SEQUENCE DETERMINATION AND GENETIC ANALYSIS OF THE LEADER REGION OF VARIOUS EQUINE ARTERITIS VIRUS ISOLATES

A. Kheyar, G. St-Laurent, M. Diouri, J. Dufresne, and D. Archambault

Université du Québec à Montréal
Département des Sciences Biologiques
C. P. 8888, Succursale Centre-Ville
Montréal, Québec, Canada, H3C 3P8

1. ABSTRACT

The entire leader sequence of ten equine arteritis virus (EAV) isolates including the Bucyrus reference strain was determined and analyzed at the primary nucleotide and secondary structure levels. The leader sequence of eight EAV isolates was determined to be 206 nucleotides (nt) in length, whereas those of the 86AB-A1 and 86NY-A1 isolates were found to be 205 and 207 nt in length, respectively. The sequence identity of the leader sequences between the different isolates and the Bucyrus reference strain ranged from 94.2 to 98.5%. An AUG start codon found at position 14 in all EAV isolates could initiate an open reading frame (ORF) that could produce a polypeptide of 37 amino acids, except for the 86NY-A1 isolate where the intraleader polypeptide would contain 54 amino acids. Five patterns of computer-predicted RNA secondary structures were identified in the ten EAV leader regions analyzed. All EAV isolates showed three conserved stem-loops (designated A, B and C). An additional conserved stem-loop (D) was observed in six EAV isolates, including the Bucyrus reference strain. Based on the presence or absence of stem-loop D, all EAV isolates analyzed in this study could be tentatively classified into two genogroups (I and II). The significance of the intraleader ORF and the predicted secondary structures has yet to be determined.

2. INTRODUCTION

Equine arteritis virus (EAV) is the etiologic agent of equine viral arteritis. The clinical outcome following EAV exposure varies widely from subclinical infection to EAV sys-

Coronaviruses and Arteriviruses, edited by Enjuanes *et al.*
Plenum Press, New York, 1998

temic disease of variable intensity, indicating that virulence varies among EAV isolates (McCollum and Swertzek, 1978). Abortion is common when pregnant mares become infected (McCollum and Swertzek, 1978).

EAV genome is a positive, polyadenylated, single-stranded RNA of 12.7 kb in length (Den Boon et al., 1991). During EAV replication, a 3' end-coterminal nested set of seven virus-specific RNAs (ORFs 1 to 7) is produced. Each ORF is preceded by the sequence motif 5'-UCAAC- 3', designated the leader-body junction site, which is involved in the formation of six subgenomic EAV RNAs with a common leader sequence derived from the extreme 5' end of the viral genome (De Vries et al., 1990). ORFs 1a/1b encode for the viral replicase, and ORF 3- and 4-encoded products are believed to be glycosylated nonstructural proteins (Den Boon et., 1991; De Vries et al., 1992). Four viron structural proteins are produced from ORFs 2, 5, 6 and 7. ORFs 6 and 7 encode an unglycosylated membrane (M) protein of 16 kDa and a 14 kda nucleocapsid (N) protein, respectively. ORFs 2 and 5 encode the glycosylated 25 kDa small (G_S) membrane protein and the heterogeneously glycosylated 30- to 42-kDa large (G_L) membrane protein, respectively (De Vries et al., 1992).

Very little information is available for the EAV leader region. We recently determined the sequence of the terminal 5' leader region of four isolates of EAV, including the Bucyrus reference strain (Kheyar et al., 1996). We, and other investigators (Van Dinten et al., 1997), showed that the leader region is 206 nucleotides (nt) in length (not including a putative 5' cap structure-associated nt). We also identified an intraleader ORF of 111 nt in length (Kheyar et al., 1996).

Leader sequences of several other families of RNA positive-sense viruses such as picornaviruses (Duke et al., 1992; Haller et al., 1996), flaviviruses (Fukushi et al., 1994) or coronaviruses (Chen and Baric, 1995; Hofmann et al., 1993) have been shown to be involved in the regulation of critical viral functions (such as replication, transcription and/or translation) by the involvement of either intraleader ORFs or sequences that, by intramolecular base pairing, form unique secondary and tertiary structures. These considerations prompted us to extend our analysis of the EAV leader region. The primary goal of this sudy was to determine the complete leader sequence of the ATCC Bucyrus strain of EAV and of nine EAV field isolates originating from both different geographic regions (United States, n = 3; Canada, n = 5; and Austria, n = 1) and years of isolation to allow a comparative genetic analysis. Second, we wished to compare the predicted RNA secondary structures in the leader region of these isolates and consider how these structures might have a role in EAV biogenesis.

3. MATERIALS AND METHODS

The EAV isolates used in this study are listed in Table 1. Virus propagation and EAV genomic RNA extraction of each isolate were performed as described (Kheyar et al., 1996). The Rapid Amplification of cDNA ends (RACE) method based on the single strand ligation to single-stranded cDNA (SLIC) was used to obtain sequences of the 5' end of each EAV isolate genome (Kheyar et al., 1996). EAV genomic RNA was reverse transcribed using the procedure described in the Amplifinder Race Kit (Clontech Laboratories, Palo Alto, CA). The reverse transcription reaction was primed with the antisense oligonucleotide primer PEV-L1, which is complementary to nt 306 to 323 of the EAV genome (Den Boon et al., 1991). After RNA hydrolysis, the single-stranded cDNA was purified and ligated to the 3' end blocked AmpliFINDER anchor sequence with T4 RNA ligase.

Table 1. Characteristics of the EAV isolates used in this study

Isolate	Origin (year of isolation)	Source	Passage history*
Bucyrus**	Ohio, USA, (1953)	Fetus lung	Horse P14/LLC-MK2, P1
87AR-A1	Arizona, USA, (1987)	Semen	RK, P4/V, P7
86AB-A1	Alberta, Canada, (1986)	Fetus	RK, P4/V, P6
86NY-A1	New-York, USA, (1986)	Semen	ED, P4/V, P2
84KY-A1	Kentucky, USA(1984)	Nasal swab	RK, P5
Vienna	Austria (1968)	Nasal swab	ED, P1/RK, P2
19933	Guelph, Canada, (1992)	Semen	RK, P5
15492	Guelph, Canada, (1991)	Semen	RK, P5
11958	Guelph, Canada, (1990)	Semen	RK, P5
T1329	Guelph, Canada, (1988)	Neonatal lung***	RK, P5

*Cells: ED; equine dermis, HK; primary horse kidney, LLC-MK2; Rhesus monkey kidney, RK; rabbit kidney-13, V; Vero, P; refers to passage number.

**ATCC number: VR-796.

***EAV was isolated from 5 days old standardbred foal.

The single stranded ligation product was then amplified in the polymerase chain reaction (PCR) procedure by using the antisense primer PEV-L02, which is complementary to an internal sequence (nt 165 to 145) of the EAV leader sequence (Den Boon et al., 1991), and the sense AmpliFINDER anchor primer (AFAP). A second set of primers (sense, nt 1 to 18; antisense, complementary to nt 165 to 145) was selected according to den Boon et al. (1991) and was used to obtain, by PCR, overlapping cDNA fragments containing the remainder of the leader sequence at the 3' end. PCR amplifications were performed with the Taq polymerase (Promega, Madison, WI). The gel-purified PCR cDNA products for each isolate were cloned into Sma I-cleaved pBluescript II KS+ (Stratagene, La Jolla, CA) or PCR^{TM}II (Invitrogen, San Diego, CA) plasmid vectors. Two or more clones of each PCR products (both strands) were sequenced using the Sanger dideoxynucleotide chain termination method (Sanger et al., 1977).

Comparison and multiple alignments of nucleic acid sequences were carried out with the University of Wisconsin Genetics Computer Groups software package (GCG, version 8.1). Putative secondary structures of the leader sequence of all EAV isolates exhibiting minimum-free energies were predicted with the FOLD and SQUIGGLES graphics output programs of the GCG software package.

4. RESULTS

The complete nucleotide sequences of the leader region of all EAV isolates including that of the ATCC Bucyrus strain are depicted in Figure 1. The RACE method successfully identified the extreme 5' ends of the genomes of all of the EAV isolates studied and revealed the presence of 17 additional nt for nine EAV isolates, not previously identified, located upstream of the published sequence (Den Boon et al., 1991). In the case of the 86AB-A1 isolate, the first of these 17 nt was missing. The 17 base sequence in the T1329 isolate was identical to that of the Bucyrus reference strain. A G→A substitution at position 1 was observed for the 87AR-A1, Vienna and 15492 isolates. Additional base substitutions of A→G at positions 6 (86AB-A1 isolate) or 7 (86AB-A1, 84KY-A1, 86NY-A1, 11958 and 19933 isolates), and C→G or T at position 2 (86AB-A1 and 19933 isolates, respectively)

```
BUCYRUS    GCTCGAAGTG TGTATGGTGC CATATACGGC TCACCACCAT ATACACTGCA    50
87AR-A1    A......... .......... .......... .....G.... ..G.......
86AB-A1    -G...GG... .......... .......... .......... ..G.......
86NY-A1    ......G... .......... .......... .......... ..........
84KY-A1    ......G... .......... .......... .......... ..........
VIENNA     A......... .......... .......... ........G. ..........
19933      .T....G... .......... .......... ..C..G.... ..........
15492      A......... .......... .......... .....G.... ..........
11958      ......G... .......... .......... .......... ..........
T1329      .......... .......... .......... .....G.... ..........

BUCYRUS    AGAATTACTA TTCTTGTGGG CCCCTCTCGG TAAATCCTAG AGGGCTTTCC    100
87AR-A1    .......... .......... .......... .......... ..........
86AB-A1    .......... .......... .......... .......... ..........
86NY-A1    .......... .......... .......... .......... ........T.
84KY-A1    .......... .......... .......T.. .......... ..........
VIENNA     .......... .......... .......T.. ...T...... ..........
19933      .......... .......... .......... .......... ..........
15492      .......... .......... .......... .......... ..........
11958      .......... .......... .......... .G........ ..........
T1329      .......... .......... .......T.. ...C...... ..........

BUCYRUS    TCTCGTTATT GCGAGATTCG TCGTTAGATA ACGGCAAGTT -CCCTTTCTT    150
87AR-A1    .......... .........T .......... .......... -.........
86AB-A1    .......... .......... .......... .......... -.........
86NY-A1    ........C. .......... .....GT... .......... C.........
84KY-A1    .......... .......... .......... .......... -.........
VIENNA     .......... .......... .......T.. .....-.T.. C..TA.....
19933      .......... .......... ......A... .......... -.........
15492      .......... .......... .......... .........C -..T......
11958      .......... .......... .......... .......... -........A
T1329      .......... .......... .......... .......A.. -.........

BUCYRUS    ACTATCCTAT TTTCATCTTG TGGCTTGACG GGTCACTGCC ATCGTCGTCG    200
87AR-A1    .......... .......... .......... .......... ..........
86AB-A1    .......... .......... .......... .......... ..........
86NY-A1    ...T...... .......... .......... .......... ..........
84KY-A1    ...T...... .......... .......... .A........ ..........
VIENNA     .TCT...... .......... .......... A......... ..........
19933      ...T...... .......... .......... .......... ..........
15492      ...T...... .......... .......... .......... ..........
11958      .......... .......... .......... .......... ..........
T1329      ...T...... .......... .......... .......... ..........

BUCYRUS    ATCTCTA
87AR-A1    .......
86AB-A1    .......
86NY-A1    .......
84KY-A1    .......
VIENNA     .......
19933      .......
15492      .......
11958      .......
T1329      .......
```

Figure 1. Alignment of the nucleotide sequences of the leader region of nine arteritis virus isolates with the Bucyrus reference strain. AUG start codons are double-underlined and in-frame stop codons are also underlined. The GenBank accession numbers of the different isolates are the following: AF001259 (ATCC Bucyrus); U46946 (87AR-A1); U65723 (86AB-A1); U46945 (86NY-A1); U65724 (84KY-A1); U46947 (Vienna); U65728 (19933); U65727 (15492); U65726 (11958); U65725 (T1329).

were also observed by comparaison with the Bucyrus reference strain. The identification of these additional nt brings the length of the leader sequence to 206 nt without the junction site motif and the putative 5' cap structure-associated nt for most EAV isolates, except the 86AB-A1 isolate which is 205 nt. Another exception is the 86NY-A1 isolate, which has a leader region of 207 nt in length due to a C insertion at position 141.

When the leader sequences were aligned and compared to that of the Bucyrus reference strain, a total of 29 base positions varied among isolates (14% of the 206 nt positions) (Figure 1). Beside the substitutions mentioned above that occurred at the extreme 5' end of the leader region, base substitutions were common at particular positions. For instance, substitutions of A→G at positions 36 and 43, C→T at position 78, and A→T at position 154 were observed with regularity. Such base mutations are common in RNA viruses (Domingo et al., 1985).

The levels of identity for the leader sequence of the nine EAV field isolates to the Bucyrus reference strain ranged from 94.2 (Vienna isolate) to 98.5% (11958 isolate) (data not shown). When the leader sequences of the EAV field isolates were compared to each other, both the 86AB-A1 and 11958, and the 84KY-A1 and T1329 isolates were the most closely related at 97.6% identity. Sequences of the North American isolates (five Canadian and four American, including the Bucyrus strain) were 95.2 to 98.5% identical when compared to each other, while all were less closely related to the Vienna European isolate where the identity of the sequence ranged from 92.2 (86NY-A1 and 19933 isolates) to 94.6% (T1329 and 15492 isolates).

A striking feature in all of the EAV leader sequences was the presence of a AUG start codon located at positions 14 to 16. This start codon has not been previously identified because it is located upstream of the published sequence (den Boon et al., 1991). The presence of an amber (UAG) or an ochre (UAA) stop codon for all EAV isolates, except the 86NY-A1 isolate, at positions 125 to 127 in frame with the first start codon predicts an ORF of 111 nt that could encode a small polypeptide of 37 amino acids. In contrast, the 86NY-A1 isolate contains an opal (UGA) stop codon at positions 176 to 178, so that its intraleader ORF is 162 nt in length and the predicted polypeptide would include 54 amino acids. When the sequences of the predicted 37 amino acid peptides for all EAV isolates were aligned and compared to that of the Bucyrus reference strain, only a few amino acid substitutions were observed and, in most cases, were common among the EAV field isolates (for instance, Arg→His at position 8, Met→Ile at position 10, and Ser→Leu at position 22) (data not shown). The significance of these variations is not known.

The putative secondary structures of the leader sequence of all EAV isolates are depicted in Figure 2. Based on the number and locations of predicted stem-loop structures, five different secondary structure patterns were obtained. Three of the predicted stem-loop structures, which would form in the first half of the leader region at nt 7 to nt 100, designated A, B and C, were found in all of the EAV isolates. Moreover, a fourth stem-loop, designated D, was predicted for the six isolates comprising patterns 1 and 2. This suggests that the EAV isolates might segregate into two major groups (I and II) based on the presence or absence of stem-loop D.

Stem-loop A (41 nt in length) forms by base-pairing nucleotides from nt 7 to nt 47, stem-loop B (20 nt in length) uses nt 48 to nt 67, stem-loop C (33 nt in length) uses nt 68 to nt 100 (except for the 86NY-A1 isolate where a shorter stem-loop C is formed from nt 72 to nt 94), and stem-loop D (16 nt in length) uses nt 101 to nt 116. For most isolates, these sets of three or four stem-loop structures were branched out from an internal loop designated I. Although some nucleic acid substitutions were observed within the se-

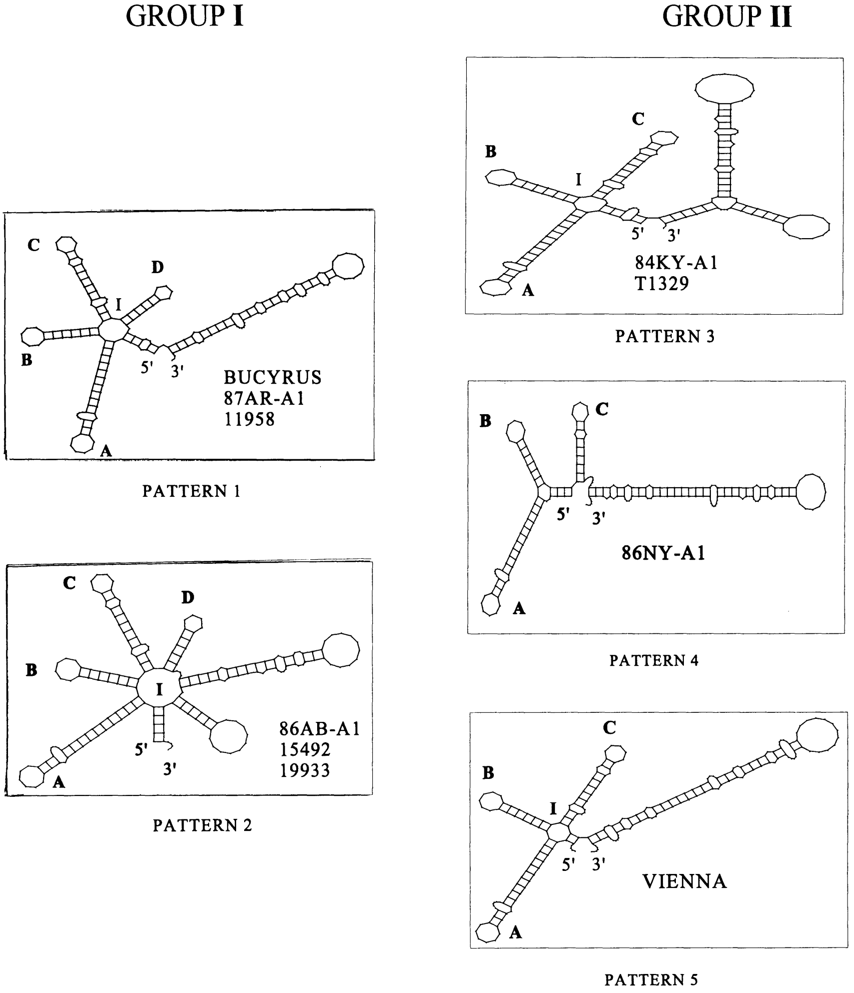

Figure 2. Predicted secondary structure of the leader sequence of ten EAV isolates using the FOLD program of the GCG software package. The conserved stem-loops (A to D) and the internal loop (I) are designated with upper cases in parenthesis.

quence of all four of these stem-loops, the general form and the length of these stem-loops were unaltered regardless of the mutations (data not shown) indicating a high degree of structural conservation for the stem-loop structures of both groups I and II. Such conservation of the general form of these stem-loop structures indicates that most mutational changes are not tolerated and suggests an important functional role of these structural domains in EAV replication or biogenesis. Furthermore, the presence of stem-loop D in six EAV isolates is of interest because it might alter or modulate EAV replication and/or virulence.

5. DISCUSSION

In this study, the complete leader region sequences of ten EAV isolates, including the ATCC Bucyrus strain were determined.The results demonstrated a high degree of nucleic acid conservation in the leader sequences among EAV isolates. Moreover, an intraleader ORF associated to a AUG start codon at residues 14 to 16 was found in all EAV isolates. A second AUG codon located at residues 41 to 43 was also observed in two EAV isolates (87AR-A1 and 86AB-A1) (Figure 1). This second AUG codon observed was in frame with the first start codon, so that a shorter ORF (nt 41 to nt 125) could theoritically produce a 28 amino acid polypeptide. However, the primary sequence in the immediate vicinity of an AUG initiation codon is important for initiation of translation (Kozak, 1991), and it is unlikely that the AUG codon at positions 41 to 43 of the EAV leader sequence would act as a start codon because, with both positions -3 and +4 occupied by a C residue, it is unlikely that translation could be initiated. In contrast, the AUG codon at positions 14 to 16 is likely to act as a functional start codon because, with the -3 and +4 positions occupied by U and G residue, respectively, it is capable, albeit at a suboptimal level, of initiating translation. Experiments to test for the presence of functional intraleader ORFs within the EAV leader region with the use of antisera to synthetic oligopeptides are in progress in our laboratory. This work may be of some interest because, as shown for coronaviruses, an intraleader ORF was reported to modulate virus translation efficiency either by attenuating (in bovine coronavirus, Hofmann et al., 1993) or by enhancing (in murine coronavirus, Chen and Baric, 1995) the translation rate of downstream ORFs.

Analysis of putative secondary structures predicted the presence of three stem-loop structures (A, B and C) in all EAV isolates. Moreover, we could theoritically segregate the EAV isolates into two major groups based on the presence or absence of the fourth stem-loop D. The nucleotide sequences and predicted secondary structures downstream from either stem-loop C (from nt 100 for group I) or stem-loop D (from nt 117 for group II) were found much more variable than the 5' end (Figures 1 and 2). This variability is seen in the sequence following the stop codon at nt 125–127 through nt 154, and is reflected by a variable pattern of additional possible stem-loop structures that would vary from one isolate to another in terms of position and sequence. Variability in this portion of the leader sequence, in marked contrast to the conserved sequence and limited variability in the secondary structures predicted for the 5' end, suggests a lesser role for this region of the genome in regulation of EAV replication. Nevertheless, such variation might also be associated with differences in EAV virulence, as seen elsewhere where variations in the secondary structure of the leader sequence of echovirus and Sabin poliovirus reflected variation in neurovirulence and the level of virus attenuation in tissue cultures, respectively (Macadam et al., 1992; Romero and Rotbart, 1995).

The role of secondary structures in leader sequences in the initiation of translation of eukaryotic mRNAs has been well documented (Kozak, 1991). Moreover, in picornaviruses (e.g. Theiler's virus), mutational analyses demonstrated that neurovirulence of certain strains was associated with a specific secondary structure in the leader sequence (Haller et al., 1996). Therefore, replication and translation strategies, as well as other features in EAV biogenesis such as virulence, might well be influenced by variations of the primary sequence, or RNA secondary structure within the leader region. Experiments with appropriate cDNA mutant clones are under way to investigate how the variation in EAV leader sequence secondary structure impacts on subgenomic mRNA translation efficiency.

ACKNOWLEDGMENTS

This work was supported by an operating grant from the National Sciences and Engineering Research Council of Canada to D. Archambault. A. Kheyar is supported by a graduate student fellowship from Université de Montréal. D. Archambault is the holder of a research scholarship from the Fonds de la Recherche en Santé du Québec (FRSQ). We thank Peter J. Timoney and William H. McCollum (Gluck Equine Research Center, Lexington, Kentucky), and Susy Carman (Veterinary Laboratory Services, Guelph, Ontario) for providing the EAV laboratory strains and field isolates. We are also grateful to John Dahlberg for reviewing the manuscript.

REFERENCES

Chen W. and Baric R.S., 1995, Function of a 5'-end genomic RNA mutation that evolves during persistent mouse hepatitis virus infection in vitro, *J. Virol.* **69:** 7529–7540.

Den Boon J.A., Snijder E.J., Chirnside E.D., De Vries A.A.F., Horzinek M.C. and Spaan W.J.M., 1991, Equine arteritis virus is not a togavirus but belongs to the coronaviruslike superfamily, *J. Virol.* **65:** 2910–2930.

De Vries A.A.F., Chirnside E.D., Bredenbeek P.J., Gravestein L.A., Horzinek M.C. and Spaan W.J.M., 1990, All subgenomic mRNAs of equine arteritis virus contain a common leader sequence, *Nucleic Ac. Res.* **18:** 3241–3247.

De Vries A.A.F., Chirnside E.D., Horzinek M.C. and Rottier P.J.M., 1992, Structural proteins of equine arteritis virus, *J. Virol.* **66:** 6294–6303.

Domingo E., Martinez-Salas E., Sobrino F., de la Torre J.C., Portela A., Ortin J., Lopez-Galindez C., Perez-Brena P., Villanueva N. and Najera R., 1985, The quasispecies (extremely heterogeneous) nature of viral RNA genome populations: biological relevance, a review, *Gene* **40:** 1–8.

Duke G.M., Hoffman, M.A. and Palmenberg A.C., 1992, Sequence and structural elements that contribute to efficient encephalomyocarditis virus RNA translation, *J. Virol.* **66:** 1602–1609.

Fukushi S., Katayama K., Kurihara C., Ishiyama N., Hoshini F.B., Ando T. and Oya A., 1994, Complete 5' noncoding region is necessary for the efficient internal initiation of hepatitis C virus RNA, *Biochem. Biophys. Res. Com.* **199:** 425–432.

Haller A.A., Stewart S.R. and Semler B.L., 1996, Attenuation stem-loop lesions in the 5' noncoding region of poliovirus RNA: neuronal cell-specific translation defects, *J. Virol.* **70:** 1467–1474.

Hofmann M.A., Senanayake S. and Brian D.A., 1993, A translation attenuating intraleader open reading frame is selected on coronavirus mRNAs during persistent infection, *Proc. Natl. Acad. Sci. USA* **90:** 11733–11737.

Kheyar A., St-Laurent G. and Archambault D., 1996, Sequence determination of the extreme 5' end of the equine arteritis virus leader region, *Virus Genes* **12:** 291–295.

Kozak M., 1991, An analysis of vertebrate mRNA sequences: intimations of translational control, *J. Cell Biol.* **115:** 887–903.

Macadam A.J., Ferguson G., Burlison J., Stone D., Skuce R., Almond J.W. and Minor P.D., 1992, Correlation of RNA secondary structure and attenuation of Sabin vaccine strains of poliovirus in tissue culture, *Virology* **189:** 415–422.

McCollum W.H. and Swertzek T.W., 1978, Studies of an epizootic of equine arteritis virus in racehorses, *J. Equine Med. Surg.* **2:** 459–464.

Romero J.R. and Rotbart H.A., 1995, Sequence analysis of the downstream 5' nontranslated region of seven echovirus with different neurovirulence phenotypes, *J. Virol.* **69:** 1370–1375.

Sanger F., Nicklen S. and Coulson A.R., 1977, DNA sequencing with chain-terminating inhibitors, *Proc. Natl. Acad. Sci. USA* **74:** 5463–5467.

Van Dinten, L.C., Den Boon J.A., Wassenaar A.L.M., Spaan W.J.M. and Snijder E.J., 1997, An infectious arterivirus cDNA clone: identification of a replicase point mutation that abolishes discontinuous mRNA transcription, *Proc. Natl. Acad. Sci. USA* **94:** 991–996.