IDENTIFICATION OF CORONAVIRAL CONSERVED SEQUENCES AND APPLICATION TO VIRAL GENOME AMPLIFICATION

Anne Bridgen, Kurt Tobler and Mathias Ackermann

Institute for Virology
University of Zürich
Winterthurerstrasse 266a
CH-8057 Zürich

Our work with the porcine epidemic diarrhoea virus (PEDV) has led us to look more closely at sequences and sequence motifs which are conserved between different coronaviral genomes. This is possible, since sequence information from at least part of the genomes of twelve coronaviruses is now available. Many of these motifs are specific to coronaviruses while others are shared with other positive strand RNA viruses, for example with the torovirus Berne virus and the flavivirus equine arteritis virus[1]. Functions can be assigned to some of these conserved regions, for example the spike (S) and membrane (M) protein transmembrane regions and the S leucine-zipper motif[2]. Other well conserved domains, including regions within the M and nucleocapsid (N) protein genes, have as yet no defined role. The identification of such conserved sequences is important for the recognition of functional domains of the viral RNA and proteins and also of regions useful in the cloning of novel coronaviruses using techniques based on the polymerase chain reaction (pcr). They can also be used to assess the evolutionary relationships of different virus groups. We plan to discuss the nature and possible functions of the conserved sequence motifs in more detail elsewhere, and to concentrate on the application of these sequences to viral genome amplification in this article.

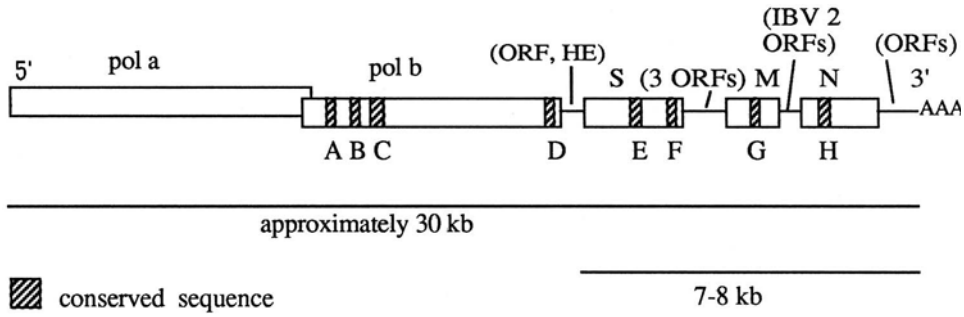

FIGURE 1

Figure 1 (not to scale) shows a representation of a standard coronavirus genome including the major open reading frames (ORFs) where HE, S, M and N represent the haemagglutinin-esterase, spike, membrane and nucleocapsid proteins, respectively. Pol a and

Coronaviruses, Edited by H. Laude and J.F. Vautherot
Plenum Press, New York, 1994

b represent the two ORFs of the polymerase gene, of which the second is expressed through ribosomal frame shifting. IBV represents Infectious Bronchitis Virus. ORFs which are not found in all coronaviruses sequenced to date are shown in parentheses. Eight sequence regions which show a high level of homology beween coronaviral genomes and which are long enough to be used for pcr amplification of the viral genome are indicated by hatching. These sequences vary in length from twenty-four to several hundred bases and represent the following regions, where amino acids (aa) are displayed in the normal one letter code:

A. polymerase region I, GDD motif[1]
B. polymerase region II, cysteine rich region[1]
C. polymerase region III, helicase domain[1]
D. polymerase region IV, conserved domain[1]
E. S gene: region of nearly 300 amino acids showing high sequence conservation
F. S gene: consensus sequence YVKWPKYVWL
G. M gene: region including the consensus sequence SWWSFNPE
H. N gene: GYW -(4aa)- R -(4-7aa)- G -(10aa)- FYYLGTGP -(11aa)- DGV(V/F)WVA

We amplified the 2.2 kbp cDNA located nearest to the 3' end of the PEDV genome using primers based on the conserved sequences in regions G and H, as shown below. The primers were synthesised to contain restriction enzyme sites, but these were not used for product cloning. Rare amino acid variations from the consensus sequence are shown in parentheses.

```
                      (F)
            F   Y   Y   L   G   T   G   P      Amino acids (H, Fig. 1)
P23  A AGC TTT TAC TAC TTA GGA ACA GGA CC      Primer, mRNA sense
5'end HindIII    T   G   C                     27mer, 18 fold degenerate
                     T   T

       D(E)  G V(I) V/F  W   V   A             Amino acids (H, Fig. 1)
P24     CTG CCA CAG AAG ACC CAG CGA GCT C      Primer, antisense
3'end    T   T     C              XhoI         25mer, 8 fold degenerate

             S  W(F) W   S  F(W) N   P   E     Amino acids (G, Fig. 1)
P25  G ACT AGT TGG TGG AGA TTT AAC CCA GA      Primer, mRNA sense
5'end   SpeI            T       T   T         27mer, 8 fold degenerate
```

Primers P24 and P25 were used to amplify a 680 bp fragment from single stranded cDNA primed with P24 or oligo dT. The 3' most sequences were amplified in a modification of the RACE (rapid amplification of cDNA ends) technique of Frohman et al.[3], which is described in more detail elswhere[4]. Briefly, the original Frohman protocol required the use of two related primers, of which one was used for cDNA synthesis from the poly A tail and the other for pcr amplification, in addition to a sequence specific primer. Instead, we created a dT containing primer for first strand cDNA synthesis by addition of thymidine residues to P25 with terminal deoxytransferase, and then used either P25 alone or P25 with P23 for the pcr amplification of a 2.2 kbp or a 1.6 kbp fragment, respectively.

This technique of genome amplification was developed for PEDV, which shows poor antigenic cross-reactivity to other coronaviruses and is still unclassified into an antigenic sub-group[5]. We thus anticipate that the method will be applicable to the cloning and sequencing of other, presently uncharacterised coronaviruses. It should be possible to clone all the structural genes using the conserved regions D-H in Figure 1, since none of these regions is located much more than 2 kb from the next. Such a means of obtaining viral clones is fast, requires less viral RNA and eliminates the need for extensive clone screening.

REFERENCES

1. Snijder, E.J., den Boon, J.A., Bredenbeek, P.J., Horzinek, M.C., Rijnbrand, R. and Spaan, W.J.M. (1990). *Nucleic Acids Res.* 18: 4535-4542.
2. Britton, P. (1991). *Nature* 353: 394.
3. Frohman, M.A., Dush, M.K. and Martin, G.R. (1988). *Proc. Natl. Acad. Sci.* 85: 8998-9002.
4. Tobler, K., Bridgen, A. and Ackermann, M. *Adv. Exp. Med. Biol.* This volume.
5. Pensaert, M.B., Debouck, P. and Reynolds, D.J. (1981). *Arch. Virol.* 68: 45-52.