Arch Virol (1989) 105: 165–178

Archives of Virology

Sequence of the coding regions from the 3.0 kb and 3.9 kb mRNA Subgenomic species from a virulent isolate of transmissible gastroenteritis virus

P. Britton[1], **C. Lopez Otin**[2], **J. M. Martin Alonso**[2], and **F. Parra**[2]

[1] Division of Microbiology, A.F.R.C. Institute for Animal Health, Compton Laboratory, Newbury, Berks., U.K.

[2] Departamento de Biologia Funcional (Area de Bioquimica), Universidad de Oviedo, Oviedo, Spain

Accepted February 20, 1989

Summary. Subgenomic mRNA from a virulent isolate of porcine transmissible gastroenteritis virus (TGEV) was used to produce cDNA clones covering the genome region from the 3′ end of the peplomer gene to the start of the integral membrane protein gene. The nucleotide sequence of this area was determined using clone pTG11 and a previously reported cDNA clone pTG22. Three open reading frames (ORFs) were identified encoding putative polypeptides of relative molecular masses (M_r) 6,600, 27,600, and 9,200. The sequence encoding the M_r 9,200 polypeptide was found to be present on the "unique" 5′ region of the 3.0 kb mRNA species whereas the other two ORFs mapped on the 3.9 kb mRNA species. Differences between the ORFs from this strain of TGEV and those from a previously reported avirulent strain of TGEV were compared.

Introduction

TGEV belongs to the family *Coronaviridae*, a large group of pleomorphic enveloped viruses with a positive-stranded RNA genome, and causes gastroenteritis in pigs resulting in a high mortality in neonates. Coronavirus proteins are expressed from a 'nested' set of subgenomic mRNAs with common 3′ termini but different 5′ extensions which are translated to produce viral proteins whose genes are absent on the smaller mRNA species. Mouse hepatitis virus (MHV) and infectious bronchitis virus (IBV) mRNA species contain short non-coding sequences at their 5′ ends which appear to be joined to the regions encoding the viral genes by discontinuous transcription. A consensus sequence identified upstream of each gene/ORF may act as a binding site for the RNA polymerase-leader complex [11, 12, 25, 36, 40]. It has been previously postulated that a heptameric sequence ACTAAAC [9, 10] or a hexameric sequence CTAAAC

[20, 33, 34] may be involved in the binding of the TGEV RNA polymerase-leader.

The TGEV virion contains three major structural polypeptides; a surface glycoprotein (spike or peplomer protein) with a monomeric M_r 200,000, a glycosylated integral membrane protein observed as a series of polypeptides of M_r 28,000–31,000 and a basic phosphorylated protein (the nucleoprotein) of M_r 47,000 associated with the viral genomic RNA [16]. TGEV infected cells, in addition to the genomic RNA, have six species of subgenomic viral mRNA [7, 18]. Expression and sequencing studies have shown that the 1.7 kb mRNA species encodes the nucleoprotein gene [7–9, 20, 34] and the 2.6 kb mRNA species encodes the integral membrane protein gene [10, 21, 26, 34]. Sequencing studies have shown that the 8.4 kb mRNA species encodes the peplomer gene [32]. The largest RNA species of about 25 kb must encode the RNA dependant-RNA polymerase as shown for IBV [5]. The 0.7 kb mRNA species [7, 9] contains a single ORF encoding a polypeptide (ORF-4) of M_r 9,000 [9, 20, 34]. Antisera raised against synthetic oligopeptides derived from ORF-4 reacted to a polypeptide of M_r 14,000 in TGEV infected cells (unpublished result). The other mRNA species of 3.0 kb and 3.9 kb [7, 18] have had no product assigned to them from either infected cells or virions. A polypeptide product of M_r 24,000 has been identified by in vitro translation of the 3.9 kb mRNA species (Purdue strain) in rabbit reticulocyte lysate [18] which was not detected in TGEV infected cells nor in purified virions and was not immuneprecipitated with anti-virion protein antibodies. Sequencing studies [34], on the avirulent Purdue strain of TGEV (Purdue-115 [2]), have shown that the 3.9 kb mRNA species has two possible ORFs, X2a and X2b, encoding putative polypeptides of M_r 7,700 and 18,800 and an ORF, X1, encoding a polypeptide of M_r 9,200 on the 3.0 kb species.

The observation that more than one ORF can be found on the 5′ 'unique' region of a particular mRNA has also been described in two other coronaviruses MHV and IBV. Sequencing studies have shown that mRNA B species from the Beaudette strain of IBV has an ORF with the translation potential for a polypeptide of M_r 7,500 [3] and that mRNA D has three ORFs encoding potential polypeptides of M_r 6,700, 7,400, and 12,400 [4]. Similarly mRNA 4 from the JHM strain of MHV has an ORF encoding a potential polypeptide of M_r 15,200 [37] and mRNA 5 has two ORFs encoding potential polypeptides of M_r 12,400 and 10,200 [38]. It appears then that some of the coronavirus mRNA species may carry and express more than one gene product, however, their in vivo detection has proven to be difficult [3, 27, 38].

Previous sequencing studies of the genome encoding the TGEV integral membrane protein and nucleoprotein genes, have shown a very high degree of homology between the virulent FS 772/70 strain [9, 10] and the avirulent Purdue strain [20, 21, 26, 34] with minor changes in their amino acid sequences, the majority of which are conservative substitutions. Thus studies were undertaken to address the question of how different are the genome regions potentially

coding for non-structural gene products from a virulent field isolate of TGEV as compared with an avirulent laboratory strain. We report in this paper the cloning and sequencing of the genome area corresponding to the 5′ coding regions of the 3.9 and 3.0 kb mRNA species from the FS772/70 strain of TGEV.

Materials and methods

Preparation of viral RNA

TGEV poly(A)-containing mRNA was prepared from LLC-PK1 cells infected with TGEV strain FS772/70 and purified from other RNA species on poly(U) Sepharose as described previously [8, 9].

Digestion and analysis of plasmid DNA

Standard recombinant DNA methods were used [29] with enzymes purchased from New England Biolabs (CP Laboratories, Bishop's Stortford) unless otherwise stated in the text. DNA fragments were isolated from agarose gels using Geneclean™ (Stratech Scientific Ltd, London). Ligation reactions were carried out as described before [6]. *E. coli* strain DH1 was used for isolation of TGEV cDNA clones. *E. coli* transformants were selected on LB plates containing ampicillin ($100\,\mu g ml^{-1}$).

cDNA synthesis

cDNA synthesis was carried out as described before [10] using a synthetic oligonucleotide to prime first strand synthesis. Transformants containing TGEV cDNA were identified by colony hybridisation, as described before [9], using two [^{32}P]-labelled TGEV cDNA fragments.

DNA sequencing

Specific restriction fragments were subcloned into M 13 mp 18 and mp 19 vectors and were sequenced using either the universal primer from the M 13 site or five synthetic oligonucleotides from within the TGEV cDNA sequences as primers. The synthetic oligonucleotide sequences were 5′-TTCTAGCTTTGTACCGC-3′(H1A), 5′-GTCATCTATGACAGTCA-3′(H1B), 5′-TGAAAAAGTGCACATCC-3′ (TG2B), 5′-TATAGCACTAACAACCT-GAT-3′(Oligo 8) and 5′-CTAAGTAGGCGAATCTTAAA-3′(Oligo 15) whose positions and directions are shown on Fig. 2. All the oligonucleotides were synthesised by the phosphoramidite method using an Applied Biosystem model 381 A synthesizer. DNA sequencing was carried out using [α-^{35}S]-dATP by the di-deoxy method [35] from single stranded DNA templates or directly from plasmid DNA [30] but using the Sequenase™ protocol.

Northern blot analysis

Specific restriction fragments from TGEV cDNA clones were separated by agarose gel electrophoresis, purified from the agarose gel using a Geneclean™ kit and labelled as described before [9]. TGEV sub-genomic mRNA species were isolated from TGEV infected LLC-PK1 cells 8 h post-infection, purified, denatured with 6 M glyoxal, electrophoresed into either 0.7% or 1% agarose gels, northern blotted onto Biodyne membranes and hybridised to the labelled cDNA fragments as described before [9]. The probes were hybridised in the presence of 50% formamide at 42 °C for 16 h. The membranes were washed three times in X2 SSC containing 0.1% SDS at room temperature, twice in X1 SSC containing 0.1% SDS at 68 °C and once in X0.2 SSC containing 0.1% SDS at 68 °C. (X1 SSC = 0.15 M NaCl, 0.015 M trisodium citrate pH 7.0).

Results

Cloning from TGEV mRNA species

TGEV poly(A)-containing mRNA species were isolated from virus infected LLC-PK 1 cells and used for the synthesis of cDNA. Production of plasmid pTG 22 was as described before [10]. Plasmid pTG 11 was produced as described for pTG 22. Plasmids pTG 22 and pTG 11 were identified by their ability to hybridise to two TGEV cDNA fragments. One of the fragments, H 11 (1.5 kbp), originated from within the peplomer gene and the other, H 26 (0.6 kbp), originated from within the 5′ coding region of the 3.9 kb mRNA species (unpublished result). pTG 22 only hybridised with H 26 whereas pTG 11 hybridised with both H 26 and H 11 indicating that pTG 11 extended further into the peplomer gene than did pTG 22. The cDNA in plasmid pTG 11 is 2.98 kbp long and from its restriction map was shown to overlap pTG 22 (Fig. 1) and extend 5.3 kb from the 3′ end into the TGEV genome. From the size and position along the TGEV genome it was deduced that the cDNA from plasmids pTG 11 and pTG 22 contained sequences from the 5′ ends of the 3.0 kb and 3.9 kb mRNA species and part of the 3′ end of the peplomer gene (Fig. 1).

Sequencing of TGEV cDNA

The strategy for sequencing the TGEV genome area from the 3′ end of the peplomer gene to the 5′ end of the integral membrane protein gene is summarised

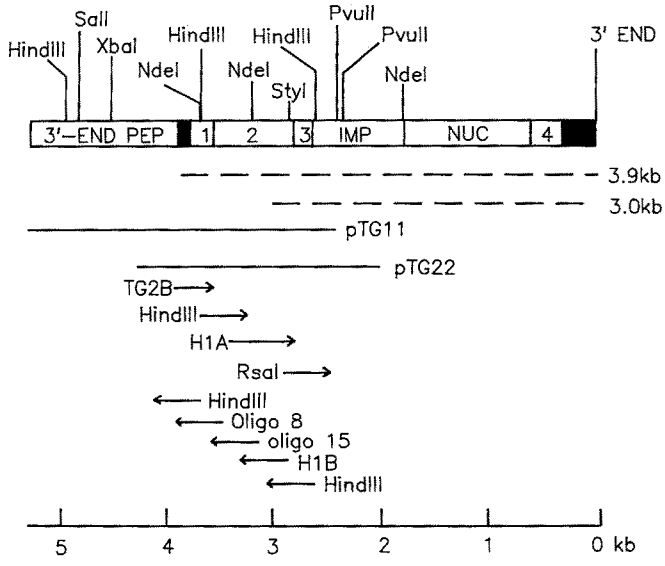

Fig. 1. Schematic diagram of the TGEV genome from the 3′ end to 5.3 kb into the genome. Relevant restriction sites, location of viral structural protein genes (*PEP, IMP, NUC*), and predicted ORFs (*1–4*) are marked. The broken lines show the areas represented in the 3.0 and 3.9 kb mRNAs. The unbroken lines show the positions of the cDNA inserts from plasmids pTG 11 and pTG 22 along the viral genome. The extent and direction of sequencing is shown by means of the arrows, where the restriction enzymes refer to the fragment subcloned and the other names the synthetic oligonucleotides used for priming

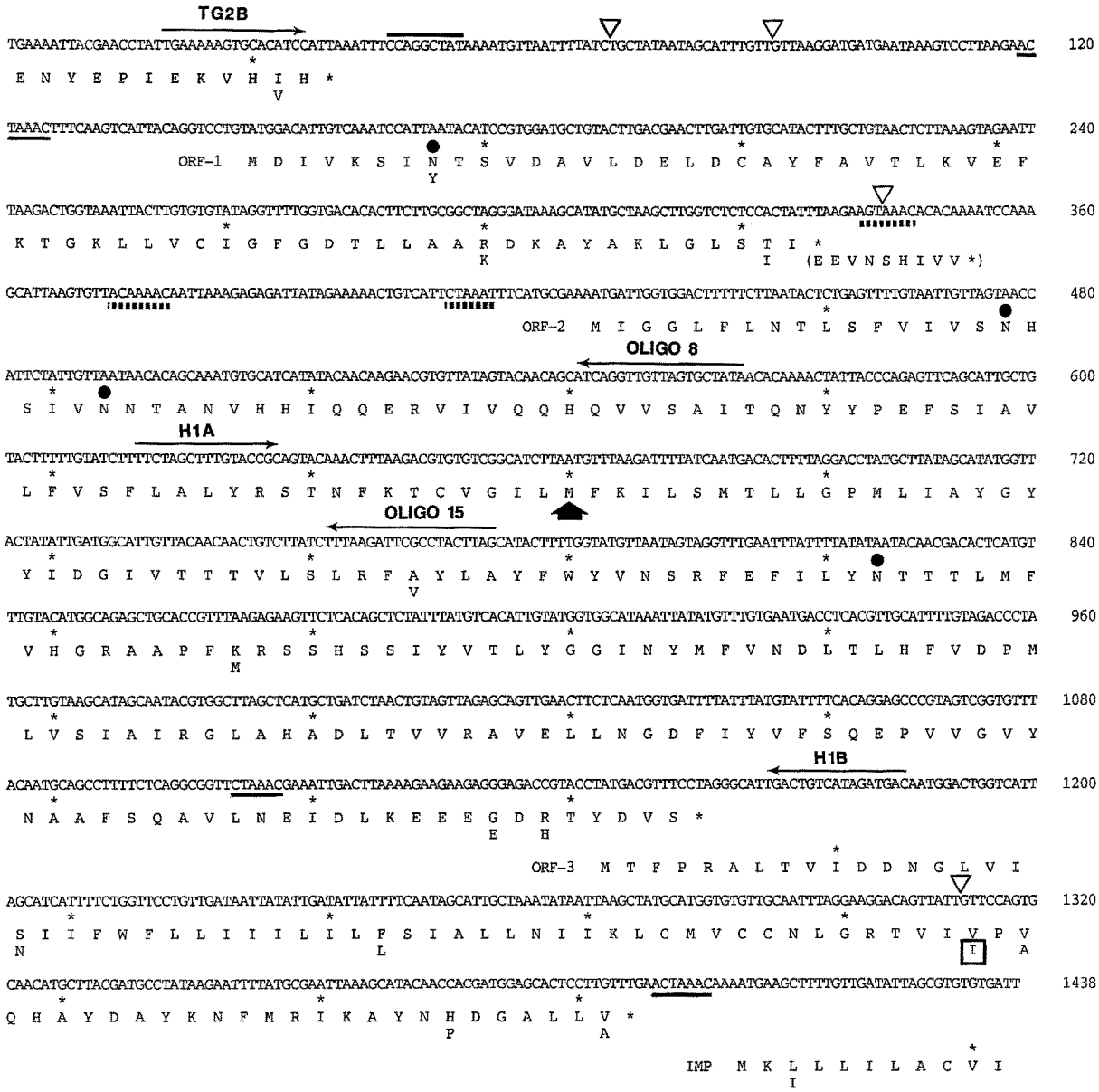

Fig. 2. The nucleotide and deduced amino acid sequences of TGEV ORFs 1–3 from the cDNA inserts in pTG 11 and pTG 22. The ACTAAAC and CTAAAC sequences postulated to be the RNA polymerase-leader complex recognition sites are underlined. The sequences preceeding the ORF-2 gene similar to the ACTAAAC or CTAAAC sequences are indicated by broken lines. Amino acids below the major sequences are substitutions found in the Purdue strain [34]. The bracketed amino acids in ORF-1 and the boxed amino acid in ORF-3 indicate amino acid insertions found in the Purdue strain. ● Potential N-glycosylation sites at asparagine residues. The vertical arrow indicates the start of the X2b ORF [34] on the genome of the Purdue strain of TGEV. The line above the sequence indicates the position of the insertion and ▽ the position of the deletions found in FS 772/70 when compared to the Purdue strain. The arrows show the positions and direction of the synthetic oligonucleotides. The end of the peplomer gene (nucleotides 1–37) and the start of the integral membrane protein gene (nucleotides 1406–1438) are shown. * These sequence data will appear in the EMBL/GenBank/DDBJ nucleotide sequence Database under the accession number X14551

in Fig. 1, in which the arrows show the regions and direction of sequencing from the M 13 clones. Thus the relevant part of the cDNA was sequenced in both directions. Several independent subclones from pTG 11, and corresponding ones from pTG 22, were sequenced with no differences between their cDNA sequences.

The resulting nucleotide sequence was translated in all six reading frames and the virus sense strand revealed three ORFs of 186 bp, 732 bp and 243 bp. The corresponding DNA sequence 148 bp from the 5′ end of the first ORF to the start of the TGEV integral membrane protein gene, present in pTG 22 [10] and pTG 11, is illustrated in Fig. 2. The three ORFs are designated ORF-1, of 62 amino acids between nucleotides 149–334; ORF-2, of 244 amino acids between nucleotides 429–1160 and ORF-3, of 81 amino acids between nucleotides 1150–1392. One ORF was identified, in the complementary strand, of 52 amino acids (M_r 5,802) between nucleotides 1,125–1,280 which was not preceeded by the potential RNA polymerase-leader complex binding site (ACTAAAC), though it does have the sequence CTAAAT 11 bases upstream of the ATG.

ORF-1, 186 bp, initiating from the ATG at position 149, has coding capacity for a putative polypeptide with a M_r 6,670. ORF-2, 732 bp, initiating from the ATG at position 429, has coding capacity for a putative polypeptide with a M_r 27,624. ORF-3, 243 bp, initiating from the ATG at position 1,150, has coding capacity for a putative polypeptide with a M_r 9,211. ORF-3 was found to overlap the 3′-end of ORF-2 by 12 nucleotides representing 4 amino acids. From their lengths and positions, from the start of the poly(A)-tail, ORFs-1 and -2 mapped within the 5′ coding region of the 3.9 kb mRNA species. The 5′ end of ORF-3 mapped within the 5′ coding region of the 3.0 kb mRNA species (see Fig. 1).

The nucleotide sequence (Fig. 2) revealed the presence, 23 bp from the ATG of ORF-1, of the heptameric sequence ACTAAAC, also found 5′ of the ATG sequences of other TGEV genes, see Table 1. Although no ACTAAAC sequence was found 5′ to the ATG of ORF-2 there are similar sequences AGTAAAC, ACAAAAC and CTAAAT (82 bp, 49 bp, and 11 bp upstream of the ATG

Table 1. Comparison of the potential RNA-polymerase binding sites, sequence contexts of initiator ATGs, and termination sequences of the genes from TGEV strain FS 772/70

Gene	Potential binding sites	Sequence context	Termination
ORF-1	ACTAAAC	(CC)TGT*ATG*G	TAA
ORF-2	see text	(CG)AAA*ATG*A	TAG
ORF-3	TCTAAAC	(TA)CCT*ATG*A	TGA
NUC	ACTAAAC	(TC)TAA*ATG*G	TAA
IMP	ACTAAAC	(AC)AAA*ATG*A	TAA
ORF-4	ACTAAAC	(AC)GAG*ATG*C	TAA

respectively) indicating possible RNA polymerase-leader complex binding sites. ORF-3 is preceded by the hexameric sequence CTAAAC, 37 bp 5′ to its ATG.

The sequence context, (CC)TGT*ATG*G (Table 1), about the ATG of the ORF-1 is not favourable for initiation by eukaryotic ribosomes, (CC)ACC*ATG*G [23, 24], due to the thymidine residue at position −3, though the presence of the guanosine residue at +4 may improve its efficiency. The sequence context, (CG)AAA*ATG*A (Table 1), for ORF-2 is favourable for initiation. The sequence context, (TA)CCT*ATG*A (Table 1), for ORF-3 is also not very favourable due to the presence of cytosine at position −3 though the efficiency may be improved by the presence of adenosine at position +4. The sequence context of the nucleoprotein gene, (TC)TAA*ATG*G (Table 1) [9], also appears not to be very favourable but is expressed efficiently in virus infected cells and this is also confirmed by expression studies in yeast [9] and vaccinia virus (unpubl. observation).

There is very little difference in the codon usage between ORFs1–3 and the nucleoprotein, integral membrane protein and ORF-4 genes. The codons GCG for alanine, TCG for serine, CGA for arginine, CTG for leucine and ATC for isoleucine are very rarely used for any of the genes suggesting that ORFs 1–3 encode genetic information and have not evolved by random incorporation of nucleotides.

Mapping of the ORFs to the TGEV subgenomic mRNA species

TGEV mRNA species were northern blotted onto Biodyne membranes as described in Materials and methods. Strips of Biodyne membrane containing the separated TGEV mRNA species were probed with various purified cDNA fragments. As can be seen from Fig. 3 the fragment from ORF-1 hybridised with the 3.9 kb mRNA species, indicating that the origin of ORF-1 is within the 5′ coding region of the 3.9 kb species. The fragment corresponding to ORF-2 hybridised with the 3.9 kb mRNA indicating that ORF-2 is also contained within the 5′ coding region of the 3.9 kb mRNA species. No mRNA species of 3.7 kb was detected, corresponding to the theoretical size of a mRNA species for ORF-2, in the infected cells under the conditions used. The fragment corresponding to ORF-3 hybridised with the 3.0 kb mRNA species indicating that the origin of ORF-3 was from within the 5′ coding region of this mRNA species. As can be seen from Fig. 3, sequences from other TGEV genes are found within the 5′ coding regions of their associated mRNA species. None of the sequences corresponding to ORF-4 or nucleoprotein (Fig. 3, track 4) and integral membrane protein genes (Fig. 3, track 5) detected a mRNA species of 3.7 kb that would correspond to a 'unique' mRNA species for ORF-2.

Discussion

The mRNA from a virulent British field isolate of TGEV was used to produce cDNA clones representing the 5′ coding regions from the 3.0 kb and 3.9 kb

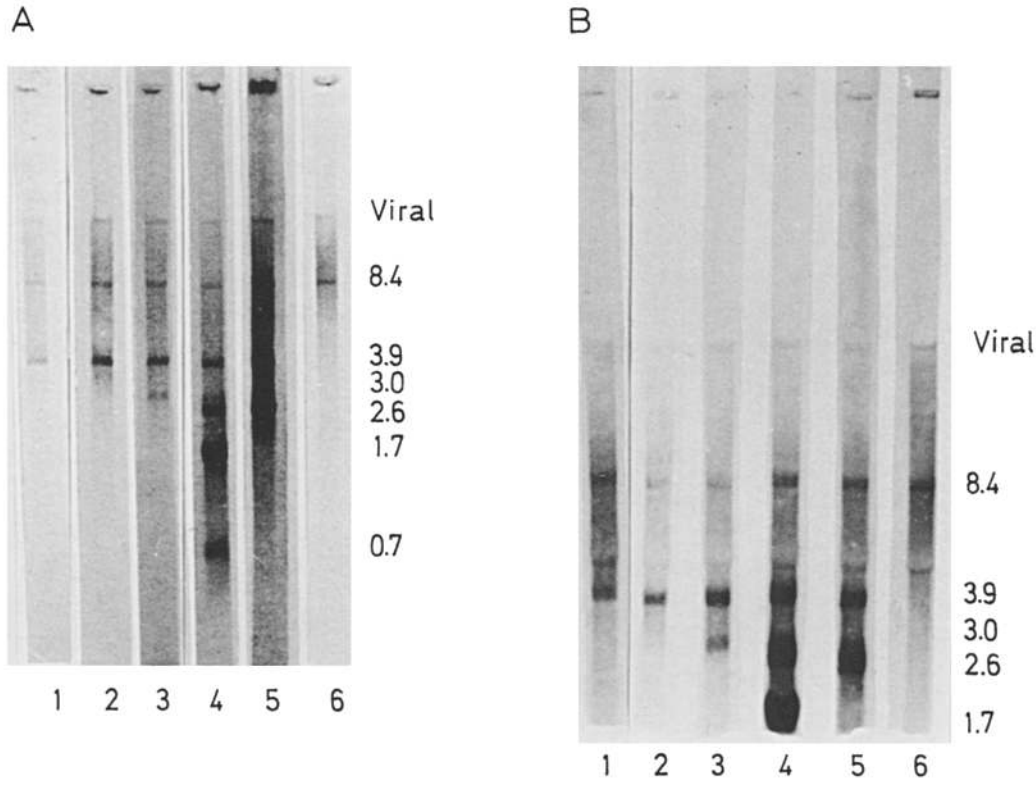

Fig. 3. Northern blot analysis of TGEV mRNA species separated **A** on a 1% agarose gel and **B** on a 0.7% agarose gel and hybridised to [^{35}S]-labelled cDNA probes derived from different ORFs. *1*, 0.17 kbp Ava II-Hind III fragment from ORF-1 (amino acids 1–56); *2*, 0.445 kbp Nde I-Sty I fragment from ORF-2 (amino acids 96–244); *3*, 0.265 kbp Rsa I-Hind III fragment from ORF-3 (amino acids 1–81); *4*, BamHI cassette of nucleoprotein gene and ORF-4 [9]; *5*, BamHI cassette of integral membrane protein gene [10]; and *6*, BamHI cassette of peplomer gene (Britton, unpublished data). The apparent band above the 3.9 kb mRNA in **B** results from the 28 S ribosomal RNA.

subgenomic mRNA species. Two independently isolated cDNA clones were sequenced allowing the identification of three ORFs, ORF-1 (186 bp), ORF-2 (732 bp) and ORF-3 (243 bp) in the viral sense strand (Figs. 1 and 2). The initiation codon of ORF-1 was preceded by the heptameric sequence, ACTAAAC, previously described as preceding the TGEV nucleoprotein, integral membrane protein and ORF-4 genes. ORF-2 is not preceded by the ACTAAAC sequence, though it has similar sequences present, and does not appear to be expressed on a separate mRNA species isolated from infected tissue culture cells. This may indicate that both ORFs may be expressed from the same message or that only one of them is translated; whether this occurs at the same or different times in the virus life-cycle is not known. The possibility that a new message be synthesized at a precise time in the virus life-cycle or that the message is only expressed in infected animals and not in tissue culture cells cannot be ruled out. ORF-3 is preceded by the hexameric sequence

CTAAAC indicating that this hexameric sequence is all that is required for recognition by the TGEV RNA polymerase-leader complex.

None of the three ORFs have any predicted N-terminal signal sequence using the weight-matrix method [41], indicating that the polypeptides, if synthesised, may reside in the cell cytoplasm and not be associated with the viral envelope or infected cell membranes. The lack of an N-terminal signal sequence does not rule out the possibility of an internal signal sequence as found in the integral membrane proteins of IBV and MHV. The predicted ORF-1 product has an overall charge of -1 and contains about 46% hydrophobic residues. The N-terminal end of the polypeptide appears to be acidic due to the presence of four aspartic and one glutamic acid residues between amino acids 12–19. The carboxyl terminus is slightly basic due to the presence of one arginine and two lysine residues between amino acids 50–56. There is one potential N-linked glycosylation site at residue 8 (Fig. 2) possibly increasing the molecular weight of the gene product to M_r 9,000 assuming that the presence a single N-linked glycan site can add 2,000 to the molecular weight of a polypeptide [22]. ORF-2 product has an overall charge of +4 and contains 46% hydrophobic residues. The N-terminus appears to be hydrophobic though the carboxyl terminus appears to be hydrophilic and acidic with four glutamic acid (three consecutive), three aspartic acid and two basic residues between amino acids 229–244. There are three potential N-linked glycosylation sites at residues 17, 22 and 132 (Fig. 2) possibly increasing the molecular weight of the gene product to M_r 33,600. ORF-3 product has an overall charge of +4 and contains 57% hydrophobic residues. The polypeptide appears to be very hydrophobic with 14.8% of the amino acid residues being leucine and 16% isoleucine. This large percentage of leucine and isoleucine residues is similar to ORF-4 which has 34.6% of the amino acid residues as leucine and 5.1% isoleucine [9]. The product does not contain any potential N-linked glycosylation sites (Fig. 2).

mRNA with coding capacity for more than one ORF have been found in other coronaviruses. IBV mRNA D contains three potential ORFs, D 1 encoding a polypeptide M_r 6,700, D 2 encoding a polypeptide M_r 7,400, and D 3 encoding a polypeptide of M_r 12,400. Chimaeric proteins have been produced [39], consisting of the ORFs fused to the *E. coli lacZ* gene, and antisera raised against them were used to immunoprecipitate proteins from virus infected cells. A polypeptide corresponding to the size of D 3 was immuneprecipitated from IBV infected cells with antisera to the D 3 chimaera and there was some evidence that the D 2 product may also be synthesised. DNA containing either D 3 or D 2 with D 3 was cloned into pSP 64 and SP 6 polymerase was used to generate RNA, which was then translated in vitro. The D 2 and D 3 products were produced from the DNA containing both genes in the wheatgerm translation system but expression of D 2 in the rabbit reticulocyte lysate system was very poor. The DNA containing D 3 alone was expressed in both systems. The D 2 ORF, like ORF-1 in TGEV, has a pyrimidine at position -3 from the initiation codon whereas D 3 and TGEV ORF-2 have a purine at -3, making expression

more favourable from the second ORF for TGEV and the third ORF for IBV. D 1 also has a pyrimidine at position -3 and there was no evidence that D 1 was produced in vivo.

MHV mRNA 5 contains two potential ORFs coding for polypeptides of M_r 13,000 and M_r 9,600 though the sizes of the products vary between the two strains of MHV that have been sequenced. The two ORFs from the A 59 strain of MHV have been cloned into the pGEM vectors and the resulting RNA translated in the wheatgerm system [13]. Polypeptides of the correct size were synthesised but no products from in vitro translation of isolated viral poly(A)-containing mRNA were detected. The second ORF, encoding the polypeptide M_r 9,600, was fused to the 5′ end of the *lacZ* gene and the resulting chimaeric protein used to raise antibodies [27]. The antibodies raised against the chimaeric protein immune precipitated a polypeptide of M_r 9,600 in MHV infected cells. The first ORF, polypeptide M_r13,000, on mRNA 5 has a pyrimidine at the -3 position from the initiation codon and the second ORF, polypeptide M_r 9,600, has a purine at the -3 position. Thus it appears that in IBV and MHV, where a mRNA species contains more than one ORF, expression appears to occur from the ORF with the more favourable initiation codon. If TGEV is analogous, it appears that the ORF-2 gene is the more likely to be expressed, though the possibility that the other ORFs may be expressed but in much lower amounts, cannot be ruled out.

The mRNA 4 species of MHV contains a single ORF, encoding a polypeptide M_r 15,200, which has been fused to the *lacZ* gene and the resultant chimaeric protein used to raised antibodies [14, 15]. The antibodies raised against the chimaeric protein immune precipitated a polypeptide of M_r 15,000 in MHV infected cells and appeared to react with a protein in the cell cytoplasm by indirect immunofluorescence. The sequence context is favourable though the putative polymerase-leader complex binding site is 53 nucleotides upstream of the ATG, about six times the distance when compared to other MHV genes. The sequence context of TGEV ORF-3 is not favourable (see Table 1) though the distance between the polymerase-leader complex binding site is not unusually long, possibly indicating some control over expression.

Comparison of the sequences between the end of the peplomer and the start of the integral membrane protein genes derived from the FS 772/70 strain of TGEV reported in this paper to those of the Purdue strain [34] shows that there are several deletions and insertions within the cDNA sequences. Nucleotides 45–53 (Fig. 2) in the FS 772/70 strain are absent from the Purdue strain. At positions 71, 90, 343, and 1312 (Fig. 2) in the FS 772/70 strain there are 3 base, 16 base, 29 base, and 3 base insertions respectively in the Purdue strain when compared to the FS 772/70 strain. The differences at positions 45, 71, and 90 are in a non-coding region of the virus genome. However the insertions in the Purdue strain, at positions 342 and 1312 of the FS 772/70 sequence, are within ORF-1 and ORF-3. A base substitution at nucleotide 335 (Fig. 2) from G in the Purdue strain to a T in the FS 772/70 strain leads to earlier termination

of the ORF in the FS 772/70 strain. An insertion at position 342 and the base substitution at nucleotide 335 in the Purdue strain leads to an increase in the size of ORF-1 (X 2a Purdue strain) by 9 amino acids (Fig. 2). The insertion at position 1312 of three bases leads to the insertion of an extra isoleucine residue at amino acid position 55 in ORF-3 (X 1 for the Purdue strain), see Fig. 2. A change from a T residue in the Purdue strain to a G residue in FS 772/70 at nucleotide position 431 results in the formation of an ATG initiation codon for ORF-2. This results in a polypeptide of 244 amino acids for ORF-2 compared to X 2b of the Purdue strain which has 165 amino acids and initiates from nucleotide 666 on Fig. 2. Both the independently isolated clones, pTG11 and pTG22, from the FS 772/70 strain showed the same sequence and neither had the 13 base deletion described in some of the Purdue cDNA clones [34] which led to the truncation of the X 2b product. The homology between the amino acid sequences for the rest of the gene products between the two strains of TGEV is very high. There are only 3, 4, and 5 amino acid substitutions in ORFs 1–3, respectively, a similar result found for the nucleoprotein, integral membrane protein and ORF-4 gene products of these TGEV strains indicating that there has been very little change between the two viral genomes.

There is very little, if any, sequence homology between the TGEV ORFs 1–3 and the ORFs from mRNAs B and D from IBV and from mRNAs 4 and 5 from MHV using the SEQHPE program from the Los Alamos package [19]. However, IBV D 3, the ORF from MHV mRNA 4 and TGEV ORF-3 all have hydrophobic N-termini and hydrophilic C-termini and are of similar amino acid length indicating that they may have similar, but as yet unknown, function. ORF-B from MHV mRNA 5 and TGEV ORF-1 have very similar overall charges and have hydrophobic N-termini suggesting some similarity in function. Neither MHV or IBV have a protein of similar size to TGEV ORF-2 indicating that this gene product may be unique to TGEV. It will be interesting to compare the sequences of the other coronaviruses belonging to the TGEV sub-group when they become available for the presence of ORF-2. Comparison of TGEV ORFs 1–3 with the PIR protein database showed that there was no significant homology with any of the proteins in the databank using the FASTA program [31] or by using the SWEEP program against the Leeds University protein database. No homology was found by comparing the nucleotide sequences of ORFs 1–3 against the EMBL [17] or GENBANK [1] nucleic acid database using the FASTN program [28]. No homologies were found by screening the amino acid sequences of ORFs 1–3 against the derived amino acids from all the nucleic acid sequences in GENBANK using the TFASTA program [31].

It is constructive to compare the sequences between avirulent and virulent viruses in order to identify regions that may be involved in pathogenicity and immunogenicity. Evidence presented in this paper indicates that there is very little homology between potential non-structural genes from the different coronaviruses. There appears to be a significant difference between the sequence of the ORF-2 gene from virulent strain, described in this paper, and the avirulent

Purdue strain of TGEV previously published [34]. However, a polypeptide of M_r 24,000 has been detected by in vitro translation of the 3.4 kb TGEV mRNA species, from the Purdue strain, using the rabbit reticulocyte lysate system [18]. Since the nucleotide sequence of this region has not been published, for the isolate of the Purdue strain that was used for in vitro translation, no conclusions can be reached with respect to variation within isolates of the Purdue strain. Experiments are under way to identify the products of ORFs 1–3 in virus infected cells.

Acknowledgements

This research was supported by the Biomolecular Engineering Programme of the Commission of the European Communities Contract No [BAP-0235-UK (HI)] at the Compton Laboratory and Contract No [BAP-0219-E(A)] at the Departamento de Biologia Funcional (Area de Bioquimica). This work, in Spain, was also supported by grant N° CE 96-0003 from C.A.I.C.Y.T. One of us, PB, would like to thank Miss K. Mawditt for synthesising oligos 8 and 15 and for her excellent technical assistance.

References

1. Bilofsky HS, Burks C, Fickett JW, Goad WB, Lewitter FI, Rindone WP, Swindell CD, Tung C-S (1986) The Genbank genetic sequence database. Nucleic Acids Res 13: 1–4
2. Bohl EH, Gupta RKP, Olquin MV, Saif LJ (1972) Antibody responses in serum, clostrum, and milk of swine after infection or vaccination with transmissible gastroenteritis virus. Infect Immun 6: 289–301
3. Boursnell MEG, Brown TDK (1984) Sequencing of coronavirus IBV genomic RNA: a 195-base open reading frame encoded by mRNA B. Gene 29: 87–92
4. Boursnell MEG, Binns MM, Brown TDK (1985) Sequencing of coronavirus IBV genomic RNA: Three open reading frames in the 5′ 'unique' region of mRNA D. J Gen Virol 66: 2253–2258
5. Boursnell MEG, Brown TDK, Foulds IJ, Green PF, Tomley FM, Binns MM (1987) Completion of the sequence of the genome of the coronavirus avian infectious bronchitis virus. J Gen Virol 68: 57–77
6. Britton P, Lee LG, Murfitt D, Boronat A, Jones-Mortimer MC, Kornberg HL (1984) Location and direction of transcription of the *ptsH* and *ptsI* genes on the *Escherichia coli* K 12 genome. J Gen Microbiol 130: 861–868
7. Britton P, Garwes DJ, Millson GC, Page K, Bountiff L, Stewart F, Walmsley J (1986) Towards a genetically-engineered vaccine against porcine transmissible gastroenteritis virus. In: Magnien E (ed) Biomolecular engineering in the European Community. Martinus Nijhoff, Dordrecht, pp 301–313
8. Britton P, Garwes DJ, Page K, Walmsley (1987) Expression of porcine transmissible gastroenteritis virus genes in *E. coli* as β-galactosidase chimaeric proteins. In: Lai MMC, Stohlman SA (eds) Coronaviruses. Plenum Press, New York London [Advances in experimental medicine and biology, vol 218, pp 55–64]
9. Britton P, Carmenes RS, Page KW, Garwes DJ, Parra F (1988a) Sequence of the nucleoprotein from a virulent British field isolate of transmissible gastroenteritis virus and its expression in *Saccharomyces cerevisiae*. Mol Microbiol 2: 89–99
10. Britton P, Carmenes RS, Page KW, Garwes DJ (1988b) The integral membrane protein from a virulent isolate of transmissible gastroenteritis virus: molecular characterization, sequence and expression in *Escherichia coli*. Mol Microbiol 2: 497–505

11. Brown TDK, Boursnell MEG, Binns MM (1984) A leader sequence is present on mRNA A of avian infectious bronchitis virus. J Gen Virol 65: 1437–1442
12. Budzilowicz CJ, Wilczynski SP, Weiss SR (1985) Three intergenic regions of coronavirus mouse hepatitis virus strain A 59 genome RNA contain a common nucleotide sequence that is homologous to the 3′ end of the viral mRNA leader sequence. J Virol 53: 834–840
13. Budzilowicz CJ, Weiss SR (1987) In vitro synthesis of two polypeptides from a non-structural gene of coronavirus mouse hepatitis virus strain A 59. Virology 157: 509–515
14. Ebner D, Siddell S (1987) Identification of the coronavirus MHV-JHM mRNA 4 gene product using fusion protein antisera. In: Lai MMC, Stohlman SA (eds) Coronaviruses. Plenum Press, New York London [Advances in experimental medicine and biology, vol 218, pp 39–45]
15. Ebner D, Raabe T, Siddell SG (1988) Identification of the coronavirus MHV-JHM mRNA 4 product. J Gen Virol 69: 1041–1050
16. Garwes DJ, Pocock DH (1975) The polypeptide structure of transmissible gastroenteritis virus. J Gen Virol 29: 25–34
17. Hamm GH, Cameron GN (1986) The EMBL data library. Nucleic Acids Res 14: 5–10
18. Jacobs L, van der Zeijst BAM, Horzinek MC (1986) Characterization and translation of transmissible gastroenteritis virus mRNAs. J Virol 57: 1010–1015
19. Kanehisa MI (1982) Los Alamos sequence analysis package for nucleic acids and proteins. Nucleic Acids Res 10: 183–196
20. Kapke PA, Brian DA (1986) Sequence analysis of the porcine transmissible gastroenteritis coronavirus nucleocapsid protein gene. Virology 151: 41–49
21. Kapke PA, Tung FYC, Brian DA, Woods RD, Wesley R (1987) Nucleotide sequence of the porcine transmissible gastroenteritis coronavirus matrix protein. In: Lai MMC, Stohlman SA (eds) Coronaviruses. Plenum Press, New York London [Advances in experimental medicine and biology, vol 218, pp 117–122]
22. Klenk HD, Rott R (1980) Cotranslational and posttranslational processing of viral glycoproteins. Curr Top Microbiol Immunol 90: 19–48
23. Kozak M (1983) Comparison of initiation of protein synthesis in prokaryotes, eukaryotes and organelles. Microbiol Rev 47: 1–45
24. Kozak M (1986) Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes. Cell 44: 283–292
25. Lai MMC, Baric RS, Brayton PR, Stohlman SA (1984) Characterization of leader RNA sequences on the virion and mRNAs of mouse hepatitis virus, a cytoplasmic RNA virus. Proc Natl Acad Sci USA 81: 3626–3630
26. Laude H, Rasschaert D, Huet JC (1987) Sequence and N-terminal processing of the transmembrane protein E 1 of the coronavirus transmissible gastroenteritis virus. J Gen Virol 68: 1687–1693
27. Leibowitz JL, Perlman S, Weinstock G, DeVries JR, Budzilowicz C, Weissmann JM, Weiss SR (1988) Detection of a murine coronavirus nonstructural protein encoded in a down stream open reading frame. Virology 164: 156–164
28. Lipman DJ, Pearson WR (1985) Rapid and sensitive protein similarity searches. Science 227: 1435–1441
29. Maniatis T, Fritsch EF, Sambrook J (1982) Molecular cloning: a laboratory manual. Cold Spring Harbor Laboratory, New York
30. Murphy G, Kavanagh T (1988) Speeding-up the sequencing of double-stranded DNA. Nucleic Acids Res 16: 5198
31. Pearson WR, Lipman DJ (1988) Improved tools for biological sequence comparison. Proc Natl Acad Sci USA 85: 2444–2448

32. Rasschaert D, Laude H (1987) The predicted structure of the peplomer protein E 2 of the porcine coronavirus transmissible gastroenteritis gastroenteritis virus. J Gen Virol 68: 1883–1890
33. Rasschaert D, Delmas B, Charley B, Grossclaude J, Gelfi J, Laude H (1987a) Surface glycoproteins of transmissible gastroenteritis virus: functions and gene sequence. In: Lai MMC, Stohlman SA (eds) Coronaviruses. Plenum Press, New York London [Advances in experimental medicine and biology, vol 218, pp 109–116]
34. Rasschaert D, Gelfi J, Laude H (1987b) Enteric coronavirus TGEV: partial sequence of the genomic RNA, its organisation and expression. Biochemie 69: 591–600
35. Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain terminating inhibitors. Proc Natl Acad Sci USA 74: 5463–5467
36. Shieh C-K, Soe LH, Makino S, Chang M-F, Stohlman SA, Lai MMC (1987) The 5'-end sequence of the murine coronavirus genome: implications for multiple fusion sites in leader-primed transcription. Virology 156: 321–330
37. Skinner MA, Siddell SG (1985) Coding sequence of coronavirus MHV-JHM mRNA. J Gen Virol 66: 593–596
38. Skinner MA, Ebner D, Siddell SG (1985) Coronavirus MHV-JHM mRNA 5 has a sequence arrangement which potentially allows translation of a second down stream open reading frame. J Gen Virol 66: 581–592
39. Smith AR, Boursnell MEG, Binns MM, Brown TDK, Inglis SC (1987) Identification of a new gene product encoded by mRNA D of infectious bronchitis virus. In: Lai MMC, Stohlman SA (eds) Coronaviruses. Plenum Press, New York London [Advances in experimental medicine and biology, vol 218, pp 47–54]
40. Spaan WJM, Delius H, Skinner M, Armstrong J, Rottier P, Smeekens S, van der Zeijst BAM, Siddell SG (1983) Coronavirus mRNA synthesis involves fusion of non-contiguous sequences. EMBO J 2: 1839–1844
41. von Heijne G (1986) A new method for predicting signal sequence cleavage sites. Nucleic Acids Res 14: 4683–4690

Authors' address: Dr. P. Britton, Division of Microbiology, A.F.R.C. Institute for Animal Health, Compton Laboratory, Compton, Newbury, Berks. RC16 ONN, U.K.

Received February 9, 1989