Elucidation of the Genomic Nucleotide Sequence of Bovine Coronavirus and Analysis of Cryptic Leader mRNA Fusion Sites

VLADIMIR N. CHOULJENKO, TIMOTHY P. FOSTER, XIAOQING LIN, JOHANNES STORZ, AND KONSTANTIN G. KOUSOULAS
Department of Veterinary Microbiology and Parasitology, School of Veterinary Medicine,Louisiana State University, Baton Rouge, LA 70803

1.   INTRODUCTION

Over the last few years numerous respiratory bovine coronaviruses (RBCV) were isolated from respiratory infections of cattle including two shipping fever epizootics (Storz *et al.*, 1996; Storz *et al.*, 2000). The RBCV strain 98TXSF-110-LU (LU) and the enteric strain 98TXSF-110-ENT (ENT) were isolated from lung and intestinal samples, respectively, from an animal experiencing fatal respiratory disease during a 1998 shipping fever epizootic. We obtained and compared the entire genomic sequences of the LU and ENT isolates to delineate nucleotide (nt) and amino acid (aa) differences between them. In addition, we analyzed the leader-mRNA junction sequences of the S and 12.7 kDa mRNAs specified by different BCV strains.

2.   MATERIALS AND METHODS

The LU and ENT BCV strains were plaque purified, gradient purified, total RNA was extracted from infected HRT-18G cells, and cDNA was produced as described previously (Chouljenko *et al.*, 1998). Primers for sequencing of unknown BCV RNA polymerase were designed based on the

The Nidoviruses (Coronaviruses and Arteriviruses).
Edited by Ehud Lavi *et al.*, Kluwer Academic/Plenum Publishers, 2001.

49

sequence of MHV-A59 c12 mutant (accession number AF 029248). To characterize the leader-mRNA junction sequences of the S and 12.7 kDa mRNAs, individual junctions were amplified by RT-PCR, cloned into plasmid pCR 2.1 (Invitrogen, Inc, Carlsbad, CA) and sequenced. Primer LBCV annealing to the leader sequence in combination with primers 3B21, 3B11, and 3B26 annealing to coding sequences of S, were used to produce PCR products containing the putative leader-body mRNA junctions (Chouljenko and Kousoulas, 2000).

3.   RESULTS AND DISCUSSION

3.1   Genomic sequences of ENT and LU strains

Sequence comparison between the intestinal and lung isolates derived from the same animal revealed that these viruses contained numerous single nucleotide substitutions. In general, the structure of the BCV RNA polymerase gene was found to be similar to the other coronaviruses. ORF 1a overlapped by 31 nt with ORF 1b located in the −1 reading frame with respect to the upstream ORF 1a. A total of 32 nt changes, 5 of which caused aa changes were detected within the predicted ORF 1a cDNA sequence between the LU and ENT isolates. Fifteen nt changes were detected within the ORF 1b and none caused aa changes. Fifty-three nucleotide changes were detected between the LU and ENT strains, 17 of which caused amino acid changes located within a 9.5 kb sequence spanning all the non-polymerase ORFs. Thus, genomic sequencing revealed that two genetically different BCV strains were isolated from lung and intestinal samples of the same animal.

3.2   Intergenic leader-mRNA junctions of S mRNA specified by different BCV strains

Thirty cloned DNA fragments containing leader-body junction sequences of the S mRNA were obtained after RT-PCR using primer pair LBCV/3B21. The majority of these mRNA junctions utilized the canonical consensus sequence located within the 5′ UTR of S. Specifically, 29 of 30 clones contained the consensus intergenic sequence TCTAAAC as part of the S leader-mRNA junction (type 1, Fig. 1). One clone contained the noncanonical leader-body sequence TA located downstream from the initiation codon of S at nucleotide positions 63-64 (type 2; Fig. 1).
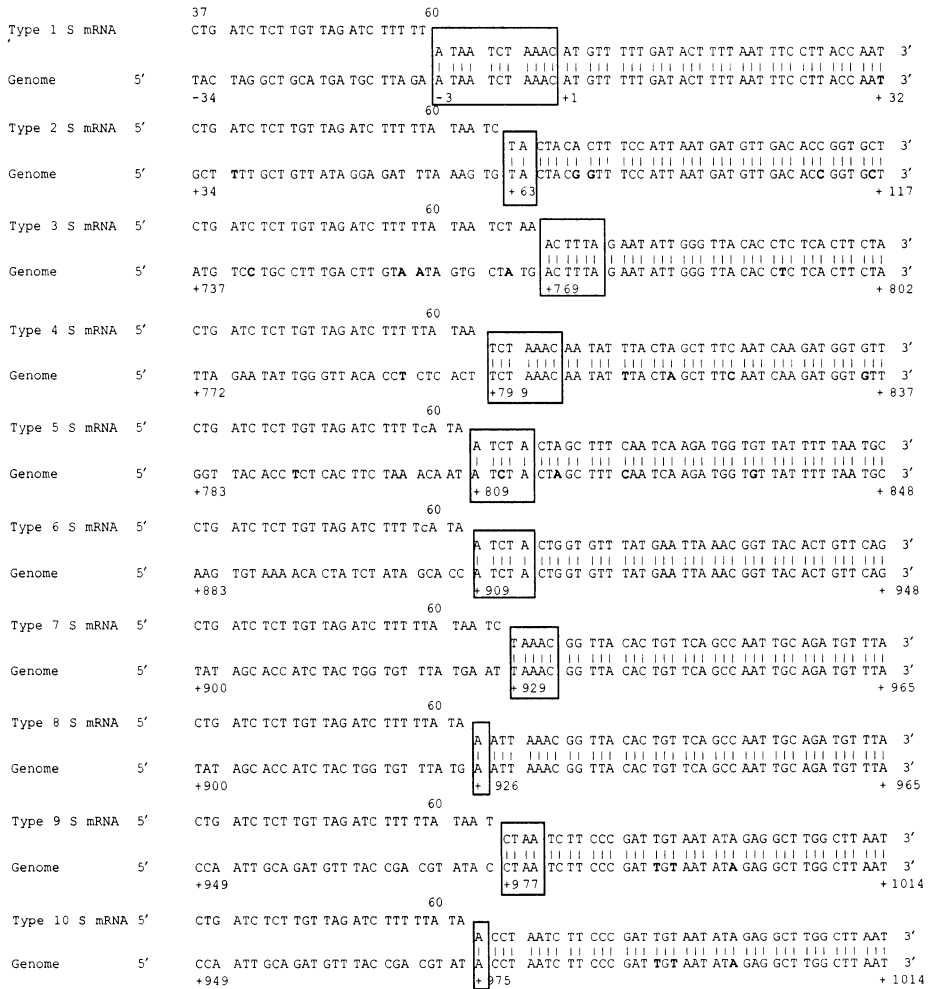
```
                    37                         60
Type 1 S mRNA          CTG ATC TCT TGT TAG ATC TTT TT
                                                      A TAA  TCT  AAAC AT GTT  TTT GAT ACT TTT AAT TTC CTT ACC AAT   3'
                                                      | | |    | | |   | | | | | | | | | | | | | | | | | | | | | | | | | |
Genome            5'   TAC  TAG GCT GCA TGA TGC TTA GA A TAA  TCT  AAAC AT GTT  TTT GAT ACT TTT AAT TTC CTT ACC AAT   3'
                      -34                              -3        +1                                            + 32
                                                           60
Type 2 S mRNA     5'   CTG ATC TCT TGT TAG ATC TTT TTA  TAA TC
                                                         T A CTA CA CTT  TCC ATT AAT GAT GTT GAC ACC GGT GCT   3'
                                                         | |  | | | | | | | |   | | | | | | | | | | | | | | | | | | | | |
Genome            5'   GCT  TTT GCT GTT ATA GGA GAT TTA AAG TG T A CTA CG GTT  TCC ATT AAT GAT GTT GAC ACC GGT GCT   3'
                      +34                               +63                                             + 117
                                                  60
Type 3 S mRNA     5'   CTG ATC TCT TGT TAG ATC TTT TTA  TAA  TCT AA
                                                         AC TT TA G AAT ATT GGG TTA CAC CTC TCA CTT CTA   3'
                                                         | | | | | |  | | | | | | | | | | | | | | | | | | | | | | |
Genome            5'   ATG  TCC TGC CTT TGA CTT GTA ATA GTG  CTA TG AC TT TA G AAT ATT GGG TTA CAC CTC TCA CTT CTA   3'
                      +737                               +769                                           + 802
                                                   60
Type 4 S mRNA     5'   CTG ATC TCT TGT TAG ATC TTT TTA  TAA
                                                         TCT AAAC AA TAT TTA CTA GCT TTC AAT CAA GAT GGT GTT   3'
                                                         | | |   | | | | | | | | | | | | | | | | | | | | | | | | | | | |
Genome            5'   TTA  GAA TAT TGG GTT ACA CCT CTC ACT TCT AAAC AA TAT TTA CTA GCT TTC AAT CAA GAT GGT GTT   3'
                      +772                               +799                                           + 837
                                                   60
Type 5 S mRNA     5'   CTG ATC TCT TGT TAG ATC TTT TcA  TA
                                                        A TCT A CTAGC TTT CAA TCA AGA TGG TGT TAT TTT TAA TGC   3'
                                                        | | | |  | | | | | | | | | | | | | | | | | | | | | | | | | |
Genome            5'   GGT  TAC ACC TCT CAC TTC TAA ACA AT A TCT A CTAGC TTT CAA TCA AGA TGG TGT TAT TTT TAA TGC   3'
                      +783                              +809                                            + 848
                                                   60
Type 6 S mRNA     5'   CTG ATC TCT TGT TAG ATC TTT TcA  TA
                                                        A TCT A CTG GT GTT TAT GAA TTA AAC GGT TAC ACT GTT CAG   3'
                                                        | | | |  | | | | | | | | | | | | | | | | | | | | | | | | |
Genome            5'   AAG  TGT AAA ACA CTA TCT ATA GCA CC A TCT A CTG GT GTT TAT GAA TTA AAC GGT TAC ACT GTT CAG   3'
                      +883                              +909                                            + 948
                                                   60
Type 7 S mRNA     5'   CTG ATC TCT TGT TAG ATC TTT TTA  TAA TC
                                                        T AAAC GG TTA CAC TGT TCA GCC AAT TGC AGA TGT TTA   3'
                                                        | | | | |  | | | | | | | | | | | | | | | | | | | | | | |
Genome            5'   TAT  AGC ACC ATC TAC TGG TGT TTA TGA AT T AAAC GG TTA CAC TGT TCA GCC AAT TGC AGA TGT TTA   3'
                      +900                              +929                                            + 965
                                                   60
Type 8 S mRNA     5'   CTG ATC TCT TGT TAG ATC TTT TTA  TA
                                                       A ATT AAAC GG TTA CAC TGT TCA GCC AAT TGC AGA TGT TTA   3'
                                                       |  | | | |  | | | | | | | | | | | | | | | | | | | | | | |
Genome            5'   TAT  AGC ACC ATC TAC TGG TGT TTA TG A ATT AAAC GG TTA CAC TGT TCA GCC AAT TGC AGA TGT TTA   3'
                      +900                              +926                                            + 965
                                                   60
Type 9 S mRNA     5'   CTG ATC TCT TGT TAG ATC TTT TTA  TAA T
                                                        CT AA TC TT CCC GAT TGT AAT ATA GAG GCT TGG CTT AAT   3'
                                                        | | | |  | | | | | | | | | | | | | | | | | | | | | | | | |
Genome            5'   CCA  ATT GCA GAT GTT TAC CGA CGT ATA C CT AA TC TT CCC GAT TGT AAT ATA GAG GCT TGG CTT AAT   3'
                      +949                              +977                                            + 1014
                                                   60
Type 10 S mRNA    5'   CTG ATC TCT TGT TAG ATC TTT TTA  TA
                                                       A CCT AATC TT CCC GAT TGT AAT ATA GAG GCT TGG CTT AAT   3'
                                                       |  | | | |  | | | | | | | | | | | | | | | | | | | | | | |
Genome            5'   CCA  ATT GCA GAT GTT TAC CGA CGT AT A CCT AATC TT CCC GAT TGT AAT ATA GAG GCT TGG CTT AAT   3'
                      +949                              +975                                            + 1014
```

Figure 1. **Cryptic leader-mRNA junction and flanking sequences of different mRNAs detected within the 5′ end of S.** Leader and genome sequences are shown for the EBCV strain L9 except types 5 and 6 detected only for the RBCV OK strain (Chouljenko *et al.*, 1998). Nucleotide sequences of subgenomic RNAs (sgRNAs) are shown in alignment with the genome sequence. Numbers above each sgRNA correspond to the leader sequence. Numbering on the genome sequence begins from the +1 (base A of the start codon of S). The putative leader-body sequences are boxed. Numbers inside the boxes denote the first nt of the putative junction sequences counting from the start codon of S. Nucleotide variations among different BCV strains noted previously (Chouljenko *et al.*, 1998; Zhang *et al.*, 1991) are shown as bold letters. The RBCV-specific change within the leader sequence is shown with a small letter.

3.3   Cryptic leader-mRNA junctions sequences within the 5′ end of the S ORF

The S ORF contains three heptameric nucleotide sequences that resemble the consensus intergenic (IG) sequence of S mRNA located within the 5′ UTR (Zhang *et al.*, 1991). The oligonucleotide primer 3B11 was used to specifically assess whether the intragenic IG homolog sequence located at nucleotide positions 799-805 of the S gene was utilized for the generation of mRNAs. A "shotgun" cloning of the heterogeneous LBCV/3B11 RT PCR products yielded a total of 33 DNA fragments encompassing S sequences ranging in size from 0.3 to 1.1 kb. Five DNA fragments contained the predicted size of RT-PCR products, while the other 28 contained shorter junctions of different sizes. DNA sequence analysis of these junctions revealed 9 different types, all of which contained the leader fused to S coding sequences (Fig. 1). The majority of the cloned DNA fragments represented type 4 junctions (14 clones) containing the targeted consensus sequence. The remaining junctions utilized nonconsensus sequences located both upstream and downstream from this consensus sequence analog. Similarly, different types of junction sequences were detected at the 3′ end of the S mRNA proximal to the other IG-like sequence, as well as within the 12.7 kDa ORF (Chouljenko and Kousoulas, 2000).

Overall, RT-PCR analysis using primer pairs that targeted two of the three intragenic consensus analog sequences located at the 5′ and 3′ ends of S, respectively, confirmed that they were utilized for the production of mRNAs in infected cells. In addition, leader mRNA junctions containing nonconsensus sequences were also detected. Therefore, leader-mRNA junction formation is not strictly dependent on sequence homology between the 3′ end of the leader and the sequences of the mRNA fusion sites or on the location of these fusion sites within the mRNA. However, sequence complementarity seems to increase the efficiency by which homologous sites are used because canonical mRNAs are the predominant species detected by northern analysis (not shown).

Cryptic or alternative junction sequences were reported for a recombinant MHV mutant in which the green fluorescence protein (GFP) gene was inserted in the viral genome (Fischer *et al.*, 1997). However, these junctions may have formed due to the GFP gene insertion within the MHV genome and in this regard, differ from our findings that cryptic junctions can be detected in infections with wild-type BCV strains.

Additional control experiments were performed to assure that the detected cryptic leader-mRNA fusion sequences were not produced by copy-choice template switching of the reverse transcriptase or Taq polymerase during the RT-PCR procedures. In these experiments, we selected the cell-adapted L9 and respiratory OK strains, because they contain a number of nucleotide differences within the genomic region bracketing the leader-mRNA junction sequences of the 12.7 kDa ORF as

well as a single nucleotide substitution between OK and L9 leader sequences. RNA was extracted from a control sample, which was prepared by mixing 1:1 virus stocks of strains L9 and OK containing approximate the same number of plaque forming units (PFU). Fourteen different DNA fragments were cloned and sequenced. Seven DNA fragments contained L9-specific leader and part of the 12.7 kDa body-mRNA sequences. Seven additional cloned DNA fragments contained OK-specific leader and body-mRNA, while two of these DNA fragments represented cryptic junction sequences. Overall, none of these cloned DNA fragments contained hybrid L9/OK sequences (not shown) proving that RT-PCR did not generate the observed cryptic junctions.

3.4 Location of cryptic fusion sites on the predicted secondary structure of the S mRNA

We generated the predicted global minimum energy secondary structure of the entire S mRNA using RNAdraw (Matzura and Wennborg, 1996). The predicted secondary structure of the BCV S mRNA revealed an unusually high level of internal homology causing extensive formation of double stranded RNA as well as multiple single stranded loops of RNA (Fig. 2). Interestingly, all cryptic leader-mRNA junctions as well as the junction within the 5′ UTR were predominantly located within single stranded portions of the S mRNA. Therefore, single stranded regions of mRNAs may be required for leader-mRNA junction formation.

3.5 Single nucleotide substitutions proximal to junction sequences within the S and 12.7 kDa ORFs specified by different virus isolates

We found that different BCV strains contained a single nucleotide substitutions at the 5′ end of S, which coincided with the location of 6 cryptic in a total of 9 detected junctions (Fig. 1). In addition, a single nucleotide substitution was detected immediately proximal to the consensus heptameric sequence within the 3′ of the S ORF specified by RBCV strains associated with the 1998 Texas "shipping fever" epizootic, while the intestinal isolates remained unchanged (Chouljenko and Kousoulas, unpublished). Similarly, the intergenic region between the 4.9 and 12.7 kDa ORFs, where the 12.7 kDa ORF cryptic junctions were detected, is the most mutated area when RBCV strains are compared among them as well as with EBCV strains (Chouljenko *et al.*, 1998). It is conceivable that the cryptic junction sequences also represent "hotspots" for recombination that may lead to sequence variation at these sites.
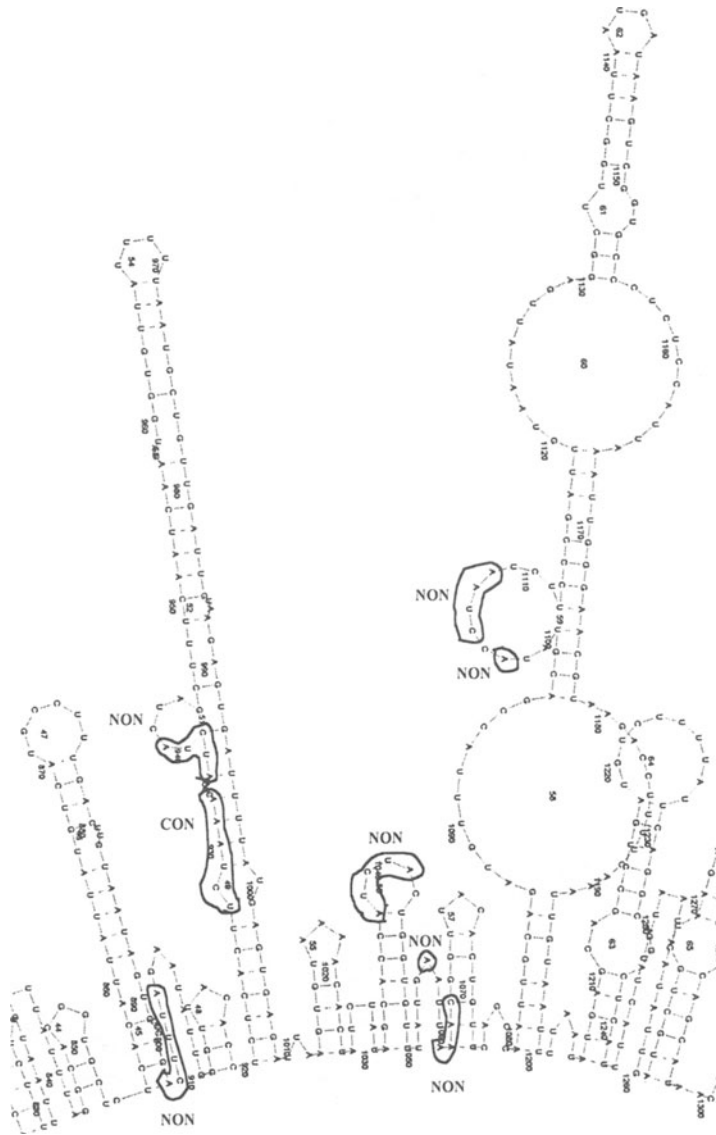
Figure 2. **Predicted secondary structures of the S mRNA.** A portion of the entire predicted secondary structure of the RBCV OK S mRNA is shown. Leader-mRNA body junctions utilizing the consensus and nonconsensus sequences are boxed and marked CON and NON, respectively.

REFERENCES

Chouljenko, V. N., and K. G. Kousoulas. 2000. Cryptic leader-mRNA fusion sites within the bovine coronavirus S and 12.7 kDa coding sequences (Submitted).

Chouljenko, V. N., K. G. Kousoulas, X. Lin, and J. Storz. 1998. Nucleotide and predicted amino acid sequences of all genes encoded by the 3' genomic portion (9.5 kb) of respiratory bovine coronaviruses and comparisons among respiratory and enteric coronaviruses. Virus Genes. **17**:33-42.

Fischer, F., C. F. Stegen, C. A. Koetzner, and P. S. Masters. 1977. Analysis of a recombinant mouse hepatitis virus expressing a foreign gene reveals a novel aspect of coronavirus transcription. J. Virol. **71**:5148-60.

Storz, J., C. W. Purdy, X. Lin, M. Burrell, R. E. Truax, R. E. Briggs, G. H. Frank, R. W. Loan. Isolation of respiratory bovine coronavirus, other cytocidal viruses, and Pasteurella spp from cattle involved in two natural outbreaks of shipping fever. J. Am. Vet. Med. Assoc. **216**:1601-1606.

Storz, J., L. Stine, A. Liem, and G. A. Anderson. 1996. Coronavirus isolation from nasal swap samples in cattle with signs of respiratory tract disease after shipping. J. Am. Vet. Med. Assoc. **208**:1452-5.

Zhang, X. M., K. G. Kousoulas, and J. Storz. 1991. Comparison of the Nucleotide and Deduced Amino Acid Sequences of the S-Genes Specified by Virulent and Avirulent Strains of Bovine Coronaviruses. Virology. **183**:397-404.

Matzura, O., and A. Wennborg. 1996. RNAdraw:an integrated program for RNA secondary structure calculation and analysis under 32-bit Microsoft Windows. Comp. Applic. Bioscie. (CABIOS). **12**:247-249.