Characterization of the L gene and 5′ trailer region of Ebola virus

Viktor E. Volchkov,1,2 Valentina A. Volchkova,1,2 Alexandr A. Chepurnov,2 Vladimir M. Blinov,2 Olga Dolnik,1 Sergej V. Netesov2 and Heinz Feldmann1

1 Institut für Virologie, Philipps-Universität, Robert-Koch-Str. 17, 35037 Marburg, Germany
2 State Research Center of Virology and Biotechnology ‘Vector’, Institute of Molecular Biology, 633159 Koltsovo, Novosibirsk Region, Russia

The nucleotide sequences of the L gene and 5′ trailer region of Ebola virus strain Mayinga (subtype Zaire) have been determined, thus completing the sequence of the Ebola virus genome. The putative transcription start signal of the L gene was identical to the determined 5′ terminus of the L mRNA (5′ GAGGAAGAUUAA) and showed a high degree of similarity to the corresponding regions of other Ebola virus genes. The 3′ end of the L mRNA terminated with 5′ AUUAUAAAAAA, a sequence which is distinct from the proposed transcription termination signals of other genes. The 5′ trailer sequence of the Ebola virus genomic RNA consisted of 676 nt and revealed a self-complementary sequence at the extreme end which may play an important role in virus replication. The L gene contained a single ORF encoding a polypeptide of 2212 aa. The deduced amino acid sequence showed identities of about 73 and 44% to the L proteins of Ebola virus strain Maleo (subtype Sudan) and Marburg virus, respectively. Sequence comparison studies of the Ebola virus L proteins with several corresponding proteins of other non-segmented, negative-strand RNA viruses, including Marburg viruses, confirmed a close relationship between filoviruses and members of the Paramyxovirinae. The presence of several conserved linear domains commonly found within L proteins of other members of the order Mononegavirales identified this protein as the RNA-dependent RNA polymerase of Ebola virus.

Introduction

The species Marburg virus (MBGV) and Ebola virus (EBOV) constitute the family Filoviridae in the order Mononegavirales, which includes all non-segmented, negative-strand (NNS) RNA viruses (Murphy et al., 1995). Both viruses cause severe haemorrhagic fever in man and monkeys with mortality rates of up to 88% (Peters et al., 1996; Feldmann & Klenk, 1996).

Analysis of genomic RNA and virus-specific mRNA showed that filoviral genes are arranged in the following linear order: 3′ (leader)-NP-VP35-VP40-GP-VP30-VP24-L-(trailer) 5′. Each gene is flanked by highly conserved transcriptional start (3′ end) and stop (5′ end) signals that have been predicted by sequence comparison analysis (Feldmann et al., 1992; Volchkov et al., 1992; Sanchez et al., 1993; Bukreyev et al., 1995) and partly by direct sequence determination of the termini of mRNA species (Mühlberger et al., 1996). With the exception of the EBOV glycoprotein (GP) gene, all filoviral genes code for single structural proteins that are translated from monocistronic mRNA species that are complementary copies of the negative-stranded corresponding genes. For the EBOV GP gene, a different organization is observed. The primary product is a small glycoprotein (sGP) which is secreted as an anti-parallel orientated homodimer (Volchkova et al., 1998; Sanchez et al., 1998), whereas the transmembrane glycoprotein (GP) is translated from an edited mRNA species (Volchkov et al., 1995; Sanchez et al., 1996). Recently, it has been shown that the transmembrane GP is proteolytically processed into disulfide-linked subunits, GP1 and GP2 (Volchkov et al., 1998a, b).

The GenBank accession number of the sequence reported in this paper is AF086833.
Viral RNA (vRNA) was isolated from purified virus by (avian myeloblastosis virus). RNA–cDNA hybrids were poly(dC)-tailed, library with random primers (hexanucleotides) and reverse transcriptase (Volchkov centrifugation on caesium chloride gradients as described previously DFG strain was cultured in E6 cells, a clone of Vero cells (ATCC CRL-1586), was received from the Institute of Microbiology and Epidemiology, Viruses and cell lines.

Methods

**Viruses and cell lines.** The Mayinga strain of EBOV subtype Zaire was received from the Institute of Microbiology and Epidemiology, Minsk, Byelorussia. After a single passage in macaque monkeys, the virus strain was cultured in E6 cells, a clone of Vero cells (ATCC CRL-1586), and purified as described previously (Chepurnov et al., 1995). Vero E6 cells were cultured in Dulbecco’s medium supplemented with 10% foetal bovine serum.

**Molecular cloning and sequencing of the L gene and the 5’ trailer.** Viral RNA (vRNA) was isolated from purified virus by centrifugation on caesium chloride gradients as described previously (Volchkov et al., 1992) and used for the construction of a cDNA–RNA library with random hexanucleotides and reverse transcriptase (avian myeloblastosis virus). RNA–cDNA hybrids were poly(dC)-tailed, cloned into the PstI-digested, poly(dG)-tailed vector plasmid pBR322, and used to transform competent E. coli cells. Recombinant plasmids were identified by in situ colony hybridization using 32P-labelled vRNA; sequence determination was performed for both DNA strands by the method of Maxam & Gilbert (1980). Results were subsequently confirmed by direct sequencing of RT–PCR fragments synthesized from vRNA (Titan Amplification kit; Boehringer) using a cycle sequencing technique based on the dideoxy chain termination method (Sanger et al., 1977). The 5’ trailer sequence of the EBOV vRNA was determined by a primer extension method and subsequent chemical sequencing of the products according to Sambrook et al. (1989). Briefly, cDNA synthesis was performed with 20 µg purified genomic RNA using an excess of [32P]dATP-labelled ssDNA primers N1 (5’ GTAAATATGAGATTAGGGT; nt 18675–18694; mRNA sense) or N2 (5’ CCTGGAAAAAATGTTCCGACAC; nt 18894–18914; mRNA sense) and reverse transcriptase (40 U) (Boehringer). cDNA bands were purified from a 6% polyacrylamide gel and subsequently sequenced by the method of Maxam & Gilbert (1980).

**Elucidation of the 5’ terminus of the L mRNA.** RNA was isolated from infected Vero cells (EBOV subtype Zaire; m.o.i. 1 p.f.u. per cell) using a commercial kit (RNAeasy kit; Qiagen). Total RNA (10 µg) was used for reverse transcription, which was carried out in a volume of 50 µl at 43 °C for 2 h with the primer N3 (5’ GGACTATGAAATCTTATGGG; complementary to nt 11812–11793; mRNA sense). The 5’ end of L mRNA was analysed with the 5’-Ampitinder RACE kit (Clontech). Briefly, an ssDNA linker carrying an EcoRI site (underlined) (3’ GGAGACTTCCAAGGGTCTATGATCATGACGACp 5’) was ligated to the purified cDNA for 16 h at 22 °C using T4 RNA ligase (Boehringer). Subsequently, PCR amplification (35 cycles at 94 °C for 1 min, 60 °C for 1 min, and 72 °C for 2 min) was performed with 1 µl of a 1/10 dilution of the ligation sample with an L gene-specific primer N4 carrying a BamHI site (underlined) (5’ CTTGGATTGGTATCATCAGTTTCAAAC; complementary to nt 11762–11734; mRNA sense) and a primer complementary to the ssDNA linker. The PCR fragment was digested with BamHI and EcoRI and cloned into the vector pGEM-3Zf(+) (Promega). The sequence was determined by the cycle sequencing technique based on the dideoxy chain termination method (Sanger et al., 1977).

**Elucidation of the 3’ terminus of the L mRNA.** To determine the 3’ end of L mRNA, first-strand cDNA synthesis was performed on total RNA using the synthetic oligonucleotide primer BamHI-oligod(T)10 containing a 5’ BamHI site. After cDNA synthesis, the sample was heated at 95 °C for 5 min, cooled on ice, and amplification was carried out (35 cycles of 94 °C for 1 min, 40 °C for 1 min and 72 °C for 1 min) using the PCR Reagent kit (Perkin-Elmer) with the L gene-specific primer N5 (5’ ACATAGATGATGTTATCTGAAATG; nt 18128–18155; mRNA sense). Nested PCR amplification was carried out using 1 µl first-round PCR products and primers BamHI-oligod(T)10 (see above) and N6 (5’ GTTTATTTCCCGATGTC; nt 18184–18201; mRNA sense). After BamHI digestion, PCR products were ligated into BamHI- and Smal-digested pGEM-3Zf(+) and transformed into E. coli. Sequence determination of recombinant clones was done by a cycle sequencing technique based on the dideoxy chain termination method (Sanger et al., 1977).

**In vitro transcription/translation.** A cDNA fragment containing the complete L ORF was generated from viral RNA by RT–PCR using the oligonucleotides N7 (5’ TTCGTCAGTTGATATTTAATGGCC-TACACATAC; nt 11561–11590; mRNA sense, 5’-terminal Sall site underlined) and N8 (5’ GGACTATGAAATCTTATGGG; nt 18675–18694; mRNA sense, 5’-terminal Sall site underlined) (GenBank accession no. AF086833). The cDNA product was cloned into pGEM-3Zf(+) using
the appropriate restriction enzyme sites. The primary structure of the L gene was verified by sequence determination. *In vitro* transcription/translation of plasmid pGEM-LEBO was performed using the TnT T7 Quick Coupled kit according to the protocol of the supplier (Promega).

**Computer analysis.** Analysis of nucleotide and amino acid sequences was performed with the Alignment Service Package (Resen'chuk & Blinov, 1995) and the Heidelberg UNIX Sequence Analysis Resources (HUSAR), release 3.0, program FOLD RNA.

**Results and Discussion**

**Complete sequence determination of the EBOV genome**

Random primers were used for the synthesis of a cDNA library from isolated genomic RNA of EBOV (subtype Zaire, strain Mayinga). Resulting cDNA fragments were cloned and recombinant clones with EBOV-specific cDNA were selected by colony hybridization using $\text{32P}$-labelled genomic RNA. Nine clones were identified by sequence determination as carrying L gene-specific sequences. These cDNA fragments overlapped and totally covered the L gene at least twice. Subsequently, the primary structure of the EBOV L gene was independently confirmed by direct sequence determination of RT–PCR products obtained from genomic RNA. To determine the nucleotide sequence of the 5′ trailer region of the genome, purified viral RNA was reverse transcribed using 5′-$\text{32P}$-labelled primers. Labelled products were purified on 6% polyacrylamide gels and sequenced according to the method of Maxam & Gilbert (1980). The 3′ end of the shortest cDNA fragment displayed the sequence 5′ TTTTTTGTGTGTCCA (plus polarity) (Fig. 1a, b). The two larger bands seen with this analysis most likely represent complete cDNA fragments that have not been cleaved after chemical modification. All larger cDNA fragments displayed the same sequence (5′ TTTTTTGTGTGTCCA) (Fig. 1b), but were extended with various lengths of nucleotides complementary to genomic RNA starting at position 18903. This may be due to a reverse complementarity of the 5′ genomic end and a region extending over positions 18904–18919 which seems to lead to a copy-back mechanism during cDNA synthesis (Fig. 1b, c). This study completed previous investigations and finally resulted in the first entire primary structure of an EBOV genome with a total length of 18957 nt.

The 3′ leader sequence of the genome of EBOV (subtype Zaire, strain Mayinga), as determined previously, is 55 nt in length and extended from the first nucleotide to the transcriptional start signal of the NP gene at position 56 (Sanchez et al., 1993). The 5′ trailer, as determined here, started at position 18282 and was 676 nt. Thus, the genomic trailer region of EBOV is much larger than that of MBGV (Mühlberger et al., 1992; Bukreyev et al., 1995). However, it is interesting to note that due to the significantly longer 3′ end non-coding sequence of MBGV L mRNA, the distance between translation stop codons and 5′ termini of viral RNA for both viruses is similar. Regulation of negative-strand RNA synthesis could be proposed as a function for this region of viral RNA. For some other NNS RNA viruses, it has been demonstrated that the 3′ regions of the genome and antigenome serve as
templates for the synthesis of small plus and minus leader RNAs (Colonna & Banerjee, 1977; Leppert et al., 1979; Shioda et al., 1986). Similar virus-specific RNAs, however, have not yet been determined for any filovirus. The genomic termini showed a high degree of complementarity, as demonstrated previously for MBGV (Feldmann et al., 1992) and other NNS RNA viruses, such as Sendai virus (Shioda et al., 1986), Newcastle disease virus (Yusoff et al., 1987), human para-influenza type 3 virus (Galinski et al., 1988) and rabies virus (Tordo et al., 1988). Computer-modelled secondary structure predictions favoured the possibility of a panhandle structure between the genomic termini starting with base paring of nt 1 and 18 957 (Fig. 2a). However, it remains questionable as to whether such a structure can form due to the expected co-transcriptional encapsidation of the genome and antigenome. The analysis further suggested the possibility of almost identical hairpin-like structures for the 3' ends of the genome and antigenome (Fig. 2b, c). Similar structures can also be found with MBGV and thus may be important elements of the genomic and antigenomic promoters. The existence of such a structure at the 5' genomic end is supported by the finding of an additional non-genomic guanosine residue (underlined) was found in cloned PCR fragments between the putative consensus transcriptional start sequence and the sequence corresponding to the DNA linker. This guanosine residue most likely represents the 5' cap structure of L mRNA. Capping of filoviral mRNA has not yet been convincingly demonstrated, but indirect evidence has been obtained from the inhibition of EBOV replication in the presence of S-adenosylhomocysteine hydrolase inhibitors (Huggins et al., 1996). The second sequence (5' ATTAAAGAAAAA, cDNA strand) was repeated seven times on the viral genome and corresponded to the previously proposed transcriptional stop signal at the 3' terminus of viral mRNAs (Sanchez et al., 1989, 1993; Volchkov et al., 1992). Sequence analysis of the 5' genomic end failed to identify the highly conserved canonical sequence downstream of the L gene ORF, but several regions were found instead that differed by two nucleotide positions. To determine the exact 3' terminus of the L mRNA, the expected region was amplified from total RNA of EBOV-infected Vero cells, clone E6, using an oligo(dT)- and a 5'-proximal L mRNA-specific primer. Direct sequence analysis of the amplification products identified the most upstream located putative region with the sequence 5' ATTAAAGAAAAA at position 18 270 as the L gene transcriptional termination signal (Fig. 3a, c). Thus, the L mRNA was headed by a 5' untranslated region of 80 nt and followed by a 3' untranslated region of 63 nt.

Based on the precise determination of the 3' and 5' ends of the EBOV L mRNA as shown here (Fig. 3), the 5' end of the EBOV NP mRNA (Sanchez et al., 1989), and the 3' and 5' ends of all MBGV mRNA (Mühlberger et al., 1996), it can be concluded that transcriptional start and termination signals are highly conserved among all filoviruses, a feature which is shared with other NNS RNA viruses. Interestingly, all signals contain the pentamer 5' AUUAA 3', with the exception of the termination signals of the MBGV VP40 (Feldmann et al., 1992;
Fig. 3. Sequence determination of the L gene mRNA termini. 
(a) Schematic representation of the L mRNA start and stop signals. 
Transcription termination signals for the VP24 and L genes are underlined. 
The GenBank accession number of the sequence of the full-length Ebola virus genomic RNA is AF086833. 
(b) 5′ terminus. The 5′ end of the L mRNA was analysed using the 5′-Amplifinder RACE technique. 
Total RNA (10 μg) from EBOV-infected Vero cells was used for reverse transcription using primer N3. 
An ssDNA linker was ligated to the purified cDNA and subsequently PCR amplification was performed using 
an L gene-specific primer and a primer complementary to the ssDNA linker. The PCR fragment was cloned and sequenced. 
An additional non-genomic guanosine residue (underlined) was found between the consensus transcriptional start signal (5′ GAGGAAGATTAA 3′) and the sequence of the linker (5′ GAATTCGTG 3′). 
(c) 3′ terminus. First-strand cDNA synthesis was performed on total RNA using the primer BamHI-oligo(dT)21. 
Subsequently, an L gene-specific primer (N5) was added for PCR amplification which was followed by a nested amplification using primers BamHI-oligo(dT)21 and N6. 
The PCR product was cloned and sequenced. The sequences of the altered transcriptional stop signal and the poly(A) stretch are underlined (cDNA sequences are presented).

Mühlberger et al., 1996) and the EBOV L gene in which alterations are found. The function of this highly conserved pentamer is still unknown. It may serve as the actual recognition signal for the polymerase complex, whereas an upstream semi-conserved region may direct initiation (start signal) and a downstream stretch of uridine residues may guide termination (termination signal) of transcription (Feldmann et al., 1992, 1993).

Based on the prediction that ATG<sub>53-57</sub> (see above) initiates translation, the L protein is 2212 aa with a calculated molecular mass of 252722 Da. Thus, the size is in agreement with that of the L protein of MBGV (267-2 kDa) and those of other RNA-dependent RNA polymerases of the Paramyxoviridae and Rhabdoviridae (242–256 kDa). SDS–PAGE analysis of in vitro transcription/translation of the EBOV L protein from plasmid pGEM-LEBO, which carries the entire L ORF under the control of the bacteriophage T7 polymerase promoter, revealed a single band migrating at approximately 220 kDa, which is close to previous molecular mass estimations by SDS–PAGE for filovirus L proteins (Elliott et al., 1985; Kiley et al., 1988). The discrepancy in defining the actual molecular mass seems to be due to the uncertainty inherent in migration of such large proteins on SDS–polyacrylamide gels (Fig. 4).

Sequence comparison analyses

The predicted L protein sequence showed significant identity with those of a subtype Sudan EBOV (EBOVS) (GenBank accession no. U23458) and two previously deposited MBGV isolates, Musoke (MBGVM) (Mühlberger et al., 1992) and Popp (MBGVP) (Bukreyev et al., 1995). The overall identities were 73-36% with subtype Sudan, 44-3% with MBGVM and 44-25% with MBGVP. As expected, the two EBOV subtypes showed a higher degree of conservation.
Fig. 5. Amino acid sequence alignment. The analysis was performed with the Alignment Service Package on an IBM computer (Resenchuk & Blinov, 1995). Three conserved motifs are underlined and marked by A, B and C. Six conserved amino acid blocks were identified which represent previously published regions of sequence conservation among Paramyxoviridae (block I, aa 226–426; block II, aa 513–614; block III, aa 623–841; block IV, aa 899–1096; block V, aa 1097–1359; and block VI, aa 1802–1878). EBOVZ, Ebola virus subtype Zaire, strain Mayinga; EBOVS, Ebola virus subtype Sudan, strain Maleo; MBGVM, Marburg virus, strain Musoke; MBGVP, Marburg virus, strain Popp. Dashes represent gaps introduced to optimize the alignment. Dots indicate identical amino acids when compared with EBOVZ.
compared with MBGV strains, but were more distinct from each other than the two MBGV isolates (96% identity). This supports previous phylogenetic and sequence comparison data on other filoviral genes demonstrating the existence of two species in the family Filoviridae, Marburg and Ebola, with a separation into subtypes only for EBOV (Sanchez et al., 1996; Feldmann et al., 1997; Volchkov et al., 1997; Georges-Courbot et al., 1997). Computer-assisted dot-plot analysis of the EBOV L protein and those of other NNS RNA viruses confirmed conservation among filoviruses (data not shown). As demonstrated previously for MBGV (Mühlberger et al., 1992), the analysis further confirmed identities to several paramyxoviruses and the closest relationship was observed with viruses of the genera Paramyxovirus and Morbillivirus. The EBOV L protein also shared several features with almost all L proteins of NNS RNA viruses, such as high leucine and isoleucine content (17-9%, which is more than 1-2 times that of an ‘average’ protein (Dayhoff et al., 1978)), and a large positive net charge (+57) at neutral pH (calculated by the convention that lysine and arginine have charges of +1, aspartic acid and glutamic acid are −1, and histidine is +0.5) with several clusters of basic amino acids interspersed with non-basic ones as found in several RNA binding proteins (Lazinski et al., 1989).

The distribution of homologous regions across filoviral L proteins showed three parts that have diverged to a greater extent: the N termini (first 350 aa); the C termini (last 210 aa); and aa 1649–1800 of the aligned sequences (Fig. 5). The alignment also revealed two blocks (aa 1147–1168 and aa 1672–1766) that are unique to MBGV. Almost all enzymatic activities involved in transcription and replication of NNS RNA viruses have been associated with the L protein. Previous studies have suggested that linear domains play an important role in certain functions of these proteins (Poch et al., 1990). The comparative analysis of filovirus L proteins shown in Fig. 5 confirmed the existence of six amino acid blocks which represent regions of significant conservation interrupted by more variable sequences. Three polymerase motifs (A–C) identified by Poch et al. (1990) for NNS viruses and previously found to be present in the MBGV L protein (Mühlberger et al., 1992) were also found with EBOV L proteins, including an RNA binding element (motif A, aa 553–571), a putative RNA template recognition and/or phosphodiester bond formation domain (motif B, aa 738–744), and an ATP and/or purine ribonucleotide triphosphate binding domain (motif C, aa 1815–1841) (Fig. 5). Furthermore, highly conserved twin cysteine residues are present in all L proteins (aa 1351–1352) which may possibly stabilize the secondary structure of the proteins and therefore be important for the proper conformation of the putative active sites (Blumberg et al., 1988; Mühlberger et al., 1992).

Finally, the completion of the genomic sequence and the availability of a full-length clone of the L gene will now allow reverse genetic systems to be established which will enable studies of transcription and replication of Ebola virus.

The authors thank Carina Eckel for excellent technical assistance. We further acknowledge I. Frolov and A. Kolykhhalov (Institute of Molecular Biology, Koltsovo, Russia) for their help with sequencing, and H.-D. Klenk, S. Becker and E. Mühlberger (Institute of Virology, Marburg, Germany) for discussions. We thank A. Sanchez, S. Trappier and P. Calain (CDC, Atlanta, USA) for providing unpublished data on the L-gene of subtype Sudan Ebola virus (strain Maleo). This work was supported by grants of the Deutsche Forschungsgemeinschaft (Fe 286/4-1, SFB 286 and SFB 535) and Russian Foundation of Basic Research (grant #93-04-00714). V.E.V. was a recipient of the Alexander von Humboldt-Stiftung.

References


Ebola virus L gene and 5’ trailer region