CLONING AND SEQUENCING THE NUCLEOCAPSID AND E1 GENES OF CORONAVIRUS

John Armstrong, Sjef Smeekens[+1], Willy Spaan[+], Peter Rottier[+] and Ben van der Zeijst[+]

European Molecular Biology Laboratory
Postfach 10.2209, 69 Heidelberg, FRG

[+]Institute of Virology, Veterinary Faculty, State University of Utrecht, 3508 TD Utrecht, The Netherlands

[1]Present address: Molecular Cell Biology Group
State University of Utrecht, Padualaan 8, 3808 CH Utrecht

INTRODUCTION

The widespread medical and veterinary importance of Coronaviruses provided the initial reason for studying these viruses at the molecular level. However, two unusual features of the viral life cycle are of particular interest for "pure" molecular biology. First, the intracellular budding site of Coronaviruses, apparently associated with the restricted intracellular distribution of the E1 glycoprotein[1] suggests that this protein may provide a model to study the transport and sorting of membrane proteins of the endoplasmic reticulum and Golgi apparatus. Secondly, the unusual "nested set" structure of the viral RNA's[2,3,4,5](Spaan et al., this volume) implies a replication mechanism unlike that of other RNA viruses.

With the aim of learning more about both these aspects of Coronavirus molecular biology, we have prepared cDNA clones whose sequences span the nucleocapsid and E1 genes of MHV-A59.

MATERIALS AND METHODS

Preparation of cDNA from mRNA of MHV-A59-infected cells, cloning and sequence determination were as described previously[6]. Further clones were isolated from the same cDNA preparation by digestion with the restriction enzymes BalI (BRL) or RsaI (New England Biolabs) and ligation to the plasmid vector pEMBL8[7]. Single-stranded DNA from

P. J. M. Rottier et al. (eds.), *Molecular Biology and Pathogenesis of Coronaviruses*

recombinant clones was prepared according to Dente et al.[7] but using wild-type phage fd for superinfection.

RESULTS

Corrected sequence of the nucleocapsid gene

Comparison of the original sequence of the nucleocapsid gene[6] with that of the corresponding gene from MHV JHM strain (Skinner and Siddell, this volume) revealed, among other differences, an apparent change of reading frame between nucleotides 478 and 795 (numbering as 6); two of the three possible reading frames lack terminator codons throughout this region. Re-examination of the sequencing autoradiograms, however, clearly shows that the apparent difference is due to two errors in the sequence reported for A59; the corrected sequence includes an additional C between bases 477 and 478, and lacks the G at position 795. The resulting amino acid sequence of the nucleocapsid is shown in Fig.1.

MSFVPGQENAGGRSSSVNRAGNGILKKTTWADQTERGPNNQNRGRRNQPKQTATTQPNSGSVVPHY
SWFSGIFTQFQKGKEFQFAEGQGVPIANGIPASEQKGYWYRHNRRSFKTPDGQQKQLLPRWYFYYL
GTGPHAGASYGDSIEGVFWVANSQADINTRSDIVERDPSSHEAIPIRFAPGTVLPQGFYVEGSGRS
APASRSGSRSQSRGPNNRARSSSNQRQPASTVKPDMAEEIAALVLAKLGKDAGQPKQVIKQSAKKV
RQKILNKPRQKRIPNKQCPVQQCFGKRGPNQNFGGSEMLKLGTSDPQFPILAELAPTVGAFFFGSK
LELVKKNSGGADEPTKDVYELQYSGAVRFDSTLPGFETIMKVLNENLNAYQKDGGADVVSPKPQRK
GRRQAQEKKDEVDNVSVAKPKSSVQRNVSERLTPEDRSLLAQILDDGVVPDGLEDDSNV

Figure 1. Amino-acid sequence of MHV-A59 nucleocapsid protein.

Sequence of the E1 gene

Digestion of cDNA with the restriction enzymes BalI and RsaI gave clones whose sequences encompassed the E1 gene (Fig.2). Clone R55 contains part of the leader region common to the 5' end of all the viral RNA's (Spaan et al., this volume). Clones R55 and PR9 have subsequently been shown to be adjacent, by subcloning and sequencing of a full-length E1 clone (Niemann, this volume). A previously unidentified clone, F11, also comes from within the E1 gene (Fig.2).

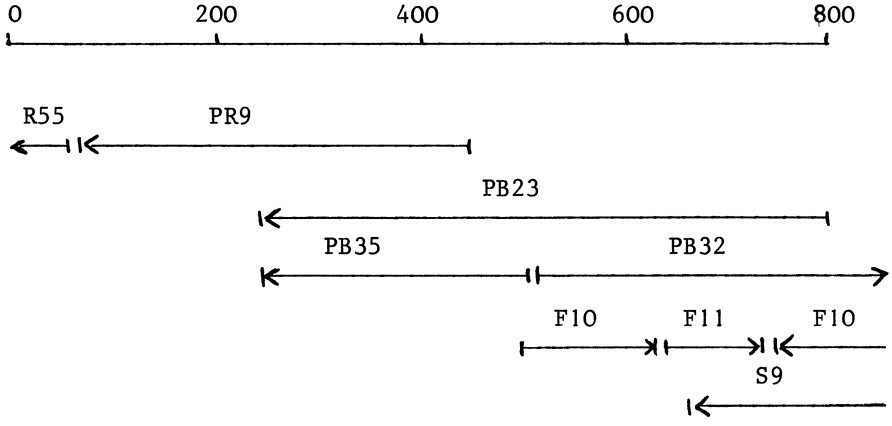

Figure 2. Arrangement of clones spanning the E1 gene of MHV-A59. Arrows show the direction in which the sequence was determined. Numbering is as Fig. 3. cDNA was digested with RsaI (R), BalI (B), Fnu DII (F) or S1 nuclease (S), and cloned in M13 mp8 or pEMBL8 (P).

Analysis of clones from the nucleocapsid region had revealed two, S9 and F10, whose sequences diverged at their 5' ends at almost exactly the inferred site of fusion between the leader sequence and the nucleocapsid-coding region of RNA7[6], suggesting that one of these clones originated from RNA7 and the other from a larger RNA. Comparison of the two sequences with those from the E1 gene, and with sequences determined directly from RNA7 and the viral genome (Spaan et al., this volume) shows that clone S9 originates from one of the six larger RNA's and covers the region from the 3' end of the E1 gene to the 5' end of the nucleocapsid gene. Clone F10, however, represents a fusion of an internal portion of the E1 gene to the 5' end of the nucleocapsid, in opposite orientations (Fig.2); presumably it arose from artefactual synthesis of a "third" DNA strand during synthesis of the second strand of the cDNA.

The sequence of the E1 gene, and the translated amino acid sequence of the protein are shown in Fig.3.

```
                                                          M  S  S  T  T  Q
ACCTCTCAACTCTAAAACTCTTGTAGTTTAAATCTAATCCAAACATTATGAGTAGTACTACTCA

  A  P  E  P  V  Y  Q  W  T  A  D  E  A  V  Q  F  L  K  E  W  N
GGCCCCAGAGCCCGTCTATCAATGGACGGCCGACGAGGCAGTTCAATTCCTTAAGGAATGGAAC
                                                40
 F  S  L  G  I  I  L  L  F  I  T  I  I  L  Q  F  G  Y  T  S  R
TTCTCGTTGGGCATTATACTACTCTTTATTACTATCATACTACAGTTCGGTTACACGAGCCGTA
                                          60
S  M  F  I  Y  V  V  K  M  I  I  L  W  L  M  W  P  L  T  I  V  L
GCATGTTTATTTATGTTGTGAAAATGATAATCTTGTGGTTAATGTGGCCACTGACTATTGTTTT
                                    80
  C  I  F  N  C  V  Y  A  L  N  N  V  Y  L  G  F  S  I  V  F  I
GTGTATTTTCAATTGCGTGTATGCGCTAAATAATGTGTATCTTGGATTTTCTATAGTGTTTACT
                                100
 I  V  S  I  V  I  W  I  M  Y  F  V  N  S  I  R  L  F  I  R  T
ATAGTGTCCATTGTAATCTGGATTATGTATTTTGTTAATAGCATAAGGTTGTTTATCAGGACTG
                           120
G  S  W  W  S  F  N  P  E  T  N  N  L  M  C  I  D  M  K  G  T  V
GTAGCTGGTGGAGCTTCAACCCCGAAACAAACAACCTTATGTGTATAGATATGAAAGGTACCGT
                      140
  Y  V  R  P  I  I  E  D  Y  H  T  L  T  A  T  I  I  R  G  H  L
GTATGTTAGACCCATTATTGAGGATTACCATACACTAACAGCCACTATTATTCGTGGCCACCTC
                    160
 Y  M  Q  G  V  K  L  G  T  G  F  S  L  S  D  L  P  A  Y  V  T
TACATGCAAGGTGTTAAGCTAGGCACCGGTTTCTCTTTGTCTGACTTGCCCGCTTATGTTACAG
               180
V  A  K  V  S  H  L  C  T  Y  K  R  A  F  L  D  K  V  D  G  V  S
TTGCTAAGGTGTCACACCTTTGCACTTATAAGCGCGCATTCTTAGACAAGGTAGACGGTGTTAG
         200
  G  F  A  V  Y  V  K  S  K  V  G  N  Y  R  L  P  S  N  K  P  S
CGGTTTTGCTGTTTATGTGAAGTCCAAGGTCGGAAATTACCGACTGCCCTCAAACAAACCGAGT
   220
  G  A  D  T  A  L  L  R  I
GGCGCGGACACCGCATTGTTGAGAATCTAATCTAAACTTTAAGGATG
```

Figure 3. Sequence of the MHV-A59 E1 gene and protein, including part of the leader region. Clone R55 contains an additional 5 nucleotides in the leader, as shown in Fig. 4. The last three bases shown correspond to the initiation codon of the nucleocapsid gene. E1 protein has a predicted molecular weight of 26000.

DISCUSSION

Topography of the E1 protein

The assembly of the E1 protein into microsomal membranes, and its disposition across the lipid bilayer, have been investigated by Rottier et al. (this volume). Several of its features can be compared with the sequence shown in Fig.3.

1. In contrast to the majority of membrane proteins, E1 lacks a cleaved "leader" peptide: inspection of the N-terminal region of the sequence shows no good candidates for a cleavage site[8].

2. The N-terminal portion carries the unusual O-linked sugars found in the mature protein[9]; assuming the terminal Met is removed (usually the case in eukaryotes), the N-terminal sequence is Ser-Ser-Thr-Thr, an obvious site for potential O-glycosylation.

3. Only approximately 2.5kD of polypeptide are susceptible to proteolysis from the N-terminus, on the inside of the microsomal vesicle, and 1.5kD from the C-terminus on the outside, implying that the rest of the protein is buried in the membrane. A sequence of 22 uncharged residues, from positions 26 to 47, represents a potential membrane-spanning region; the first 25 residues would then correspond to the portion removed by protease. A further sequence of uncharged residues, from positions 57 to 106, is sufficiently long to represent second and third membrane-spanning segments. There are no further hydrophobic sequences, implying that the region from residues 107 to approximately 190 is either folded in the membrane to neutralise charges, or, more plausibly, is adjacent to the membrane, but resistant to proteolysis. The remaining C-terminal portion would then correspond to the proteolysed terminus.

Thus, the sequence in general confirms the various unusual characteristics of the E1 glycoprotein, any of which may be related to its restricted intracellular distribution. The availability of cDNA to the E1 gene, in a single clone (Niemann, this volume) presents the opportunity to dissect the functions of the molecule, by mutagenesis and expression of the gene.

Synthesis of Coronavirus mRNA's

It is now clear that the subgenomic RNA's of MHV-A59 share a short 5' "leader" region, probably corresponding to the 5' end of the genome RNA[5,10] (and Spaan et al., this volume). Assuming that all the RNA's are synthesized from a full-length negative-strand template[11], and that they are not produced from genome-length

precursors[12], a possible mechanism for RNA synthesis would be completion of the leader region, followed by "jumping" of the viral polymerase to one of several sites within the template, upstream of each gene: synthesis would then continue until the end of the template was reached. How could such sites be recognized, and how would their efficiencies be regulated to ensure synthesis of the RNA's in the correct proportions?

```
                                                    Met(Nucleocapsid)
RNA7          uuuaAAUCUAAUCUAAACuuuaaggaug
Clone S9      ugagAAUCUAAUCUAAACuuuaaggaug
                   (E1)Stop
                                          Met(E1)
RNA6          uuuaAAUCUAAUCcAAACauuaug
RNA6,
clone R55     uuuaAAUCUAAUCUAAUCcAAACauuaug
```

Figure 4. Conserved sequences upstream of coding regions in MHV-A59. The sequence adjacent to the nucleocapsid gene in RNA7 (Spaan et al., this volume) is aligned with the corresponding region in the larger RNA's (clone S9), and with the two sequences upstream of the E1 gene; from a full-size E1 clone (Niemann, this volume) and from clone R55.

Comparison of the various sequences upstream of the nucleocapsid and E1 genes shows some features of interest (Fig.4). A sequence of 14 bases, with one mismatch, is present on the 5' side of the E1 gene in RNA6, and in the intergenic region between the E1 and nucleocapsid genes, suggesting that this represents a site for re-initiating RNA synthesis, and will be found on the 5' side of all the viral genes. However, one clone from the 5' end of the E1 gene, clone R55, contains an additional 5 nucleotides next to the sequence of 14 (Fig.4), reminiscent of the sequence of RNA7 from the JHM strain (Skinner and Siddell, this volume). Thus, there is apparent heterogeneity in the site of fusion between the leader and the E1-coding region in RNA6, consistent with the low molar yield of an oligonucleotide from this region isolated by Lai et al.[10].

A possible mechanism which accommodates these data is that the sites for re-initiation of RNA synthesis are recognized by base-pairing of the leader sequence to internal sites within the negative-strand template. This would be possible if part of the 14-base sequence of Fig.4 was also present in the leader itself. This is illustrated, for RNA6, in Fig.5. A consequence of the sequence of this region, however, is that the leader could base-pair in an alternative position, generating the lengthened sequence found in clone R55 (Fig.5).

```
                3'-UUAGAUUAGGUUUGUAAUAC-5'   Template
          5'-uuuaaaucuaaucCAAACAUUAUG-3'     RNA6
5'-uuuaaaucuaaucUAAUCCAAACAUUAUG-3'          RNA6,CLONE R55
```

Figure 5. Hypothetical base-pairing between the leader RNA (small letters) and the negative-strand template. Alternative positions of base-pairing could generate the alternative sequences observed in two RNA6 clones.

Clearly this model is at present very speculative, but it implies a possible mechanism for regulating the relative levels of synthesis of the RNA's: variations in the length, precision and number of positions of base-pairing between the leader and the template could determine the probability of re-initiation at a particular site. Further sequence analysis of the MHV-A59 RNA's, in particular of the 5' end of the genome, are now required to test this, and other possible models for the generation of Coronavirus mRNA's.

REFERENCES

1. K. Holmes and J.N. Behnke, Biochemistry and Biology of Coronaviruses, in "Advances in Experimental Medicine", V. Ter Meulen, S. Siddell and H. Wege, eds., vol. 142, Plenum Press, New York (1981).
2. D.F. Stern and S.I.T. Kennedy, Coronavirus multiplication strategy II. Mapping the avian infectious bronchitis virus intracellular RNA species to the genome, J. Virol. 36, 440-449 (1980)
3. M.M. Lai, P.R. Brayton, R.C. Armen, D.D. Patton, C. Pugh and S.A. Stohlman, Mouse hepatitis virus A59: mRNA structure and genetic localization of the sequence divergence from hepatotropic strain MHV-3, J. Virol. 39, 823-834 (1981)
4. S. Cheley, R. Anderson, M.J. Cupples, E.C.M. Lee Chan and V.L. Morris, Intracellular murine hepatitis-virus-specific RNA's contain common sequences, Virology 112, 596-604 (1981)
5. W.J.M. Spaan, P.J.M. Rottier, M.C. Horzinek and B.A.M. van der Zeijst, Sequence relationships between the genome and the intracellular RNA species 1,3,6 and 7 of mouse hepatitis virus strain A59, J. Virol. 42, 432-439 (1982)
6. J. Armstrong, S. Smeekens and P. Rottier, Sequence of the nucleocapsid gene from murine coronavirus MHV-A59, Nucl. Acids Res. 11, 883-891 (1983)

7. L. Dente, G. Cesareni and R. Cortese, pEMBL: a new family of single stranded plasmids, Nucl. Acids Res. 11, 1645-1655 (1983)
8. G. von Heijne, Patterns of amino acids near signal-sequence cleavage sites, Eur. J. Biochem. 133, 17-21 (1983)
9. H. Niemann and H.-D. Klenk, Coronavirus glycoprotein E1, a new type of viral glycoprotein, J. Mol. Biol. 153, 993-1010 (1981)
10. M.M. Lai, C.D. Patton and S.A. Stohlman, Further characterisation of mRNA's of mouse hepatitis virus: presence of common 5'-end nucleotides, J. Virol. 41, 557-565 (1982)
11. M.M. Lai, C.D. Patton and S.A. Stohlman, Replication of mouse hepatitis virus: negative-stranded RNA and replicate form RNA are of genome length, J. Virol. 44 , 487-492 (1982)
12. L. Jacobs, W.J.M. Spaan, M.C. Horzinek and B.A.M. van der Zeijst Synthesis of subgenomic mRNA's of mouse hepatitis virus ls initiated independently: evidence from UV transcription mapping, J. Virol. 39, 401-406 (1981)

ACKNOWLEDGEMENTS

We thank Michael Skinner and Heiner Niemann for sharing data and Annie Steiner for preparation of the manuscript. J.A. was the recipient of a European Fellowship from the Royal Society. P.R. was supported at the EMBL by a short-term EMBO fellowship.