GMR
Genetics and Molecular Research
ISSN 16765680

Positive evolution of the glycoprotein (*GP*) gene is related to transmission of the Ebola virus

Y.X. Jing[1], L.N. Wang[2], X.M. Wu[2] and C.X. Song[1]

[1]Department of Computer Science, Qinghai Normal University, Qinghai, China
[2]The Key Laboratory of Biomedical Information Engineering of Ministry of Education, Xi'an Jiaotong University, Xi'an, China

Corresponding author: C.X. Song
E-mail: scx@qhnu.edu.cn

Genet. Mol. Res. 15 (1): gmr.15017173
Received September 14, 2015
Accepted November 5, 2015
Published March 28, 2016
DOI http://dx.doi.org/10.4238/gmr.15017173

ABSTRACT. Ebola hemorrhagic fever is a fatal disease caused by the negative-strand RNA of the Ebola virus. A high-intensity outbreak of this fever was reported in West Africa last year; however, there is currently no definitive treatment strategy available for this disease. In this study, we analyzed the molecular evolutionary history and attempted to determine the positive selection sites in the Ebola genes using multiple-genomic sequences of the various Ebola virus subtypes, in order to gain greater clarity into the evolution of the virus and its various subtypes. Only the glycoprotein (*GP*) gene was positively selected among the 8 Ebola genes, with the other genes remaining in the purification stage. The positive selection sites in the *GP* gene were identified by a random-site model; these sites were found to be located in the mucin-like region, which is associated with transmembrane protein binding. Additionally, different branches of the phylogenetic tree displayed different positive sites, which in turn was responsible for differences in the cell adhesion ability of the virus.

In conclusion, the pattern of positive sites in the *GP* gene is associated with the epidemiology and prevalence of Ebola in different areas.

Key words: Ebola disease; Evolutionary analysis; *GP* gene; Positive selection

INTRODUCTION

The year 2014 was witness to a severe outbreak of Ebola virus disease in various portions of West Africa, including Guinea, Liberia, Nigeria, Senegal, and Sierra Leone; this was the largest outbreak of Ebola in history (Dallatomasina et al., 2015). This disease has a high mortality rate of up to 90%, and is characterized by hemorrhagic fever and multiple organ failure (Goeijenbier et al., 2014). Ebola virus (EBOV) belongs to the genus *Filovirus*, which also includes the Cuevavirus and Marburg virus (Tseng and Chan, 2015). EBOV is divided into four subtypes: the Zaire (EBO-Z), Sudan (EBO-S), Reston (EBO-R), and Bundibugyo (EBO-B) viruses, each of which present different biological features and virulence (Ikegami et al., 2001).

The Ebola genome is a non-segmented negative-strain RNA that contains seven genes in the order *NP*, *VP35*, *VP40*, *GP*, *VP30*, *VP24*, and *L* (Ikegami et al., 2001; Dallatomasina et al., 2015). These genes encode eight proteins, with the *GP* gene encoding a 676-residue glycoprotein (GP), as well as a 364-residue secreted glycoprotein (Lee et al., 2008). EBOV GP is a type I transmembrane glycoprotein composed of two disulphide-linked subunits (GP1 and GP2) (Volchkov et al., 1998).

In this study, the evolutionary characteristics and differences in the Ebola genome were analyzed, with the objective of identifying the mechanism of disease transfer and geographic specifications of Ebola. The analysis methods used were based on the complete open-reading frames of GP.

MATERIAL AND METHODS

Data preparation

Genomic sequences of all EBOVs were obtained from the GenBank database (National Center for Biotechnology Information, NCBI; www.ncbi.nlm.nih.gov). The gene sequences were derived using the Perl script, through GenBank annotation. Stop codons in each sequence were excluded in all future analyses. A total of 53 *EBOV GP* genes were selected for this study, among which 31 belonged to the Zaire strain, and 10, 7, and 5 genes were obtained from the Sudan, Reston, and Bundibugyo strains of the virus.

Research methods

The nucleotide sequences of *GP* genes extracted from all EBOVs excluding EBOV-R contained 2028 sites (EBOV-R contains 2031 sites). All 53 sequences were aligned using the CLUSTALW program in the MEGA 6 platform (Kumar et al., 1994).

The nucleotide substitution rates and the most recent common ancestors (TMRCAs) of all sequences were estimated by the Bayesian method, using BEAST v1.8.1 (Drummond et al., 2012). The sequences were analyzed using an HKY model and an uncorrelated lognormal relaxed clock model with TipDates; all models were selected using the Modeltest program implemented in MEGA 6. The relative substitution rates of all three codon positions were also estimated. Non-informative priors were calculated using the MCMC algorithm; the sequences were run for 100 million generations, with the first 10 million being discarded as burn-in. TMRCA and effective sample sizes were determined using Tracer v.1.5 (http://evolve.zoo.ox.ac.uk). A consensus tree was created for each run, and the maximum clade credibility tree was selected from the posterior tree distribution using TreeAnnotator v.1.8.

Positive selection analyses

The selective pressures among sites were revealed by performing a maximum-likelihood analysis of the *EBOVGP* gene. A Bayesian tree was constructed with MrBayes v.3.1.1, using the GTR + Γ + I model; random site models were constructed using the codeml program, implemented in the PAML v.4.8 software platform (Yang, 2007), in order to access the different selective pressures.

The selective pressure was measured by the non-synonymous/synonymous rate ratio ω (dN/dS). Positive selection was detected when $\omega > 1$ (Yang et al., 2000) Comparisons between different random-site models were used to calculate the variations in ω (M0 *vs* M3) and to discover the presence of selected positive sites (M1 *vs* M2 and M7 *vs* M8). The gene tree of each subtype was subjected to all analyses, with the branch length being estimated by MrBayes. The models were compared using likelihood ratio tests (LRTs), with a chi-square distribution. Positive selective sites were identified by Bayes' Empirical Bayes analysis, performed on the PAML v.4.8 software platform.

The selective pressure exerted on different subtypes of the *EBOVGP* gene was tested by constructing branch site models. The ω values of these models varied along the different branches, classified into foreground and background branches. The sites in the genome sequences were classified into four types: 1) sites with identical ω values ($\omega_0 = 0$) that were conserved in all branches; 2) neutral sites with $\omega = \omega_1$, and with the remaining sites expressing ω_1 values of 1; 3) sites in background branches with $\omega_{2a} = \omega_0$, and $\omega_{2a} = \omega_{2b} \geq 1$ in the foreground branches; and 4) sites in background branches with $\omega_{2a} = \omega_1$, and $\omega_{2a} = \omega_{2b} \geq 1$ in the foreground branches (Yang and Nielsen, 2002). In this study, the branch site model was applied to all 53 strains. The different branch site models were then compared with the site-specific model M1 (neutral) (Zhang et al., 2005).

RESULTS

Evolutionary rate

The mean evolutionary rate of all strains, as estimated by BEAST v.1.8.1, was 6.884E-4 [95% highest probability density (HPD); 2.7137E-4-11.285E-4] substitutions per site per year. The relative substitution rates of the positions of all three codons were also estimated; the evolutionary rates at codon positions 1, 2, and 3 were 0.709 (95% HPD; 0.6534-0.7641), 0.49 (95% HPD; 0.4458-0.5359), and 1.801 (95% HPD; 1.73-1.87) substitutions per site per year.

The most recent common ancestor of the various subtypes existed 676.38 years ago (95% HPD; 252.40-1236.54); the most recent common ancestor of EBOV-R and EBOV-S and EBOV-Z and EBOV-B existed 429.8497 and 327.9389 years ago (Figure 1). The trees were summarized using TreeAnnotator, and were viewed using the FigTreev.1.4.2 program.

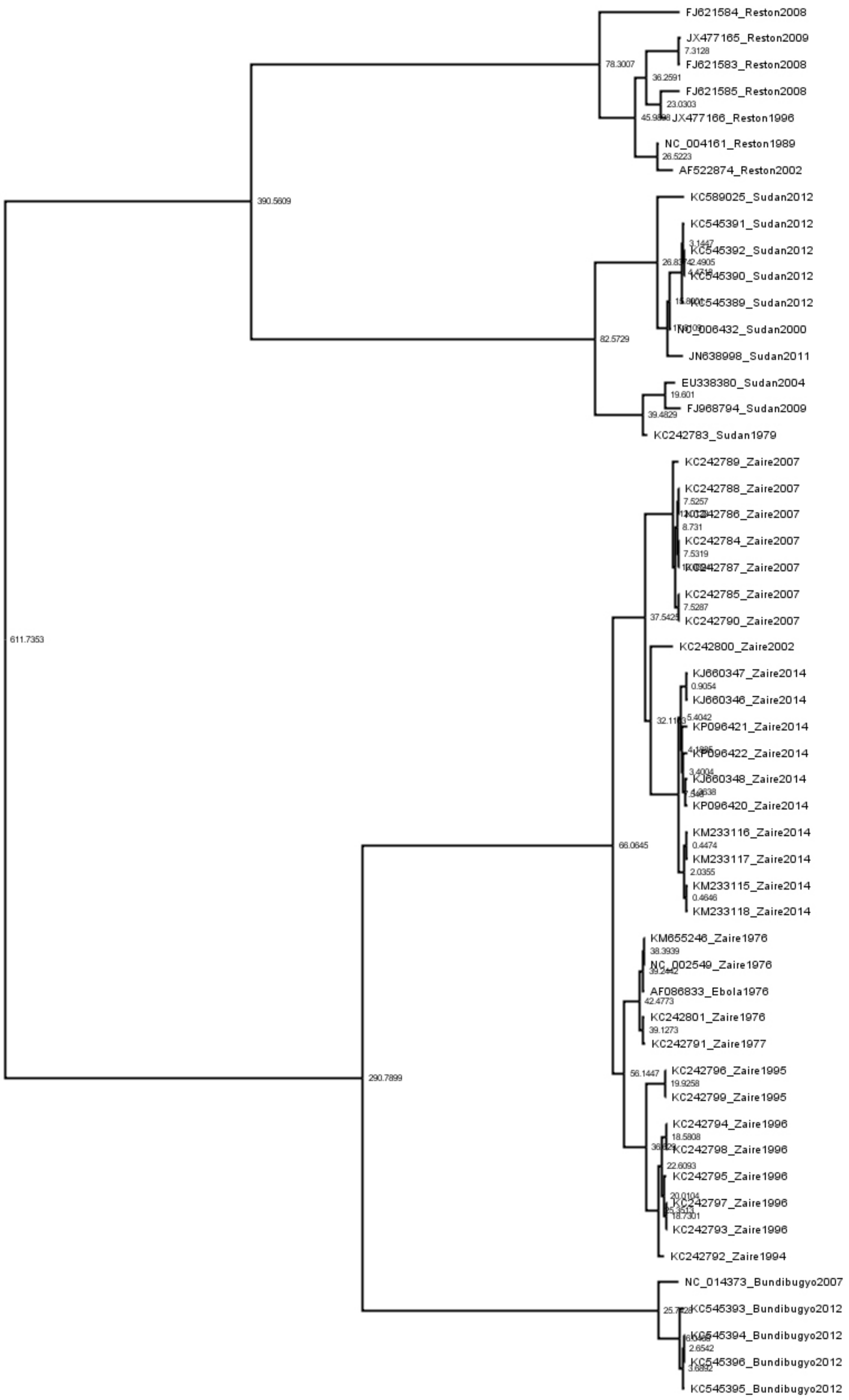

Figure 1. Evolution of various Ebola viruses based on the evolution of the glycoprotein (*GP*) genes.

Positive selected sites

The positive selected sites were analyzed using the codeml program of PAML v.4.8; the sites were identified using the random-site model. The models and LRTs are summarized in Table 1 and the sites are shown in Table 2. Eleven positive selection sites were identified in EBOV-R, and 6, 4, and 6 sites were observed in EBOV-Z, EBOV-S, and EBOV-B; the specific sites in the M8 (PAML) models were identified with a posterior probability >50%. The selection sites did not overlap in the different subtypes; however, these sites were concentrated in the 300-400-amino acid region, coding for mucin-like properties, indicating that this region may be positively selected.

Table 1. Radom-site models and likelihood ratio test (LRT) for all subtypes of the Ebola virus.

Model	np	lnL	κ	ω_0	ω_1	ω_2	null	LRT	d.f.	P
EBOV-Z										
M0	45	-3648.90	11.64	0.28						
M1	46	-3644.79	11.75	0.36 (74.79%)	1 (25.21%)		M0	8.22	1	0.004
M2	48	-3643.70	11.95	0.21 (98.18%)	1 (0%)	5.02 (1.82%)	M1	2.18	2	0.336
M3	49	-3643.70	11.95	0.19 (0%)	0.20 (98.18%)	5.02 (1.82%)	M0		4	0.034
M7	46	-3644.86	11.82	P = 0.005	q = 0.014					
M8	48	-3643.71	11.95	P0 = 0.982, P1 = 0.018, P = 26.26, w = 5.03, q = 98.77			M7	2.3	2	0.317
EBOV-B										
M0	7	-2948.33	30.27	0.33						
M1	8	-2948.10	30.71	0 (66.7%)	1 (33.3%)		M0	0.23	1	0.639
M2	10	-2947.80	34.77	0.29 (99.6%)	1 (0%)	43.83 (0.3%)	M1	1.06	2	0.588
M3	11	-2947.80	34.77	0.29 (94.2%)	0.29 (5%)	43.82 (0.3%)	M0	1.06	4	0.900
M7	8	-2948.12	30.62	P = 0.013	q = 0.025					
M8	10	-2947.80	34.76	P0 = 0.996, P1 = 0.004, P = 40.408, w = 43.83, q = 99.00			M7	0.64	2	0.726
EBOV-R										
M0	13	-3762.44	10.02	0.52						
M1	14	-3760.29	10.16	0.02 (47%)	1 (52%)		M0	4.3	1	0.038
M2	16	-3760.17	10.20	0.34 (89.4%)	1 (0%)	2.26 (20%)	M1	0.24	2	0.887
M3	17	-3760.17	10.20	0.35 (42.2%)	0.35 (47.2%)	2.25 (10.5%)	M0	4.54	4	0.338
M7	14	-3760.29	10.13	P = 0.027	q = 0.0226					
M8	16	-3760.27	10.21	P0 = 0.896, P1 = 0.104, P = 52.80, w = 2.27, q = 99.00			M7	0.34	2	0.843
EBOV-S										
M0	16	-3394.14	12.88	0.25						
M1	17	-3390.62	13.02	0 (74.39%)	1 (25.60%)		M0	7.44	1	0.0079
M2	19	-3390.42	13.15	0.13 (93.90%)	1 (0%)	2.44 (6%)	M1	0.4	2	0.8187
M3	20	-3390.42	13.15	0.13 (8.82%)	0.13 (85.17%)	2.44 (6%)	M0	7.44	4	0.1144
M7	17	-3390.72	12.97	p = 0.01635	q = 0.04870					
M8	19	-3390.42	13.15	P0 = 0.9404, P1 = 0.0596, P = 14.8371, w = 2.4481, q = 99.0			M7	0.6	2	0.7408

Table 2. Positive selected sites and their probabilities, identified by the random-site model.

Partition	Sites
EBOV-R	13E, 395E, 409A, 413D, 424Y, 426S, 434S, 462A, 504V (>50%); 229N (>60%); 430P (>70%)
EBOV-Z	544I, 430L (>80%); 443S, 377P, 331E (>70%); 455Y (>50%)
EBOV-S	374S, 403I, 503T (>70%); 432G (>80%)
EBOV-B	151F, 239P, 310L, 452Q, 489V (>50%); 367L (>70%)

The values in parentheses indicate the posterior probabilities of the respective sites.

Additionally, branch-site models were used to identify the positive selection pressure exerted on the various branches (Table 3). The significance of these results were determined by the LRT. Table 4 summarizes the various LRTs of the branch-site models; the branch-site models were compared with the site model M1 (neutral), with the conserved sites being denoted as $\omega = 0$ in both models (Zhang et al., 2005). The fit of the branch-site model was better than that of the site model M1 in EBOV-R and EBOV-S ($P < 0.0001$), which indicated the increased significance of the positive selection in EBOV-S and EBOV-R.

Table 3. Branch-site models for each partition.

Partition	Model	np	lnL	κ	Foreground	Background
	M1	85	-10947.17	4.79	ω_0 = 0.0399 (69.18%) ω_1 = 1.0000 (30.82%)	
EBOV-R	branch-site A	87	-10936.74	5.03	ω_0 = 0.03656 (67.03%) ω_1 = 1 (29.03%) ω_{2a} = 21.79 (2.75%) ω_{2b} = 21.79 (1.19%)	ω_0 = 0.03656 (67.03%) ω_1 = 1 (29.03%) ω_{2a} = 0.03656 (2.75%) ω_{2b} = 0.01190 (1.19%)
EBOV-Z	branch-site A	87	-10946.82	4.78	ω_0 = 0.0398 (68.93%) ω_1 = 1.0000 (30.00%) ω_{2a} = 4.80 (0.74%) ω_{2b} = 4.79 (0.32%)	ω_0 = 0.0398 (68.93%) ω_1 = 1.0000 (30.00%) ω_{2a} = (0.0398.74%) ω_{2b} = 1.0000 (0.32%)
EBOV-S	branch-site A	87	-10938.23	5.05	ω_0 = 0.3660 (66.93%) ω_1 = 1.0000 (29.47%) ω_{2a} = 35.12 (2.5%) ω_{2b} = 35.12 (1.1%)	ω_0 = 0.3660 (66.93%) ω_1 = 1.0000 (29.47%) ω_{2a} = 0.3660 (2.5%) ω_{2b} = 1.0000 (1.1%)
EBOV-B	branch-site A	87	-10946.52	4.83	ω_0 = 0.03935 (64.07%) ω_1 = 1 (28.19%) ω_{2a} = 1 (5.38%) ω_{2b} = 1 (2.37%)	ω_0 = 0.03935 (64.07%) ω_1 = 1 (28.19%) ω_{2a} = 0.03935 (5.38%) ω_{2b} = 1 (2.37%)

Partitions indicate the four subtypes (EBOV-Z, EBOV-R, EBOV-B, and EBOV-S).

Table 4. Likelihood ratio tests (LRTs) of branch-site models (BrS) compared to the random-site model M1.

Models	Partition	d.f.	LRT	P
BrS *vs* M1	EBOV-R	2	20.86	0.00002
BrS *vs* M1	EBOV-Z	2	0.7	0.7047
BrS *vs* M1	EBOV-S	2	17.88	0.0001
BrS *vs* M1	EBOV-B	2	1.3	0.5220

Statistical significance was indicated by $P < 0.05$ and $\chi^2 = 3.8$.

DISCUSSION

The evolutionary rate of the various codon positions revealed that these sequences underwent purifying selection, despite the third position of a codon being a wobble position [where the evolutionary rate is significantly (approximately 3 times) higher than those of other positions in a codon]. Additionally, the ω_0 of all subtypes was <1 in the M0 random-site model, which indicated the occurrence of a purifying selection.

However, the positive selection sites, detected using the random site model, were located in the mucin-like region (306-486 in EBOV-R, 305-485 in other subtypes), which is associated with a cytotoxic function in the EBOV (all subtypes); the functionality of the other sites remains unknown. The mucin-like region is also responsible for binding the virus to human C-type lectin domain family 10 member A (*CLEL10A*) gene, which encodes a member of the C-type lectin/C-type lectin-like domain (CTL/CTLD) superfamily. Members of the CTL/CTLD superfamily share a common protein fold and have diverse functions, such as cell-cell signaling, cell adhesion, and glycoprotein turnover. According to the annotation data, CTL/CTLD family members also play important roles in inflammation and immune response. Therefore, positive selection sites may play an important role in enhancing the viral binding and infection capacity. Additionally, these sites may also be responsible for immune response; these correlations must be defined and corroborated in future studies.

Conflicts of interest

The authors declare no conflict of interest.

ACKNOWLEDGMENTS

Research supported by the National Natural Science Foundation of China (#61450006), the Natural Science Foundation of Qinghai Province (#2014-ZJ-724), and the Fundamental Research Funds for the Central Universities (#2012jdhz47, #xjj2013117).

REFERENCES

Dallatomasina S, Crestani R, Sylvester Squire J, Declerk H, et al. (2015). Ebola outbreak in rural West Africa: epidemiology, clinical features and outcomes. *Trop. Med. Int. Health* 20: 448-454.http://dx.doi.org/10.1111/tmi.12454

Drummond AJ, Suchard MA, Xie D and Rambaut A (2012). Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol. Biol. Evol.* 29: 1969-1973.http://dx.doi.org/10.1093/molbev/mss075

Goeijenbier M, van Kampen JJ, Reusken CB, Koopmans MP, et al. (2014). Ebola virus disease: a review on epidemiology, symptoms, treatment and pathogenesis. *Neth. J. Med.* 72: 442-448.

Ikegami T, Calaor AB, Miranda ME, Niikura M, et al. (2001). Genome structure of Ebola virus subtype Reston: differences among Ebola subtypes. Brief report. *Arch. Virol.* 146: 2021-2027.http://dx.doi.org/10.1007/s007050170049

Kumar S, Tamura K and Nei M (1994). MEGA: Molecular Evolutionary Genetics Analysis software for microcomputers. *Comput. Appl. Biosci.* 10: 189-191.

Lee JE, Fusco ML, Hessell AJ, Oswald WB, et al. (2008). Structure of the Ebola virus glycoprotein bound to an antibody from a human survivor. *Nature* 454: 177-182.http://dx.doi.org/10.1038/nature07082

Tseng CP and Chan YJ (2015). Overview of Ebola virus disease in 2014. *J. Chin. Med. Assoc.* 78: 51-55.http://dx.doi.org/10.1016/j.jcma.2014.11.007

Volchkov VE, Feldmann H, Volchkova VA and Klenk HD (1998). Processing of the Ebola virus glycoprotein by the proprotein convertase furin. *Proc. Natl. Acad. Sci. USA* 95: 5762-5767.http://dx.doi.org/10.1073/pnas.95.10.5762

Yang Z (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* 24: 1586-1591.http://dx.doi.org/10.1093/molbev/msm088

Yang Z and Nielsen R (2002). Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol. Biol. Evol.* 19: 908-917.http://dx.doi.org/10.1093/oxfordjournals.molbev.a004148

Yang Z, Swanson WJ and Vacquier VD (2000). Maximum-likelihood analysis of molecular adaptation in abalone sperm lysin reveals variable selective pressures among lineages and sites. *Mol. Biol. Evol.* 17: 1446-1455.http://dx.doi.org/10.1093/oxfordjournals.molbev.a026245

Zhang J, Nielsen R and Yang Z (2005). Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol. Biol. Evol.* 22: 2472-2479.http://dx.doi.org/10.1093/molbev/msi237