Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Genomic characterization of equine coronavirus

Jianqiang Zhang, James S. Guy, Eric J. Snijder, Doug A. Denniston, Peter J. Timoney, Udeni B.R. Balasuriya

Department of Veterinary Science, 108 Maxwell H. Gluck Equine Research Center, University of Kentucky, Lexington, KY 40546, USA
Department of Population Health and Pathobiology, College of Veterinary Medicine, North Carolina State University, Raleigh, NC 27606, USA
Molecular Virology Laboratory, Department of Medical Microbiology, Leiden University Medical Center, Leiden, The Netherlands

Received 16 April 2007; returned to author for revision 8 May 2007; accepted 30 June 2007

Available online 13 August 2007

Abstract

The complete genome sequence of the first equine coronavirus (ECoV) isolate, NC99 strain was accomplished by directly sequencing 11 overlapping fragments which were RT–PCR amplified from viral RNA. The ECoV genome is 30,992 nucleotides in length, excluding the polyA tail. Analysis of the sequence identified 11 open reading frames which encode two replicase polyproteins, five structural proteins (hemagglutinin esterase, spike, envelope, membrane, and nucleocapsid) and four accessory proteins (NS2, p4.7, p12.7, and I). The two replicase polyproteins are predicted to be proteolytically processed by three virus-encoded proteases into 16 non-structural proteins (nsp1–16). The ECoV nsp3 protein had considerable amino acid deletions and insertions compared to the nsp3 proteins of bovine coronavirus, human coronavirus OC43, and porcine hemagglutinating encephalomyelitis virus, three group 2 coronaviruses phylogenetically most closely related to ECoV. The structure of subgenomic mRNAs was analyzed by Northern blot analysis and sequencing of the leader–body junction in each sg mRNA.

© 2007 Elsevier Inc. All rights reserved.

Keywords: Equine coronavirus; Entire genome; Subgenomic RNA; Transcription regulatory sequence; Non-structural protein 3

Introduction

Coronaviruses are mainly associated with respiratory and gastrointestinal disease in humans (Drosten et al., 2003; Holmes, 2001; Ksiazek et al., 2003; Peiris et al., 2003; van der Hoek et al., 2004; Woo et al., 2005) and respiratory, enteric, neurological, or hepatic disease in animals (Holmes, 2001). Coronaviruses have also been isolated from bats, poultry and other birds (Cavanagh, 2005; Chu et al., 2006; Poon et al., 2005; Ren et al., 2006). On the basis of antigenic and genetic analyses, coronaviruses are divided into three groups (Gonzalez et al., 2003; Gorbalenya et al., 2004; Snijder et al., 2003). Group 1 viruses include human coronaviruses 229E (HCoV-229E) and NL63 (HCoV-NL63), canine coronavirus (CCoV), feline coronavirus (FCoV), porcine transmissible gastroenteritis virus (TGEV), porcine epidemic diarrhea virus (PEDV), and bat coronavirus. Group 2 viruses are subdivided into group 2a which includes murine hepatitis virus (MHV), human coronaviruses OC43 (HCoV-OC43) and HKU1 (HCoV-HKU1), bovine coronavirus (BCoV), porcine hemagglutinating encephalomyelitis virus (PHEV), and rat coronavirus (RCov), and group 2b which includes SARS-coronavirus (SARS-CoV). Group 3 viruses include avian viruses, such as avian infectious bronchitis virus (IBV), and turkey coronavirus (TCoV).

Members of the family Coronaviridae are enveloped, positive-stranded RNA viruses with exceptionally large, polycistronic genomes (27–32 kb). The 5′-proximal two-thirds of the genome comprises two open reading frames (ORFs), ORF1a and ORF1b, which encode the replicase polyproteins (pp) 1a and pp1ab (Ziebuhr, 2005). Expression of the pp1ab requires a −1 ribosomal frameshift during translation of the genomic RNA (Brierley et al., 1987). The two replicase polyproteins are processed extensively by two or three viral proteases encoded by ORF1a to generate up to 16 end-products termed nonstructural proteins (nsp) 1 to 16 and multiple processing intermediates (Ziebuhr, 2005; Ziebuhr et al., 2000). The N-proximal region of the polyproteins is processed by one or two papain-like proteases (PLpro), whereas the central and C-proximal region is processed...
by the viral main protease, 3C-like protease (3CLpro) (Ziebuhr, 2005; Ziebuhr et al., 2000). The 3′-proximal one-third of the genome encodes structural proteins and various accessory proteins. Genes encoding the four structural proteins present in all coronaviruses occur in the 5′ to 3′ order as spike (S), envelope (E), membrane (M), and nucleocapsid (N) proteins (Brian and Baric, 2005; Lai et al., 2006). Some coronaviruses contain an additional structural protein, the hemagglutinin–esterase (HE) protein which is located upstream of the S protein gene (Lai et al., 2006). In contrast to the replicase proteins which are directly translated from the genomic RNA, coronavirus structural and accessory proteins are expressed from a nested set of 3′ c-terminal subgenomic (sg) mRNAs that also possess a common 5′ leader sequence derived from the 5′ end of the genome (Pasternak et al., 2006; Sawicki et al., 2007). The common 5′ leader is fused to the 3′ body segments through a mechanism that is presumed to involve discontinuous minus strand RNA synthesis to produce subgenome-length templates for subgenomic mRNA synthesis, with the transcription regulatory sequence (TRS) elements determining the fusion sites of leader and body segments (see recent review of Pasternak et al., 2006; Sawicki et al., 2007 for details).

Equine coronavirus (ECoV) was first isolated from feces of a diarrheic foal in 1999 (ECoV-NC99) in North Carolina, USA (Guy et al., 2000). Little is known about ECoV and its clinical significance. Molecular characterization of ECoV and development of diagnostic and prophylactic reagents necessitate sequencing of ECoV. In this study, we determined the full-length nucleotide sequence of the ECoV-NC99 strain of equine coronavirus. The viral genome and proteome were analyzed and the predicted features of ECoV nonstructural, structural, and accessory proteins were compared to those of other coronaviruses. Synthesis of sg mRNAs in ECoV-infected cells was analyzed by Northern blotting. The leader–body junction sequence in each sg mRNA was determined and the exact position of TRS used for synthesis of each sg mRNA was mapped on the genome. The evolutionary relationship between ECoV and other phylogenetically closely related group 2a coronaviruses was explored.

Results and discussion

ECoV genome sequence analysis

We report here the full-length genomic sequence of the first ECoV isolate, the NC99 strain, and this is also the first reported complete genome sequence of ECoV. The nucleotide sequence was determined by directly sequencing 11 overlapping cDNA fragments which were RT–PCR amplified from viral RNA. The ECoV-NC99 genome comprises 30,992 nucleotides (nt), excluding the 3′ poly (A) tail, and has a GC content of 37.2%. The nucleotide sequence data have been deposited in GenBank under accession number EF446615.

Both 5′ and 3′ ends of the ECoV genome contain short untranslated regions (UTR). The 5′ UTR comprises 209 nt (1–209) and includes a potential short internal ORF of 8 codons (nt 99–125). Four stem–loop structures (I, II, III, and IV) were identified in the 5′ UTR and a short stretch of nucleotides that are part of the ORF1a (see Supplementary Fig. S1). The bulged stem–loop III (96–115) and IV (189–208) closely resemble the stem–loop III and IV that have been identified as replication signaling elements in bovine coronavirus and other group 2 coronaviruses (Raman and Brian, 2005; Raman et al., 2003; Wu et al., 2003). The 3′ UTR of the ECoV genome comprises 289 nt (30,704–30,992) and contains a putative bulged stem–loop structure (nt 30,703–30,770) and a putative pseudoknot structure (30,766–30,819) (see Supplementary Fig. S2). Similar putative bulged stem–loop structure and pseudoknot structure have been identified in murine hepatitis virus and other group 2 coronaviruses; these have been shown to be essential for viral replication (Goebel et al., 2004a,b; Hsu and Masters, 1997; Hsu et al., 2000; Williams et al., 1999).

Analysis of the ECoV-NC99 genome reveals 11 potential ORFs (1a, 1b, 2–8, 9a and 9b) as shown in Fig. 1 and Table 1. The ORFs 1a and 1b encode the replicase polyproteins pp1a and pp1ab. The ORFs 2–8, 9a and 9b encode structural and accessory proteins NS2, HE, S, p4.7, p12.7, E, M, N, and I, respectively.

The replicase ORF1a (nt 210–13,499) and replicase ORF1b (13,478–21,595) occupy 21.4 kb (69%) of the ECoV-NC99 genome. The translation of ORF1a generates a precursor pp1a of 4,429 amino acids. Similar to other coronaviruses, translation of ORF1b involves a –1 ribosomal frameshift, generating a 7128-amino acid pp1ab. The ribosomal frameshift is assumed to be directed by two signals in the ORF1a/1b overlapping region: a slippery sequence 5′UUUAAC3′ (nt 13,472–13,478) and a predicted downstream RNA pseudoknot structure (nt 13,484–13,559) (see Supplementary Fig. S3). The pp1a and pp1ab proteins are predicted to be proteolytically processed by viral-encoded proteases into 16 non-structural proteins (nsp1–16, Table 2) required for viral replication and transcription. By comparison to other coronaviruses, a number of putative functional domains are predicted in the ECoV pp1a and pp1ab and these are summarized in Fig. 1 and Table 2 (Gorbalenya et al., 1991, 2006; Snijder et al., 2003; Ziebuhr, 2005; Ziebuhr et al., 2001). Enzymatic activities of nsp3, nsp5, nsp12, nsp13, nsp14 and nsp15 have been experimentally confirmed for some coronaviruses (Barretto et al., 2005; Cheng et al., 2005; Guarino et al., 2005; Heusipp et al., 1997; Ivanov et al., 2004a,b; Ivanov and Ziebuhr, 2004; Lindner et al., 2005; Minskaia et al., 2006; Putics et al., 2005, 2006; Seybert et al., 2000, 2005; Tanner et al., 2003; Ziebuhr, 2005; Ziebuhr et al., 2001). The 3CLpro (catalytic residues His-3333 and Cys-3437) is predicted to cleave the C-terminal half of the ECoV pp1a and the ORF1b-encoded part of pp1ab. The putative PL1 (catalytic residues Cys-1078 and His-1229) and PL2 (catalytic residues Cys-1675 and His-1832) are predicted to process the N-proximal regions of the ECoV pp1a (Fig. 1 and Table 2).

The most striking differences between the ECoV replicase and other group 2 coronaviruses replicases were identified in ns3p. The ECoV ns3p protein has 3 aa deletions and 55 aa insertions compared to the ns3p proteins of BCoV, HCoV-OC43, and PHEV, three viruses phylogenetically most closely related to ECoV. These insertions and deletions are clustered at two
regions: the Ac domain and the region between the PL2\textsuperscript{pro} and the Y domain. The functional significance of these insertions and deletions is unknown as yet; however, the functions of PL1\textsuperscript{pro}, PL2\textsuperscript{pro}, and ADRP are not anticipated to be affected since insertions and deletions are not located in the functional domains of these enzymes (Fig. 1).

ORF2 (nt 21,610–22,446) of ECoV-NC99 encodes the predicted NS2 protein with 278 amino acids. The NS2 of
Table 2
Predicted end-products of proteolytic processing of the ECoV replicase polyproteins pp1a and pp1ab

| Cleavage product | Nucleotide positiona | Polyprotein | Position in pp1a/pp1ab (aa) | Length (aa) | Putative functional domain(s)b | Putative proteases predicted to release protein from polyproteins |
|------------------|----------------------|-------------|-----------------------------|-------------|-------------------------------|---------------------------------------------------------------|
| nsp1             | 210–941              | pp1a/pp1ab  | 1Met-Gly244                  | 244         |                               | PL1pro                                                       |
| nsp2             | 942–2744             | pp1a/pp1ab  | 245Val-Ala845                | 601         |                               | PL1pro, ADRP, PL2pro                                          |
| nsp3             | 2745–8597            | pp1a/pp1ab  | 846Gly-Gly2796               | 1951        |                               | PL2pro                                                       |
| nsp4             | 8598–10,085          | pp1a/pp1ab  | 2797Ala-Gln3292              | 496         |                               | TM2, PL2pro, 3CLpro                                          |
| nsp5             | 10,086–10,994        | pp1a/pp1ab  | 3293Ser-Gln3595              | 303         |                               | 3CLpro                                                       |
| nsp6             | 10,995–11,855        | pp1a/pp1ab  | 3596Ser-Gln3882              | 287         |                               | 3CLpro                                                       |
| nsp7             | 11,856–12,122        | pp1a/pp1ab  | 3883Ser-Gln3971              | 89          | Part of RNA binding            | 3CLpro                                                       |
| nsp8             | 12,123–12,713        | pp1a/pp1ab  | 3972Ala-Gln4168              | 197         | Part of RNA binding            | 3CLpro                                                       |
| nsp9             | 12,714–13,043        | pp1a/pp1ab  | 4169Asn-Gln4278              | 110         | ssRNA-binding protein          | 3CLpro                                                       |
| nsp10            | 13,044–13,454        | pp1a/pp1ab  | 4279Ala-Gln4415              | 137         | 2 zinc fingers                 | 3CLpro, 3CLpro                                               |
| nsp11            | 13,455–13,496        | pp1a        | 4416Ser-Ser4429              | 14          |                               | 3CLpro                                                       |
| nsp12            | 13,455–16,237        | pp1b        | 4416Ser-Gln5343              | 928         | RdRp                          | 3CLpro                                                       |
| nsp13            | 16,238–18,034        | pp1b        | 5344Ser-Gln5942              | 599         | ZBD, HEL                      | 3CLpro                                                       |
| nsp14            | 18,035–19,597        | pp1b        | 5943Cys-Gln6463              | 521         | Exonuclease (ExoN)            | 3CLpro                                                       |
| nsp15            | 19,598–20,695        | pp1b        | 6464Ser-Gln6829              | 366         | NendoU                        | 3CLpro                                                       |
| nsp16            | 20,696–21,592        | pp1b        | 6830Ala-Ile7128              | 299         | 2′-O-MT                       | 3CLpro                                                       |

Domains Ac and Y are described by Ziebuhr et al. (2001).
a Nucleotide position means the location of the nucleotides encoding corresponding proteins in the entire genome of equine coronavirus-NC99 strain.
b Putative proteases predicted to release protein from polyproteins.

ECoV has 67%, 67%, and 45% amino acid identity with the respective NS2 proteins of BCoV, HCoV-OC43, and PHEV. The lower amino acid identity with PHEV may be attributable to the fact that PHEV has a truncated NS2 protein (Vijgen et al., 2006). Sequence analysis revealed that the ECoV NS2 protein contains a domain (aa 46–135) with similarity to the putative cyclic phosphodiesterase (CPD, Martzen et al., 1999). The CPD domain has also been identified in the NS2 proteins of other group 2a coronaviruses as well as in the 3′ end of the pp1a protein of toroviruses (Gorbalenya et al., 2006; Snijder et al., 1991, 1993). The NS2 of ECoV was predicted to contain 9 potential phosphorylation sites. The NS2 of ECoV does not contain a signal peptide and is a non-secretory protein. The function of the NS2 protein in coronaviruses has not been studied in detail. It is known that the NS2 gene is non-essential for MHV replication in continuous cell culture (Sperry et al., 2005). ORF3 (nt 22,458–23,729) of ECoV-NC99 encodes the predicted spike (S) protein containing 1363 amino acids. Eighteen potential N-glycosylation sites were predicted. An N-terminal signal peptide was identified with a potential cleavage site between amino acids 14 and 15 predicted by SignalP-NN or between amino acids 17 and 18 predicted by SignalP-HMM. The ECoV S protein was predicted to be a typical type I membrane protein with the N-terminal 1307 residues exposed on the outside of the cell surface or virus particle, a transmembrane domain near the C terminus (residues 1308–1330), followed by a cytoplasmic tail (residues 1331–1363). Following multiple alignments with the S proteins of other group 2a coronaviruses, a potential cleavage recognition sequence (RRQKR) was identified at residues 764–768 which would predict a cleavage between amino acids 768 and 769, separating the ECoV S protein into S1 and S2 subunits (Fig. 1). The ECoV S1 subunit is expected to contain a receptor-binding domain whose position has not yet been determined. The S2 subunit is predicted to mediate membrane fusion. Two heptad repeat (HR) regions, which are conserved in position and sequence among the three groups of coronaviruses and play important roles in membrane fusion (see reviews of Eckert and Kim, 2001; Hernandez et al., 1997), were identified in the ECoV S2 subunit (HR1: aa 991–1092; HR2: aa 1259–1304) (Fig. 1). The ECoV S2 subunit is anticipated to possess a fusion peptide whose position is yet unknown. Some coronavirus S proteins have been shown to contain important neutralization epitopes (Godet et al., 1994; Kubo et al., 1994; Yoo et al., 1991) and mutations in the S protein have been associated with altered viral antigenicity and pathogenicity (Ballesteros et al., 1997; Bernard...
and Laude, 1995; Dalziel et al., 1986; Gallagher and Buchmeier, 2001; Leparc-Goffart et al., 1997). Whether the S protein of ECoV has such properties remains to be determined.

ORF5 (nt 27,825–27,947) of ECoV-NC99 is predicted to encode a hypothetical protein of 40 amino acids with an estimated molecular weight of 4.7 kDa (termed p4.7 protein). It was predicted to be a non-secretory protein and did not contain any transmembrane helix. This protein is not closely matched to any known protein based on a search using BLASTP, PSI-BLAST, or FASTA programs.

ORF6 (nt 28,076–28,405) of ECoV-NC99 is predicted to encode a protein of 109 amino acids corresponding to the BCoV 12.7 kDa non-structural protein (p12.7). This ORF overlaps by 15 nucleotides with the ORF7 that encodes the E protein. No N-glycosylation site was found.

ORF7 (nt 28,392–28,646) of ECoV-NC99 encodes the predicted E protein containing 84 amino acids. No N-glycosylation site was identified. It was predicted to contain a signal anchor (probability 0.999). One transmembrane domain was predicted at residues 18–36 by TMpred analysis or at residues 15–37 by TMHMM analysis. Both programs predicted the N-terminus of the protein to be external to the cell surface or viral envelope. In the case of other coronaviruses, there is increasing evidence that the E protein together with the M protein is instrumental in viral assembly and budding; the cytoplasmic tails of both proteins have an important interactive role in this process (Corse and Machamer, 2000, 2002, 2003; Vennema et al., 1996).

ORF8 (nt 28,661–29,353) of ECoV-NC99 encodes the predicted M protein containing 230 amino acids. It was predicted to contain a signal anchor (probability 0.947). Three transmembrane domains were predicted to be present at positions 25–46, 57–78, and 81–102 by TMpred analysis or at positions 25–44, 49–71, and 81–103 by TMHMM analysis. The N-terminal 24 amino acid residues were predicted to be outside and the C-terminal 127 or 128-amino acid hydrophilic domain was predicted to be inside the virus. One potential N-glycosylation site was predicted at position 26 (NFS). The presence of potential O-glycosylation sites was predicted at the extreme N-terminus of the M protein (MSSTPTPAPGYT). Whether these sites are glycosylated or not needs to be experimentally verified. Previous studies have shown that the M protein of group 1 and 3 coronaviruses (e.g. TGEV and IBV) are N-glycosylated, whereas the M protein of group 2 coronavirus MHV is only O-glycosylated (de Haan et al., 2002; Lai et al., 2006). The M protein is the most abundant envelope component and plays a key role in coronavirus assembly by interacting with the E, S, N and HE proteins (Bosch et al., 2005; de Haan and Rottier, 2005, and references therein).

---

**Fig. 2.** Northern blot analysis of intracellular RNA isolated from ECoV-infected HRT-18G cells. A DIG-labeled probe which was complementary to the 3′ end (nt 30,660–30,946) of ECoV genome was used to detect the genomic and subgenomic mRNAs in ECoV-infected (lane 2) and mock-infected (lane 1) HRT-18G cells at 72 h p.i.

---

**Table 3**

| Primer ID | Position  | Sequence (5′–3′) | Use |
|-----------|-----------|------------------|-----|
| 22813N    | 22,792–22,813 | GCGTTATCAACCAGAAGCGGTGC | Reverse transcription for mRNA2 (NS2) and reverse primer for mRNA3 (HE) PCR |
| 25095N    | 25,076–25,095 | CGCCTATTCCAGGCAGAAGG | Reverse transcription for mRNA3 (HE) and mRNA4 (S) PCR |
| 29010N    | 29,078–29,101 | GGCAGTAAAAGTTATGATGTTCCCTC | Reverse transcription for mRNA5 (p4.7), mRNA6 (p12.7) and mRNA7 (E) PCR |
| 30945N    | 30,921–30,945 | CTGGGTGTTAATCTAATCATGCTTGCC | Reverse transcription for mRNA8 (M) and mRNA9 (N) PCR |
| 1P        | 1–21     | GATTGTGAGCGAATTGCGTGC | Forward primer for all sg mRNA PCR |
| 21982N    | 21,958–21,982 | GACGCGGACTGACCAACTAACAACC | Reverse primer for mRNA2 (NS2) PCR |
| 24283N    | 24,262–24,283 | GCGTTGCGGACCAATACCACTAGG | Reverse primer for mRNA4 (S) PCR |
| 28100N    | 28,078–28,100 | TCCCTCAGGTCTCCACAGATGCTC | Reverse primer for mRNA5 (p4.7) PCR |
| 28334N    | 28,312–28,331 | CAGCCTTCTTATATGATGTTGCC | Reverse primer for mRNA6 (p12.7) PCR |
| 28641N    | 28,617–28,641 | CGTCATCCACATTAAAGGACTGTTGG | Reverse primer for mRNA7 (E) PCR |
| 29016N    | 28,992–29,016 | GGTTGAAATCACTCAACACTCA | Reverse primer for mRNA8 (M) PCR |
| 29710N    | 29,691–29,710 | GCGTTGATTGCCATCGGCTG | Reverse primer for mRNA9 (N) PCR |
ORF9a (nt 29,363–30,703) of ECoV-NC99 encodes the predicted N protein containing 446 amino acids. It was predicted to contain 36 potential phosphorylation sites. No signal peptide or any transmembrane helix was present. The N protein of coronaviruses has been shown to be multifunctional, e.g. interaction with the viral RNA genome to form a viral nucleocapsid, interaction with the M protein, and the ability for self-association (Masters, 1992; Narayanan et al., 2000, 2003). Recently it has also been reported that the N protein may play a role in coronavirus replication (Almazan et al., 2004; Schelle et al., 2005).

ORF9b (nt 29,424–30,044) of ECoV-NC99 encodes a hypothetical protein (I) containing 206 amino acids within ORF9a which encodes the N protein. It was predicted to contain 10 potential phosphorylation sites. No signal peptide or any transmembrane helix was present. In the case of MHV, expression of the protein I has been detected in virus-infected cells but this protein is nonessential for viral replication and viral production (Fischer et al., 1997).

Northern blot analysis of ECoV genomic and subgenomic mRNAs

It is generally accepted that the replicase proteins are directly synthesized from the coronavirus genome, whereas the structural and accessory proteins are expressed from a nested set of subgenomic mRNAs. However, the number of sg mRNAs and the characteristics and expression pattern of the
proteins they encode (e.g. a sg mRNA may sometimes express multiple proteins) varies for each virus. In order to investigate ECoV sg mRNA synthesis, Northern blot analysis was performed to evaluate the synthesis of genomic and subgenomic RNAs in ECoV-infected cells. A digoxigenin-labeled RNA probe complementary to the 3′ end (nt 30,660–30,946) of the ECoV genome was used for a Northern blot hybridization analysis. As shown in Fig. 2, nine mRNAs were detected in ECoV-infected HRT-18G cells at 72 h p.i. Absence of such mRNAs in mock-infected cells confirms that these mRNAs are ECoV-specific. According to the estimated sizes of the mRNAs, it is reasonable to assume that sg mRNAs 2–8 express the NS2, HE, S, p4.7, p12.7, E, and M proteins, respectively and that mRNA 9 expresses the N protein and probably the I protein as well.

**Determination of leader–body junction sequences of sg mRNAs**

There is a general agreement that the TRS elements determine the fusion sites of the 5′ leader and the 3′ body segments in coronavirus sg mRNAs. In order to determine the precise location of the leader and body TRSs used for ECoV sg mRNA synthesis, the leader–body junction and flanking sequences of each ECoV sg mRNA were determined using sg mRNA-specific RT–PCRs (see Table 3 and Materials and methods for details). The sg mRNA sequences were aligned to the leader and corresponding ‘body’ genomes as shown in Fig. 3. Analysis of the leader–body junction sequences revealed that the core sequence of the TRS motifs is 5′UCUAAAC′3′. The leader TRS (5′UCUAAC3′) and the body TRS (5′UCUAAC3′) used for synthesizing HE mRNA, S mRNA, and N mRNA exactly match each other. There is one mismatch between the leader TRS and the body TRS (5′UCUAAC3′) used for generating the mRNA of the NS2 protein. There is also one mismatch between the leader TRS and the body TRS (5′UCUAAC3′) used for generating E mRNA and M mRNA. There are two mismatches between the leader TRS and the body TRS (5′UCUAAC3′) used for generating the mRNA of the p4.7 protein. Interestingly, in the case of the mRNA of the p12.7 protein, the leader and the body segment is joined at the unusual consensus variant 5′UAAA-CUUAUAAC′3′. Previously it has been shown that the mRNA of the p12.7 protein of BCoV also utilizes an unusual consensus variant for joining the leader and body segment (Hofmann et al., 1993). From the sequence data, we conclude that the ECoV common leader on sg mRNAs is the first 64 nucleotides of the ECoV genome.

**Phylogenetic analysis of ECoV**

Phylogenetic analyses of ECoV and other coronaviruses were performed based on the amino acid sequences of replicase polyprotein pp1a, the ORF1b-encoded part of the pp1ab, S, E, M, and N. Phylogenetic analysis clustered coronaviruses into three major groups (G1, G2a, and G3) irrespective of the gene used for analysis (Fig. 4). The SARS-CoV forms a separate branch and is classified as subgroup 2b (G2b) as suggested previously (Gorbalenya et al., 2004; Snijder et al., 2003). Phylogenetic analysis clearly demonstrated that ECoV falls into the cluster of group 2a coronaviruses and is most closely related to BCoV, HCoV-OC43, and PHEV.

To further explore the possible evolutionary relationships among ECoV, BCoV, HCoV-OC43, and PHEV, the genetic distances of ECoV, BCoV, and PHEV to HCoV-OC43 were determined over the entire genome using the SimPlot analysis (Lole et al., 1999). As shown in Fig. 5, the BCoV strains and HCoV-OC43 had lowest genetic distances over the complete genome; the genetic distance between PHEV and HCoV-OC43 was similar to the distance between BCoV and HCoV-OC43 in most regions of the genome with exception of the spike gene where the genetic distance of PHEV to HCoV-OC43 was significantly greater than the distance of either BCoV to HCoV-OC43; the genetic distance of ECoV to HCoV-OC43 was significantly greater than the distance of either BCoV or PHEV to HCoV-OC43 in the regions of the first half of ORF1a, the central part of ORF1b, NS2 and HE genes; the genetic distance with respect to the spike gene between ECoV and HCoV-OC43 was similar to the distance between PHEV and HCoV-OC43 but greatly higher than the distance between BCoV and HCoV-OC43. The genetic distances of BCoV and PHEV to HCoV-OC43 observed in this study are consistent with previously reported findings (Vijgen et al., 2005, 2006). Vijgen et al. (2006, 2005) concluded that PHEV diverged from the common ancestor before BCoV and HCoV-OC43. Our analysis suggested that ECoV had diverged earlier than PHEV from a common ancestor. In summary, ECoV had emerged earlier than PHEV, BCoV, and HCoV-OC43, notwithstanding the fact that ECoV was not isolated until 1999 from a diarrheic foal in USA.

**Conclusion**

In this study, we have determined the first complete genome sequence of ECoV and provided the first comprehensive analysis of the ECoV genome. Completion of the genome sequence of
ECoV will contribute to our understanding of this virus at the molecular level and also enrich the database of coronaviruses. The sequence data are expected to aid in the development of diagnostic and prophylactic reagents. The sequence data of ECoV-NC99 will also help identify and characterize other ECoV isolates and enhance our understanding of the molecular epidemiology of coronavirus. Neonatal enterocolitis is an economically significant disease for horse breeders. Further studies are needed to determine the prevalence of ECoV infection in equine populations and the relative role of ECoV as a cause of enteric disease in horses.

Materials and methods

Cells and virus

The human rectal tumor cell line HRT-18G (American Type Culture Collection [ATCC, CRL-11663]) was grown in Dulbecco’s modified Eagle’s medium (DMEM) supplemented with 4 mM L-glutamine, 5% fetal bovine serum, and penicillin/streptomycin at 37 °C in the presence of 5% CO2. The equine coronavirus-NC99 (Guy et al., 2000) was propagated once in HRT-18G cells to produce the working virus stocks.

Isolation of viral RNA, RT–PCR amplification and sequencing

The complete genome of ECoV was determined by sequencing 11 overlapping RT–PCR products encompassing the entire genome (nt 1–3615; nt 3446–5458; nt 4953–6600; nt 5497–9678; nt 9347–13,021; nt 12,451–15,736; nt 15,425–19,307; nt 19,039–22,812; nt 22,566–26,390; nt 26,065–29,662; and nt 29,363–30,992). Viral RNA was isolated from ECoV stocks using the QIAamp viral RNA mini kit (Qiagen). Viral RNA was first reverse transcribed with AccuScript reverse transcriptase (Stratagene) following the manufacturer’s instructions. Then, PCR amplification was performed with proof-reading PfuUltra high-fidelity DNA polymerase (Stratagene) in a volume of 50 μl: 5 μl PfuUltra PCR buffer (10×), 1.0 μl dNTP mix (10 mM each), 1 μl of each primer (20 μM), 2 μl cDNA template, 1 μl PfuUltra DNA polymerase, and 39.0 μl nuclease-free water. The reaction mixtures were incubated at 95 °C for 2 min, followed by 35 cycles of amplification at 95 °C for 45 s, 50–53 °C for 45 s, and 72 °C for 4.5 min, with a final incubation at 72 °C for 10 min. The PCR products were gel-purified using QIAquick gel extraction kit (Qiagen). Both sense and anti-sense strands were sequenced using the Applied Biosystems Big Dye Terminator V3.0 sequencing chemistry on ABI 3730 DNA sequencers (Davis Sequencing Center). Partial genomic sequence (9487 nucleotides) of ECoV had
been previously determined by two groups (Guy et al., 2000, GenBank accession number AF251144; Wu et al., 2003, AF523846 and AF523850. H.Y. Wu, J.S. Guy, and D.A. Brian, unpublished data, AY316300). These regions were re-sequenced in this study. To determine the remaining genomic sequence of ECoV-NC99, initial RT–PCR and sequencing primers were designed based on multiple alignments of the genomes of BCov (GenBank accession number NC_003045), HCoV-OC43 (NC_001846), and MHV (NC_001846); additional primers were designed based on the results of the first and subsequent rounds of sequencing. All of the primer sequences are attached in the Supplementary Table S1.

DNA and protein sequence analysis

The nucleotide sequences were assembled and manually edited using CodonCode Aligner version 1.5.2 to produce the complete sequence of the viral genome. ORF analysis was performed using Vector NTI Advance 10 (Invitrogen). RNA secondary structures of 5' and 3' UTRs and the ribosomal frameshift signals were predicted using the Mfold program with the default parameter settings (Mathews et al., 1999; Zuker, 2003). Potential 3C-like protease cleavage sites were predicted using the default parameter settings (Mathews et al., 1999; Zuker, 2003). Potential N-glycosylation of signal peptides and their cleavage sites was conducted using the NetCorona 1.0 server (Kiemer et al., 2004). Prediction 2003). Potential 3C-like protease cleavage sites were predicted with the default parameter settings (Mathews et al., 1999; Zuker, 2003). Frameshift signals were predicted using the MFOLD program (Zuker, 2003).

Analysis of viral RNA by Northern blotting

One anti-sense RNA probe base pairing to the 3' end of the ECoV genome (nt 30,660–30,946) was developed to evaluate the synthesis of genomic and subgenomic RNAs in ECoV-infected cells by Northern blotting. The ECoV RNA was amplified using two primer pairs (forward primer 30660P: 5'AGCAGATGATGGATC3'; reverse primer 30946N: 5'ACTGGGTGGTAACTTACATGG3') and the QIagen One-step RT–PCR kit (Qiagen). The gel-purified RT–PCR products were cloned into a linearized plasmid vector with overhanging 3'T residues (pDrive Cloning Vector, Qiagen). The authenticity and orientation of the insert was determined by sequencing both strands of DNA with M13 reverse and forward primers. Plasmid DNA was linearized with BamHI (Roche), phenol/chloroform extracted, ethanol precipitated, and resuspended in nuclease-free water. A digoxigenin (DIG)-labeled RNA probe was prepared using the DIG RNA labeling kit (Roche) according to the manufacturer’s instructions.

Intracellular RNA was extracted at 72 h p.i. from ECoV-infected HRT-18G cells using the RNaqueous-4PCR kit (Ambion). Northern hybridization with the DIG-labeled RNA probe was carried out following the protocols that had been previously described for equine arteritis virus (Balasuriya et al., 2004).

Determination of the leader–body junction sequence

The leader–body junction sites of all ECoV sg mRNAs were RT–PCR amplified and sequenced. Briefly, intracellular RNA was extracted from ECoV-infected HRT-18G cells using the RNaqueous-4PCR kit (Ambion). Reverse transcription was carried out with an RT primer located downstream to the body TRS region in a sg mRNA (Table 3) using SuperScriptIII reverse transcriptase (Invitrogen) following the manufacturer’s instructions. Due to the nested nature of sg mRNAs, such an RT primer also binds to the corresponding positions in all larger viral mRNAs, including the genomic RNA. Subsequently, cDNA was PCR amplified with a forward primer (1P) located in the leader sequence and a reverse primer located just upstream of the RT primer in the body of the mRNA (Table 3). Amplification was performed in a volume of 50 μl: 5 μl PfuTurbo PCR buffer (10×), 0.4 μl dNTP mix (25 mM each), 1 μl of each primer (20 μM), 2 μl cDNA template, 1 μl PfuTurbo® DNA polymerase, and 39.6 μl nuclease-free water. The reaction mixtures were incubated at 95 °C for 2 min, followed by 35 cycles at 95 °C for 45 s, 50–56 °C for 45 s, and 72 °C for 3 min, with a final incubation at 72 °C for 10 min. RT–PCR products corresponding to each mRNA species could be distinguished by size differences on agarose gel. PCR products were gel-purified and sequenced to obtain the leader–body junction sequences for each sg mRNA.

Nucleotide sequence accession number

The nucleotide sequence of ECoV was deposited in GenBank under the accession number EF446615.

Acknowledgments

This work was partly supported by funds from Fort Dodge Animal Health and Kentucky Agricultural Experiment Station, College of Agriculture, University of Kentucky.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.virol.2007.06.035.
human coronavirus OC43: molecular clock analysis suggests a relatively recent zoonotic coronavirus transmission event. J. Virol. 79 (3), 1595–1604.

Vijgen, L., Keyaerts, E., Lemey, P., Maes, P., Van Reeth, K., Nauwynck, H., Pensaert, M., Van Ranst, M., 2006. Evolutionary history of the closely related group 2 coronaviruses: porcine hemagglutinating encephalomyelitis virus, bovine coronavirus, and human coronavirus OC43. J. Virol. 80 (14), 7270–7274.

Williams, G.D., Chang, R.Y., Brian, D.A., 1999. A phylogenetically conserved hairpin-type 3’ untranslated region pseudoknot functions in coronavirus RNA replication. J. Virol. 73 (10), 8349–8355.

Woo, P.C., Lau, S.K., Chu, C.M., Chan, K.H., Tsoi, H.W., Huang, Y., Wong, B.H., Poon, R.W., Cai, J.J., Luk, W.K., Poon, L.L., Wong, S.S., Guan, Y., Peiris, J.S., Yuen, K.Y., 2005. Characterization and complete genome sequence of a novel coronavirus, coronavirus HKU1, from patients with pneumonia. J. Virol. 79 (2), 884–895.

Wu, H.Y., Guy, J.S., Yoo, D., Vlasak, R., Urbach, E., Brian, D.A., 2003. Common RNA replication signals exist among group 2 coronaviruses: evidence for in vivo recombination between animal and human coronavirus molecules. Virology 315 (1), 174–183.

Yoo, D.W., Parker, M.D., Song, J., Cox, G.J., Deregt, D., Babiuk, L.A., 1991. Structural analysis of the conformational domains involved in neutralization of bovine coronavirus using deletion mutants of the spike glycoprotein S1 subunit expressed by recombinant baculoviruses. Virology 183 (1), 91–98.

Ziebuhr, J., 2005. The coronavirus replicate. Curr. Top Microbiol. Immunol. 287, 57–94.

Ziebuhr, J., Snijder, E.J., Gorbalenya, A.E., 2000. Virus-encoded proteinases and proteolytic processing in the Nidovirales. J. Gen. Virol. 81 (Pt. 4), 853–879.

Ziebuhr, J., Thiel, V., Gorbalenya, A.E., 2001. The autocatalytic release of a putative RNA virus transcription factor from its polyprotein precursor involves two paralogous papain-like proteases that cleave the same peptide bond. J. Biol. Chem. 276 (35), 33220–33232.

Zuker, M., 2003. Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res. 31 (13), 3406–3415.