used to trace the exact position of the amino acids found under selection.

2.3. Amino acid mutation frequency analysis

Alignment between the Ebola protein reference sequences derived from NC_002549 (NP: P18272, GP: Q05320, VP24: