Biochemical and Biophysical Research Communications 463 (2015) 756–761

Contents lists available at ScienceDirect

Biochemical and Biophysical Research Communications

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

The crux and crust of ebolavirus: Analysis of genome sequences and glycoprotein gene

Kiran Narasinha Mahale*, Milind S. Patole*

National Centre for Cell Science, SP Pune University Campus, Pune, 411007, India

ARTICLE INFO

Article history:
Received 19 May 2015
Accepted 2 June 2015
Available online 4 June 2015

Keywords:
Genome phylogeny
GP-gene
RNA polymerase slippage
Glycosylation

ABSTRACT

The recent 2013–15 epidemic of Ebola virus disease (EVD) has initiated extensive sequencing and analysis of ebolavirus genomes. All ebolavirus genomes available until December 2014 have been collated and analyzed in this study to obtain phylogenetic relationship and uncover the variations amongst them. The terminal 'leader' and 'trailer' nucleotide sequences of the genomes were omitted and analysis of the intermediate region accommodating the sole seven genes (hepta-CDS region) of the virus showed relative stability of the genome, including the ones isolated from the current epidemic. The genome information was scrutinized to detect the variation in the surface glycoprotein gene and annotate its three protein products, resulting from its atypical transcription. This study will make an easy understanding of the genomes for those who desire to exploit the genome sequences for different investigations in EVD.

1. Introduction

Ebola virus disease (EVD) is a severe haemorrhagic fever caused by ebolavirus (genus *Ebolavirus*, family *Filoviridae*) with a very high case-fatality rate ranging from 25 to 90% [1]. EVD was first identified in 1976 with few cases restricted to sub-Saharan Africa followed by sporadic reports in the ensuing years. However, in 2014, WHO confirmed the worst EVD outbreak causing 8004 deaths globally, including 7989 deaths in the west African countries, namely Liberia, Guinea, and Sierra Leone (as on December 31, 2014; http://www.cdc.gov/vhf/ebola/outbreaks/2014-west-africa/case-counts.html). The genus *Ebolavirus* has five members namely, Taï Forest virus (TAFV; formerly Côte d'Ivoire), Reston virus (RESTV), Sudan virus (SUDV), Bundibugyo virus (BDBV) and Ebola virus (EBOV; member of *Zaire ebolavirus* species) [2]. RESTV is infective only in non-human primates; and causes seroconversion in humans exposed to the virus with no association of illness, indicating its avirulence in humans [3]. On the contrary, *EBOV* has caused at least 12 major outbreaks since 1976 including the recent epidemic of 2013–15 and in the light of the rapid accumulation of genomic data, intelligible investigations of the sequence data are crucial.

The ebolavirus genome is a linear, negative-sense, single-stranded, ~19 kb RNA containing seven protein-coding genes. The terminal ends of the genome have short non-transcribed 3′ 'leader' (*Ldr*) and 5′ 'trailer' (*Tlr*) sequences that contain the signals for replication, transcription and encapsidation [4]. One of the viral proteins responsible for its virulence is its surface glycoprotein (*GP*-gene product) present in the envelope that aids viral entry into the host cell [5] and is shown to induce production of the virus-neutralizing antibody [6]. Anti-glycoprotein antibody cocktail was found to be protective in laboratory animals infected with EBOV [7] and without the safety and efficacy testing of this cocktail in humans, clinicians used this cocktail to treat a small number of EVD patients during the 2013–15 epidemic. Therefore, a comparative analysis of the different ebolavirus glycoprotein gene and its protein products is crucial for the investigation of its antigenic efficacy. The *GP*-gene has been previously used by many investigators for sequence comparison and phylogeny [8,9]. The *GP*-gene is also remarkable for its polyadenosine transcription slippage (stuttering) site leading to the coding of three proteins from three different frames [5]. Some ebolavirus genomes deposited into NCBI lack correct annotations for the *GP*-gene.

Here, we have compared all complete genome sequences of ebolavirus available until 2014 to demonstrate the overall variation in the genome. Additionally, the *GP*-gene and its three protein products have also been investigated for their sequence variation and glycosylation potential.

* Corresponding authors.
E-mail addresses: kkmahale@gmail.com (K.N. Mahale), patole@nccs.res.in (M.S. Patole).

http://dx.doi.org/10.1016/j.bbrc.2015.06.008
0006-291X/© 2015 Elsevier Inc. All rights reserved.

2. Material and methods

2.1. Genome datasets

A total of 169 complete genome sequences of ebolaviruses from different geographical locations and outbreaks were downloaded from the NCBI database (http://www.ncbi.nlm.nih.gov/nucleotide). Each genome has been assigned a 10 character code for clarity and uniformity (Supplementary Table 1).

2.2. Phylogenetic analysis

The terminal '*Ldr*' and '*Tlr*' sequences from ebolavirus genomes were excised and the intermediate region containing seven genes (hence forth referred as the hepta-CDS region) were aligned using MAFFT v7.186, a multiple sequence alignment tool. This was further subjected to 1000 randomly assorted bootstrap replicates and a consensus tree was built by the *neighbor-joining method (Phylip v3.695)*. A single nucleotide polymorphism (SNP)-based phylogeny was executed by manual alignment of the hepta-CDS region by placing the sequences one above the other and gaps were introduced at the appropriate locations to match their lengths. Further analysis was done by 'dnadist.exe' and 'neighbor.exe' programs of Phylip v3.695. The *GP*-gene nucleotide sequence was also subjected to a similar manual alignment and analysis.

2.3. Glycosylation prediction

Motifs for N-glycosylation and W-mannosylation were scanned in the three proteins of *GP*-gene by searching for **N{P}[ST]{P}** [10] and **WXXW** [11] motifs respectively.

3. Results

3.1. Genome-based phylogeny of hepta-CDS region

As it is difficult to obtain the exact DNA sequence of the terminal ends of the ebolavirus genome [12], they show greater variations in the length of their terminal ends. Therefore, we have excluded the terminal non-coding '*Ldr*' and '*Tlr*' sequences in the present analysis, which accounts for 6.38% of the genome. This study utilized the intermediate hepta-CDS region (~17.7 kb), which includes sequence from the start codon of the first gene (*NP* gene) till the stop codon of the rearmost 7th gene (*L* gene).

An initial alignment of ebolavirus genomes showed that many genomes had identical hepta-CDS sequences (Supplementary Tables 1A and 1B). For instance, two isolates from the 1996 outbreak in the Gabonese Republic (Gabon), 'z96_G1Ikot' (NCBI: KC242798) and 'z96_Gb2Nza' (NCBI: KC242794), had identical hepta-CDS sequences. Whereas, **'zMR_EM106x'** (NCBI: KM233036), an isolate of the Makona variant from the 2013–15 epidemic was identical to 26 other reported viruses from the same epidemic [13]. The genomes of the first ebolavirus isolate, **'z6_ymMayg'** (NCBI: NC_002549) and 'zMR_EM106x' (representative of Makona variant) are used as reference strains for comparisons in this analysis. Such identical sequences may arise from close neighbors in a transmission chain and are meaningful in terms of temporal context in epidemiology. Inclusion of a large number of identical sequences in phylogenetic construction often leads to spurious results. Therefore, we have taken only 94 unique hepta-CDS sequences out of 169 ebolavirus genomes for phylogenetic analysis. The phylogenetic tree constructed using the hepta-CDS region for different ebolaviruses (Supplementary Fig. 1) was very similar to that observed in previous reports [14,15]. The tree clearly indicated that various ebolavirus members have split during evolution and isolated genetically from one another evolving separately. Lineages like TAFV or RESTV have split because of the variance effect as these viruses were isolated from distinct regions like Ivory Coast or macaques imported from Philippines. The EBOV isolates from the recent epidemic of 2013–15 from west Africa (Makona variant) clustered in a well defined clade, barring the four genomes of Lomela variant isolated from the Democratic Republic of the Congo (DR Congo) [16]. The members of the Lomela variant fell in the clade between the isolates from Gabon (1994) and Kikwit, DR Congo (1995).

Since the genomes of different ebolavirus are divergent and have variable lengths, automated sequence alignment tools tend to add many gaps making it difficult for analysis. Also, the close identity of the genome sequences reported from a single epidemic (Supplementary Fig. 1) leads to a weaker bootstrap value. This prompted us to construct a SNP-based phylogenetic tree by manually aligning the genomes. The genomes of SUDV, EBOV, BDBV and RESTV were separately analyzed [15] for SNPs and the results are as shown in Supplementary Fig. 2 A–E.

3.1.1. Reston virus (RESTV)

Amongst the eight genomes of RESTV (Supplementary Fig. 2D), hepta-CDS had a variable length (17721–17728 bases), as against the constant size of hepta-CDS region in other ebolaviruses. Even amongst the RESTV members, the genomes of Pennsylvania strain (r89_mPenns, NCBI: NC_004161; and r04_pennGF, NCBI: AY769362) isolated from a single location in 1989 showed a variation of 41 SNPs and an insertion of three nucleotides in 'r04_pennGF' (NCBI: AY769362). Viral isolates from the same swine farm A collected in 2008 and 2009 showed a variation of only 14 SNPs. Whereas, the C and E isolates of RESTV collected from the same farm CE in the same year (2008) displayed a variation with more than 200 SNPs. Interestingly, the RESTV isolate E (r08_swnSpE, NCBI: FJ621585) was found closer to the isolate from infected monkeys in Alice, Texas, 1996 (r96_mPHLtx, NCBI: JX477166). The hepta-CDS based SNP-phylogenetic tree showed that it is difficult to ascertain the genetic diversity of RESTV to any particular niche.

3.1.2. Sudan virus (SUDV)

Four SUDV isolates from **Uganda** (Supplementary Fig. 2C; representative s12_ebo602; NCBI: KC545389) isolated from Kibaale (western Uganda) in 2012, had a difference of only 1–2 nucleotides amongst them and had close identity with isolates previously obtained from Gulu (northern Uganda) in 2000 (125 SNPs compared to s00_gulu92, NCBI: NC_006432) and Nakisimata (Luwero district, central Uganda) in 2011 (135 SNPs compared to s11_nakism, NCBI: JN638998). This analysis confirms the ability of SUDV to preserve their genetic makeup. However, another viral genome 's12_ebo639' (NCBI: KC589025) isolated in the same year, 2012, from Luwero showed large variations in the genome sequences as compared to the aforementioned four isolates (176 SNPs with respect to s12_ebo602, NCBI: KC545389).

The SUDV isolates from **Sudan** (s79_hMaleo, NCBI: KC242783; s08_bonifc, NCBI: FJ968794; s04_yambio, NCBI: EU338380) form a well-separated clade from the Ugandan SUDV isolates. Inclusion of five isolates from the 2012 outbreak reconfirms the spatial clustering of the SUDV. However, the two clades showed great variation in the number of their SNPs as compared to other members of ebolavirus. For example, 's08_bonifc' (NCBI: FJ968794; isolated from Sudan, 1976; sequenced in 2008) shows a variation of 863 SNPs when compared to 's00_gulu92' (NCBI: NC_006432).

3.1.3. Bundibugyo virus (BDBV)

A minimal amount of variation is seen amongst the genomes of four BDBV isolates obtained from the 2012 outbreak in Isiro, DR

Congo [17]. However, they show a variation with ~225 SNPs and a deletion of one cytosine residue with respect to the genome isolated from the first BDBV isolate (b07_butlya, NCBI: NC_014373) from Bundibugyo (in 2007; western Uganda bordering DR Congo; distance of 590 km s from Isiro; Supplementary Fig. 2B) [18]. These results are also quite similar to the ones obtained from SUDV analysis showing the spatial distribution of virus isolates.

3.1.4. Ebola virus (EBOV)

EBOV **Yambuku** variant caused the first outbreak in 1976 in northern DR Congo and its three isolates (Mayinga, deRoover and Ecran) showed very little variation of 1–3 SNPs amongst themselves. Another genome of Mayinga isolate (z02_mayigZ, NCBI: AY142960) showed only a single SNP change with respect to ‘z76_ymMayg’ (NCBI: NC_002549). An isolate from Bonduni, DR Congo (1977) also showed a variation of only 2 SNPs when compared to ‘z76_ymMayg’. However, two guinea pig (z00_gpigMy, NCBI: AF272001; z07_gpigMy, NCBI: EU224440) and one mouse (z02_maygMS, NCBI: AF499101) passaged strains of EBOV showed a slightly higher variation (9–21 SNPs) with respect to ‘z76_ymMayg’. Although this variation altered the virulence phenotype [19], the passaged strains share the same cluster as the Yambuku variant.

The EBOV genome from **Kikwit** (southern DR Congo) outbreak in 1995 showed a variation of 196 SNPs from the Yambuku variant, 1976. Similarly, six EBOV genomes reported from 1994 to 1996 outbreak in **Gabon** show a variation of ~250 SNPs with respect to ‘z76_ymMayg’ and 162 SNPs as compared to ‘z95_kik709’ (NCBI: KC242799; Kikwit, 1995). The viral hepta-CDS sequences from earlier outbreaks in Yambuku, Kikwit, and Gabon had very low genetic diversity amongst themselves, and on the basis of their sequence resemblance, these viral genomes form well defined clusters (Supplementary Fig. 2A). Interestingly, four isolates of **Lomela** variant from Boende (midway between Yambuku and Kikwit) in 2014 also clustered along with the mid-1990 isolates from Kikwit and Gabon (125 SNPs compared to z95_kik709, NCBI: KC242799; and 157 SNPs as against z94_hGabon, NCBI: KC242792).

Genome sequences of nine EBOV isolates from the 2007 outbreak of **Luebo** (DR Congo) were also found to be form a well defined cluster as a result of very high homology amongst them. A maximum divergence of .000958 (calculated by dnadist.exe, F84 method, Phylip package) [20] was seen between the seven Luebo isolates with corresponding 17 SNPs between the most distant isolates (z07_v034KS, NCBI: HQ613402; and z07_Lubo43, NCBI: KC242788). Nonetheless, viruses isolated from the southern cities of Luebo and Kikwit, which are at a road distance of 600 km, do not fall into the same clade.

The set of 113 genomes of **Makona** variant from the recent epidemic of 2013–15 in western Africa form a very close cluster and a separate clade with a common ancestor. Interestingly, this ancestor is also shared by the viral isolates from 2007 Luebo outbreak. Maximum divergence seen amongst the members of the Makona variant of 2014 is .001353 with corresponding 24 SNPs observed between the most distant isolates (z14_WPGc05, NCBI: KP096420; and z14_MLImm4, NCBI: KP260802). Whereas, the combined set of genomes isolated from 43 Makona and 7 Luebo variants showed a maximum divergence of .028525, with a corresponding 483 SNPs between ‘z14_MLImm4’ (NCBI: KP260802) and ‘z07_Lubo43’ (NCBI: KC242788). Although the Makona variant members form a very tight cluster (Supplementary Fig. 2A), isolates of 2014 from Guinea, Mali, and Sierra Leone, when analyzed separately (Supplementary Fig. 2E) showed cluster formation among the different members isolated from these countries. Malian genome sequences also formed a separate clade (with 7–12 SNPs as against zMR_EM106x) and when compared with a single Liberian isolate (z14_LBRm07, NCBI: KP178538; 3 SNPs compared to zMR_EM106x) showed different spectra of SNPs. The first three isolates of Makona variant reported from Guinea [21] and their in vitro passaged viruses in Vero E6 cell line [12] form a separate clade within the viruses isolated from Sierra Leone [13]. However, a single genome reported from Guinea (z14_GIN192, NCBI: KP342330) does not club with the aforementioned Guinea isolates and in fact, has only a single SNP variation from ‘zMR_EM106x’ (c1380t). This exceptional variation may be accounted for the movement of the related patient from Sierra Leone to neighboring Guinea. The genome of a single isolate reported from Great Britain (patient infected in west Africa; z14_GBRm1x, NCBI: KP120616) and its passaged virus (having an identical hepta-CDS sequence) were also used in the phylogenetic analysis and has a variation of only 3 SNPs with respect to ‘zMR_EM106x’. It is worth mentioning that the Lomela variants show a greater variation of ~550 SNPs from the Makona variant isolated in the same epidemic 2013–15 from

Table 1
Numerical summary of the ebolavirus genomes, *GP*-gene CDS and its three protein products.

Member of ebolavirus	Total no. of genome sequences	No. of unique hepta-CDS representatives	No. of unique *GP*-gene CDS representatives	No. of unique preGP protein representatives	No. of unique sGP protein representatives	No. of unique ssGP protein representatives
EBOV from Luebo (DR Congo), 2007 (Carroll et al., 2013)	9	9	3	1	3	3
EBOV (Makona variant) from Guinea, 2014: 3 unpassaged samples: Baize et al., 2014; 3 tissue culture samples: Hoenen et al., 2014	6	6	5	5		
EBOV (Makona variant) from Sierra Leone, 2014: 99 unpassaged samples (Gire et al., 2014); Other EBOV (Makona variant), 2014: (Mali: 4; Guinea: 1; Liberia: 1; Great Britain: 2)	99 + 8	37	4	2	2	2
EBOV (Lomela variant) from DR Congo, 2014 (Maganga et al., 2014)	4	2	2	2	1	1
EBOV genomes from other outbreaks	19	16	10	10	6	6
Total Ebola virus (EBOV) genomes	**145**	**70**	**24**	**20**	**12**	**11**
Taï Forest virus (TAFV) genomes	**1**	**1**	**1**	**1**	**1**	**1**
Bundibugyo virus (BDBV) genomes	**5**	**5**	**3**	**3**	**3**	**3**
Sudan virus (SUDV) genomes	**10**	**10**	**7**	**7**	**4**	**4**
Reston virus (RESTV) genomes	**8**	**8**	**8**	**7**	**7**	**7**
Total ebolavirus genomes	**169**	**94**	**43**	**38**	**27**	**26**

The bold rows show the number of isolates of each species of virus. The un-bold rows give the numerical distribution of isolates from the epidemics (including 2013–15) related to the first species (EBOV) described in the table.

Table 2
Annotation of the CDS of the *GP*-gene of ebolavirus members. The details for the three frames of translation along with their respective protein products are shown. Numbering of the nucleotides is initiated from the first nucleotide of the *GP*-gene start codon.

GP-gene CDS of ebolavirus	preGP (membrane-bound form of GP)			sGP (soluble form of GP)			ssGP (small soluble form of GP)		
	Reading frame	Slippage (if any)	preGP protein (length)	Reading frame	Slippage (if any)	sGP protein (length)	Reading frame	Slippage (if any)	ssGP protein (length)
EBOV (typical *GP*-gene CDS)	Frame −1	1..885, 885..2030	295 + 381 = 676	Frame +1	1..1095	364	Frame +2	1..885, 887..895	295 + 002 = 297
EBOV (z95_kFotDt, z95_kikwit)	Frame +1	1..2031	676	Frame +2	1..0885, 887..1096	295 + 069 = 364	Frame +3	1..885, 888..896	295 + 002 = 297
TAFV (typical *GP*-gene CDS)	Frame −1	1..885, 885..2030	295 + 381 = 676	Frame +1	1..1098	365	Frame +2	1..885, 887..910	295 + 007 = 302
BDBV (b12_ebo122, b12_ebo112)	Frame −1	1..885, 885..2030	295 + 381 = 676	Frame +1	1..1101	366	Frame +2	1..885, 887..910	295 + 007 = 302
BDBV (b07_butlya)	Frame −1	1..885, 885..2030	295 + 381 = 676	Frame +1	1..1122	373	Frame +2	1..885, 887..910	295 + 007 = 302
SUDV (typical *GP*-gene CDS)	Frame −1	1..885, 885..2030	295 + 381 = 676	Frame +1	1..1119	372	Frame +2	1..885, 887..958	295 + 023 = 318
RESTV (typical *GP*-gene CDS)	Frame −1	1..885, 885..2033	295 + 382 = 677	Frame +1	1..1104	367	Frame +2	1..885, 887..997	295 + 036 = 331
RESTV (r08_swnLuA, r09_swinFA)	Frame −1	1..885, 885..2033	295 + 382 = 677	Frame +1	1..1146	381	Frame +2	1..885, 887..1189	295 + 100 = 395

western Africa. This indicates that each ebolavirus has been evolving independently in each geographical location and there is minimal genetic exchange between them.

3.2. Comparative analysis of GP-gene sequence

Although there are 94 unique hepta-CDS sequences available from various isolates of ebolavirus, many of them have exact identity with respect to their *GP*-gene CDS sequence (Table 1). A total of 43 *GP*-gene homologs can be seen in the dataset of 169 genomes. Interestingly, all eight Reston virus genomes showed different *GP*-gene sequence, while highly pathogenic EBOV with 145 genome sequences in database show only 24 unique CDS for *GP*-gene (Table 1).

As *GP*-gene encodes for three different proteins, DNA sequence was chosen instead of protein sequences for the derivation of SNP-based comparative analysis. *GP*-gene phylogeny showed that each member was distinctly clustered indicating that each ebolavirus harbors specific SNPs (Supplementary Fig. 3 A–B). Within each ebolavirus member, there is very little variation at DNA sequence level which, at times, may not get reflected at the amino acid level in GP proteins. However, the *GP*-gene of recently isolated Makona (zMR_EM106x) and Lomela (z14_CODBLk, NCBI: KM519951; z14_CODL19, NCBI: KP271020) variants show a variation of 67–69 SNPs between them.

RESTV had an atypical *GP*-gene CDS length of 2033 bases, as compared to 2030 bases of all other ebolaviruses. The length of Reston preGP, sGP and ssGP were 677, 367 and 331 amino acids respectively as compared to EBOV GP-proteins which had lengths of 676, 364 and 297 amino acids (Table 2). Hence the *GP*-gene of RESTV was analyzed separately (Supplementary Fig. 3B) and this was essentially very similar to the SNP-based phylogenetic tree of hepta-CDS region of RESTV genomes.

3.3. Transcriptional slippage of GP-gene

The ebolavirus *GP*-gene contains a polyadenosine stretch, which is susceptible to transcriptional slippage. Unedited transcripts (seven adenosine residues at the editing site) of the *GP*-gene, encode a soluble form of the glycoprotein (sGP, 364 amino acids in EBOV). The unedited transcript (+1 Frame) encodes sGP protein containing 364 amino acids in EBOV, 365 amino acids in TAFV, 366 or 373 amino acids in BDBV and 372 amino acids in SUDV. In EBOV, the stop codon for sGP is either ‘taa’ or ‘tga’. However, the 1,093rd residue (1st residue of the stop codon) is mutated (t1093c in TAFV and BDBV; t1093a in SUDV) resulting in the variable length of sGP protein (Supplementary Fig. 4 and Supplementary Table 2).

When the slippage occurs in the +2 frame, the 886th adenosine residue is not transcribed (1..885, 887..895) encountering a stop codon prematurely, leading to an early termination of transcription. This mRNA translates to a smaller ssGP protein with a highly variable length of 297 (EBOV), 302 (TAFV and BDBV), 318 (SUDV) and 331–395 (RESTV) amino acids (Table 2). Alterations in the stop codon position leading to variable lengths of sGP and ssGP proteins has been very well depicted for the non-RESTV members in Supplementary Fig. 4.

When the slippage occurs in −1 frame, with the 885th (adenosine) residue transcribed twice (1..885, 885..2030), a 2031 bases mRNA is formed, which encodes for 676 (= 295 + 381) amino acid long protein (preGP or $GP_{1,2}$; or virion spike protein). This membrane form of GP has a constant length of 676 amino acids in all non-RESTV members (RESTV preGP: 677 amino acids). Each monomer of preGP protein contains two subunits, GP_1 and GP_2. The GP_1 subunit contains the core of the glycoprotein, its receptor binding domain, a glycan cap, and a large mucin-like domain [5,22].

One of the exceptions to the *GP*-gene annotation are the sequences reported for the extensively passaged isolates, 'z95_kFotDt' (NCBI: AY354458) and 'z95_kikwit' (NCBI: JQ352763). The annotation of these genomes suggest that the unedited transcript (+1 frame) leads to the formation of preGP protein; while the +2 and +3 frames code for the sGP and ssGP proteins respectively (Table 2).

Although there are 43 homologs for *GP*-gene at DNA sequence level, the resulting protein homologs are 38 for preGP ($GP_{1,2}$), 27 for sGP and 26 for ssGP proteins (Supplementary Table 2). In the case of Luebo 2007 isolates, the three *GP*-gene representatives encode identical GP proteins. A similar case was observed in the isolates of RESTV, Reston08-A isolated in two consecutive years (2008 and 2009), where the *GP*-gene CDS showed a single SNP change (c501a) without any change at the protein level.

As the *GP*-gene encodes three different proteins (with the first 295 amino acids being identical), the two variants (Luebo of 2007 and Makona of 2013–15) show differences in the longer preGP and sGP proteins, but may have an identical ssGP protein (Supplementary Table 2). In Makona variants, the most common amino acid substitution is T262A (with respect to z76_ymMayg; also shared by the Luebo 2007 variants). There is an additional set of 17 amino acid substitutions among the 296–676 amino acids of preGP protein of Makona variants with respect to that of 'z76_ymMayg' (Supplementary Fig. 2E).

3.4. Glycosylation of GP-gene products

The translation of *GP*-gene mRNAs occurs on polysomes attached to endoplasmic reticulum and the protein is processed in Golgi bodies, where it undergoes glycosylation [23]. Minor changes in glycosylation pattern may pose challenges to the antigen-neutralizing antibodies and needs to be viewed seriously.

The number of N-glycosylation sites in EBOV are 17 (exception: z02_Ilembe, NCBI: KC242800; 15 N-glycosylation sites), 12 in SUDV, 11 in BDBV, 10 in TAFV, and 15–17 in RESTV with significant variation specific to each ebolavirus (Supplementary Table 3). The preGP protein of the EBOV Makona variant shows a significant variation from NTTT to NTTN (positions 333–336) and NHSE to NYSE (positions 454–457) as compared to other EBOV members. Another significant observation is the absence of the W-mannosylation site WAFW (at position 288–291) in the 'z14_GNgC05' (NCBI: KJ660348) isolate (and its passaged isolate, z14_WPGc05: NCBI: KP096420) of Makona variant from Guinea, which is in fact present in all members of *Ebolavirus* genus.

The Lomela variant members show a specific amino acid substitution of NDST to NAST at 413–416 positioned glycosylation site. This change is also shared by the isolates of 1996 EBOV Gabon outbreak ('z96_G1Ikot', NCBI: KC242798; and 'z96_Gb2Nza', NCBI: KC242794; having identical hepta-CDS sequences).

The predicted N-glycosylation sites in mucin domain (N228, N238, N257 and N268) were found to be well conserved in different EBOV members. However, in SUDV and TAFV, the predicted N-glycosylation site N228 is absent and is compensated by a commensurate site N208.

4. Discussion

Absence of proofreading system during RNA virus replication leads to higher mutation rates, but genome sequence comparisons in this study show very few SNPs amongst different variants, implying the stable genome of ebolavirus [9,24]. This relative stability suggests that therapeutic interventions based on protein sequences like vaccines or monoclonal antibodies should retain functionality over a long period of time and during forthcoming epidemics, if any.

This study includes genome sequences reported for isolates from most outbreaks and geographical regions. The phylogenetic analysis performed using automated alignment and SNP-based manual alignment showed that hepta-CDS based analysis yields proper linkage between different isolates showing a spatial variation among the different members of ebolavirus. Nonetheless, the EBOV isolates fall correctly into separate clades, not only with respect to their geographical location, but also with respect to their temporal origin of outbreak. The variants from each outbreak, namely Yambuku (1976), Kikwit (1995), Gabon (1996), Luebo (2007) and the latest Makona (2013–15) variants, group into distinct clades. An exception to this is the Lomela variant (2013–15, DR Congo) which clusters with Kikwit (1995) and Gabon (1996) variants. The recent Liberian isolates [25] were not included in this study, but the preliminary analysis of their hepta-CDS region shows only 5–11 SNPs when compared to 'zMR_EM106x' (Supplementary Fig. 2A). Regardless of genome stability, a simple comparison of the 1976 Yambuku variant with that of 2013–15 Makona variant shows that there is a change of ~500 SNPs, indicating accumulation of mutations in EBOV. Surprisingly, Makona variant shows an equivalent level of variation (~475 SNPs) with its closest relative, the Luebo isolates of 2007. However, the Lomela variant (2013–15) shows a drastically low variation with ~200 SNPs with respect to Yambuku variant. These unusual differences between the Yambuku, Luebo, Makona and Lomela variants from different timeframes hinder any molecular clock investigation.

A recent study [26] showing the presence of virus in the ocular fluid from an EVD recovered patient indicated the viral existence in unanticipated tissues of the human and animal body. Therefore, EVD demands a constant surveillance by isolation or molecular detection of ebolavirus from different tissues of humans and animals residing in geographical niches that are hotspots for EVD.

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.bbrc.2015.06.008.

Transparency document

Transparency document related to this article can be found online at http://dx.doi.org/10.1016/j.bbrc.2015.06.008.

References

[1] R.B. Martines, D.L. Ng, P.W. Greer, et al., Tissue and cellular tropism, pathology and pathogenesis of Ebola and Marburg viruses, J. Pathol. 235 (2015) 153–174.

[2] J.H. Kuhn, S. Becker, H. Ebihara, et al., Proposal for a revised taxonomy of the family *Filoviridae*: classification, names of taxa and viruses, and virus abbreviations, Arch. Virol. 155 (2010) 2083–2103.

[3] G.A. Marsh, J. Haining, R. Robinson, et al., Ebola Reston virus infection of pigs: clinical significance and transmission potential, J. Infect. Dis. 204 (2011) 804–809.

[4] E. Mühlberger, M. Weik, V.E. Volchkov, et al., Comparison of the transcription and replication strategies of Marburg virus and Ebola virus by using artificial replication systems, J. Virol. 73 (1999) 2333–2342.

[5] J.D. Cook, J.E. Lee, The secret life of viral entry glycoproteins: moonlighting in immune evasion, PLoS Pathog. 9 (2013) e1003258.

[6] S.T. Agnandji, A. Huttner, M.E. Zinser, et al., Phase 1 trials of rVSV Ebola vaccine in Africa and Europe – preliminary report, N. Engl. J. Med. (2015), http://dx.doi.org/10.1056/NEJMoa1502924.

[7] X. Qiu, G. Wong, J. Audet, et al., Reversion of advanced Ebola virus disease in nonhuman primates with ZMapp, Nature 514 (2014) 47–53.

[8] A. Sanchez, S.G. Trappier, U. Ströher, et al., Variation in the glycoprotein and VP35 genes of Marburg virus strains, Virology 240 (1998) 138–146.

[9] G. Dudas, A. Rambaut, Phylogenetic analysis of Guinea 2014 EBOV Ebolavirus outbreak, PLoS Curr. 6 (2014) 1229–1235.

[10] R. Apweiler, H. Hermjakob, N. Sharon, On the frequency of protein glycosylation, as deduced from analysis of the SWISS-PROT database, Biochim. Biophys. Acta 1473 (1999) 4–8.

[11] J. Krieg, S. Hartmann, A. Vicentini, et al., Recognition signal for C-mannosylation of Trp-7 in RNase 2 consists of sequence Trp-x-x-Trp, Mol. Biol. Cell. 9 (1998) 301–309.

[12] T. Hoenen, A. Groseth, F. Feldmann, et al., Complete genome sequences of three ebola virus isolates from the 2014 outbreak in west Africa, Genome Announc.. 2 (2014) e01331–14.

[13] S.K. Gire, A. Goba, K.G. Andersen, et al., Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak, Science 345 (2014) 1369–1372.

[14] R.W. Barrette, L. Xu, J.M. Rowland, M.T. McIntosh, Current perspectives on the phylogeny of Filoviridae, Infect. Genet. Evol. 11 (2011) 1514–1519.

[15] S.A. Carroll, J.S. Towner, T.K. Sealy, et al., Molecular evolution of viruses of the family *Filoviridae* based on 97 whole-genome sequences, J. Virol. 87 (2013) 2608–2616.

[16] G.D. Maganga, J. Kapetshi, N. Berthet, et al., Ebola virus disease in the Democratic Republic of Congo, N. Engl. J. Med. 371 (2014) 2083–2091.

[17] C.G. Albariño, T. Shoemaker, M.L. Khristova, et al., Genomic analysis of filoviruses associated with four viral hemorrhagic fever outbreaks in Uganda and the Democratic Republic of the Congo in 2012, Virology 442 (2013) 97–100.

[18] J.S. Towner, T.K. Sealy, M.L. Khristova, et al., Newly discovered ebola virus associated with hemorrhagic fever outbreak in Uganda, PLoS Pathog. 4 (2008) e1000212.

[19] E. Subbotina, A. Dadaeva, A. Kachko, A. Chepurnov, Genetic factors of Ebola virus virulence in guinea pigs, Virus Res. 153 (2010) 121–133.

[20] J. Felsenstein, G.A. Churchill, A hidden Markov model approach to variation among sites in rate of evolution, Mol. Biol. Evol. 13 (1996) 93–104.

[21] S. Baize, D. Pannetier, L. Oestereich, et al., Emergence of Zaire Ebola virus disease in Guinea, N. Engl. J. Med. 371 (2014) 1418–1425.

[22] A. Sanchez, S.G. Trappier, B.W. Mahy, et al., The virion glycoproteins of Ebola viruses are encoded in two reading frames and are expressed through transcriptional editing, Proc. Natl. Acad. Sci. U S A. 93 (1996) 3602–3607.

[23] V.E. Volchkov, H. Feldmann, V.A. Volchkova, H.D. Klenk, Processing of the Ebola virus glycoprotein by the proprotein convertase furin, Proc. Natl. Acad. Sci. U S A 95 (1998) 5762–5767.

[24] T. Hoenen, D. Safronetz, A. Groseth, et al., Mutation rate and genotype variation of Ebola virus from Mali case sequences, Science 348 (2015) 117–119.

[25] J.R. Kugelman, M.R. Wiley, S. Mate, et al., Monitoring of Ebola virus Makona evolution through establishment of advanced genomic capability in Liberia, Emerg. Infect. Dis. 21 (2015).

[26] J.B. Varkey, J.G. Shantha, I. Crozier, et al., Persistence of Ebola virus in ocular fluid during convalescence, N. Engl. J. Med. (2015), http://dx.doi.org/10.1056/NEJMoa1500306.