# Cloning and sequencing of a 8.4-kb region from the 3'-end of a Taiwanese virulent isolate of the coronavirus transmissible gastroenteritis virus

Chi-Min Chen [1], David Cavanagh, Paul Britton [*]

*Division of Molecular Biology, Institute for Animal Health, Compton, Newbury, Berkshire RG20 7NN, UK*

## Abstract

The nucleotide sequence (8396 nucleotides) was determined, from the 3'-end of the putative polymerase gene to the poly(A) tail, for a Taiwanese virulent isolate, TFI, of transmissible gastroenteritis virus (TGEV). The TFI nucleotide sequence had very high identity to the British virulent field isolate FS772/70 (98.3%), the attenuated Purdue 115 (96.7%) and from the S gene to ORF-4 gene region, to the low passaged virulent Miller (98.3%) strains of TGEV. Comparison of the TFI S protein sequence with those determined from other TGEV strains and those of the TGEV variant, porcine respiratory coronavirus, isolated from Europe and North America showed that they had changed very little over a period of 4 decades. The two extra amino acids found to be present in the spike proteins of the virulent FS772/70 and Miller strains when compared to the avirulent Purdue strain were found to be present in the TFI strain. The genomic organisation of the TFI strain was the same as that of the other TGEV viruses.

*Keywords:* Transmissible gastroenteritis virus (TGEV); Coronavirus; Porcine; Swine

Transmissible gastroenteritis virus (TGEV) is an enteropathogenic porcine virus that preferentially replicates in and destroys the enterocytes covering the tips of the villi in the small intestine, causing diarrhoea and dehydration in pigs resulting in high mortality in neonates. The virus is a member of the Coronavirus genus

[*] Corresponding author: Fax: +44 (1635) 577263. E-mail: Britton@BBSRC.AC.UK.
[1] Present address: Department of Pathology, Pig Research Institute Taiwan, P.O. Box 23, Chunan, Miaoli, Taiwan, Republic of China.

family Coronaviridae and has a single-stranded, positive-sense RNA genome of $\approx 28,000$ nucleotides. TGEV virions contain at least 4 proteins: a phosphorylated nucleoprotein (N; $M_r$ 47,000) and 3 membrane proteins, the integral membrane glycoprotein (M; $M_r$ 28,000–31,000), the spike glycoprotein (S; $M_r$ 200,000) and the small membrane protein (sM; $M_r$ 9200).

Sequence data obtained for several strains of TGEV and the variant virus porcine respiratory coronavirus (PRCV) have identified sequence polymorphism within 3 regions of the genome. Two regions are located in the 5′-end of the S protein gene. One region consists of a 6 nucleotide (nt) deletion in some attenuated strains of TGEV, namely Purdue 115 (Rasschaert and Laude, 1987) and Nebraska 72 (Sánchez et al., 1992), when compared to other TGEV (Britton and Page, 1990; Wesley, 1990) and PRCV (Rasschaert et al., 1990; Britton et al., 1991; Wesley et al., 1991) S sequences. The other region in the S protein gene consists of 672 or 681 nt deleted from the PRCV S protein gene when compared to the TGEV sequence (Rasschaert et al., 1990; Britton et al., 1991; Wesley et al., 1991). The third region is within and downstream of the ORF-3a gene and in some cases extends to within the ORF-3b gene (Rasschaert et al., 1990; Wesley et al., 1990; Page et al., 1991; Britton et al., 1993). Sequence analysis of all PRCV strains to date has identified that a deletion within the region of the genome corresponding to the putative TGEV ORF-3a gene has converted it to a pseudogene (Page et al., 1991; Rasschaert et al., 1990; Wesley et al., 1991). However, a gene equivalent to TGEV ORF-3b is intact in PRCV. A small plaque variant of the Miller TGEV strain has 462 nucleotides deleted downstream of the S gene resulting in the loss of mRNA-3 and consequently ORFs 3a and 3b when compared to the parent Miller strain (Wesley et al., 1990). A small plaque and acid-resistant attenuated TGEV strain, 188-SG, derived from a virulent French isolate, D52, was found to have 268 nucleotides deleted from within ORF-3a to within the 5′-end of ORF-3b (Britton et al., 1993). Thus those viruses with a small plaque phenotype and shown to be avirulent in pigs have been shown to contain a deletion within the ORF-3a/3b region indicating that these genes may be involved in pathogenicity, but are not required for growth in vitro or in vivo.

Work presented in this paper describes sequence data obtained for a virulent field strain, originally isolated in 1983 at Pingtong in Southern Taiwan from the duodenum of a piglet with severe diarrhoea. The virus was passaged 50 times on swine testis (ST) cells and shown to be virulent in pigs (Chen et al., 1989). Previous strains of TGEV that have been sequenced had been isolated between 1946 and 1972 and had predominantly originated either from the USA or Western Europe, with the exception of one high passaged strain isolated in 1956 from Japan. The sequence data available for the PRCV strains are from viruses isolated in the mid 1980s from Western Europe. Thus sequence information from the TFI strain not only represented data for another virulent field isolate, but also of a virus from a later period than the other TGEV strains and which was also from Asia.
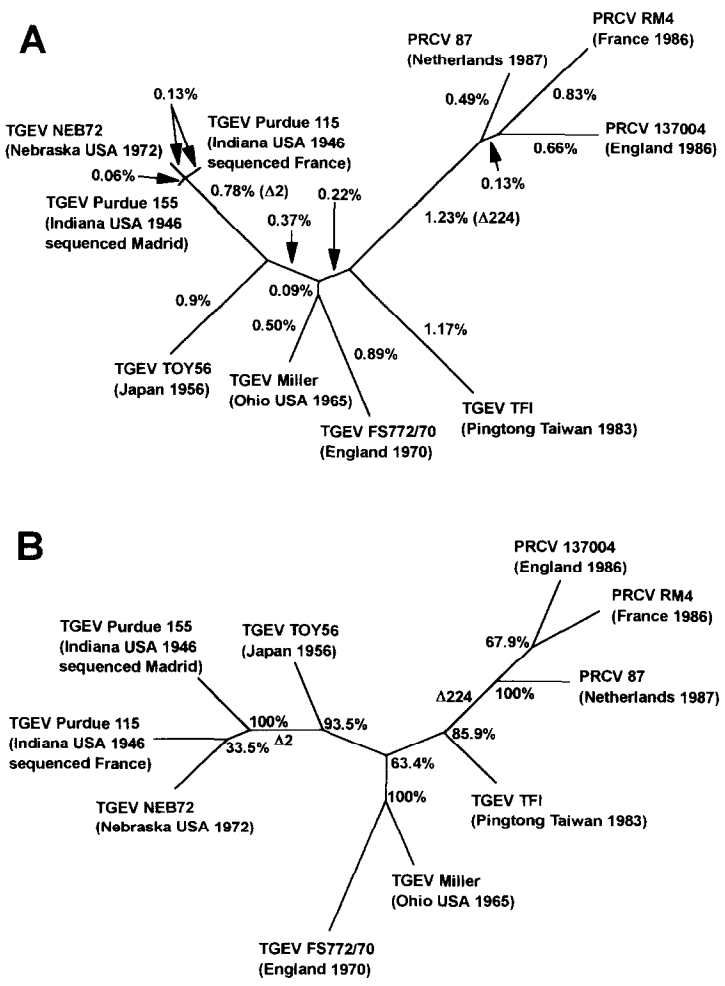
The TFI strain was grown in ST cells in Taiwan and virus samples, corresponding to $10^9$ PFU, lyophilised in glass vials for transportation to the UK. TFI genomic RNA was prepared from the lyophilised virus and TFI cDNAs generated

from TFI RNA by RT-PCR using SuperScript™ RNase H⁻ reverse transcriptase (Gibco BRL), *Taq* polymerase (Boehringer Mannheim) and several sets of oligonucleotides derived from FS772/70 sequence data. The RT-PCR products were purified from agarose gels, end-repaired, ligated into the *Sma*I site of pGEM®-3Z (Promega) and transformed into DH5α competent *Escherichia coli* cells (Gibco BRL). Transformants containing TFI cDNA were identified by colony hybridization using ³²P-labelled oligonucleotides derived from FS772/70 TGEV sequence data. Plasmid DNA was isolated, purified by CsCl/ethidium bromide centrifugation and the cDNA inserts sequenced by the dideoxy chain termination method using Sequenase™ (United States Biochemical Corporation) and oligonucleotides derived from FS772/70.

Sequence data obtained from 12 overlapping TFI cDNA clones were compiled using the Staden package (Staden, 1982) to produce a consensus sequence of 8396 bp (EMBL/GenBank/DDBJ accession number Z35758). Analysis of the sequence data, using the University of Wisconsin Genetics Computer Group (UWGCG) version 7 programs (Devereux et al., 1984) identified 8 potential ORFs; ORF-1 (nt 3–248; $M_r$ 9266), ORF-2 (S; nt 245–4594; $M_r$ 160,117), ORF-3a (nt 4694–4912; $M_r$ 7847), ORF-3b (nt 4937–5671; $M_r$ 27,693), ORF-4 (sM; nt 5658–5903; $M_r$ 9229), ORF-5 (M; nt 5914–6702; $M_r$ 29,479), ORF-6 (N; nt 6715–7863; $M_r$ 43,559) and ORF-7 (nt 7869–8105; $M_r$ 9111). Due to the position on the consensus sequence, the fact that it lacked a 5'-initiation codon and similarity to other TGEV and PRCV sequences, ORF-1 was identified as the 3'-end of the TFI polymerase gene. The other ORFs were identified by comparison with sequences observed for other TGEVs. All the ORFs, except the polymerase gene and the ORF-3b gene, were preceded by the sequence, CTAAAC, predicted to be involved in the synthesis of TGEV subgenomic mRNAs.

Comparison of the TFI consensus sequence with sequences obtained from other TGEV strains identified that 38 nt, position 4928 on the TFI sequence, were deleted from the non-coding region upstream of the ORF-3b in TFI. No other differences, apart from various base substitutions, were identified between the TFI and other TGEV sequences. Interestingly the 3'-end of the TFI ORF-3a, a region that shows polymorphism amongst the viruses sequenced to date, had the same sequence as obtained for the Miller strain. Comparison of the amino acid sequences deduced from the TFI genes with those from other viruses showed that there was a high identity between the various TFI gene products and those of the other isolates of TGEV and PRCV. The TFI S protein sequence contained the two amino acids, residues 375 and 376, deleted from the avirulent Purdue-115 and Nebraska 72 strains, when compared to other TGEV and PRCV S sequences. The linear monoclonal antibody binding site SSFFSYGEI (site C of Delmas et al., 1990 or site D of Correa et al., 1990 and Posthumus et al., 1990) identified on the Purdue S sequence, but absent in the FS772/70, Miller and PRCV 137004, RM4 and 87 sequences due to a Phe → Ser amino acid substitution at the second phenylalanine residue was also absent from the TFI S sequence due to the same substitution as for other TGEVs. The S sequences of the avirulent TGEV strains NEB72 and TOY56 do not have this substitution (Sánchez et al., 1992).

To determine any correlation between the virulent and avirulent viruses a clustal analysis was performed on the published S protein amino acid sequences of TGEV and PRCV. The differences analysed are based on true amino acid substitutions and that the amino acids in the TGEV S sequences deleted from PRCV cannot be taken into account. The results of a distance based analysis and a maximum parsimony analysis are shown in Fig. 1A and B, respectively. Both analyses only represent clustering of the sequences due to similar or different amino acid changes without any evolutionary significance attached. The analyses showed that although the S sequences were very similar (identities ranging from 95.8 to 98.0%) amino acid substitutions between the sequences resulted in the viruses forming some preliminary groupings. The individual virus sequences di-

verged from common consensus sequences by varying degrees. For example, the TFI sequence diverged from the FS772/70 and Miller consensus sequence by 0.31% (0.09 + 0.22%; Fig. 1A), representing 4 amino acid substitutions. These 4 amino acid substitutions were also present in the PRCV sequences. The TFI sequence then contains a further unique change of 1.17% (equivalent to 17 amino acid substitutions) from the FS772/70 and Miller consensus sequence. The S sequences of the avirulent strains of TGEV diverged from the virulent FS772/70 and Miller strain consensus sequence by one amino acid in common with TFI and the PRCV sequences and then by a further 0.37% different to TFI and the PRCV consensus sequence. It is interesting from the data that viruses isolated over a period of 40 years have only diverged by 2.67% (TFI → Purdue 115; Fig. 1A), equivalent to 39 amino acid substitutions. This is not a feature of coronaviruses as a group. Infectious bronchitis virus (IBV), which replicates in the mucosa of the respiratory tract and other organs, exhibits substantial amino acid variation (for references see Wang et al., 1994). Isolates of IBV commonly differ by 20–25% in the N-terminal (S1) half of the S protein. Even closely related strains, isolated within 5 years of each other, can differ by 3%.

If amino acid changes in the S protein are responsible for loss of virulence, it is possible that the amino acid substitution(s) resulting in virulence could be part of the difference observed between the avirulent and virulent virus consensus sequences. However, loss of virulence could be associated with changes unique to each of the avirulent viruses. Interestingly, the PRCV strains did not contain the same amino acid substitutions, potentially involved in loss of virulence, observed for the avirulent TGEV strains. Moreover, changes or a combination of changes in another TGEV gene(s) might be associated with virulence.

Fig. 1. Clustal analysis of the S protein sequences of the TGEV and PRCV S protein sequences based on a distance analysis (A) and parsimony (B). The sequences were aligned, using the PAM 250 matrix, using the CLUSTALV programme (Higgins and Sharp, 1988) and analysed using the programs in the Phylogeny Inference Package (PHYLIP) version 3.52c (Joseph Felsenstein, Department of Genetics, University of Washington, Seattle, 1993; Felsenstein, 1988). PROTDIST (using the tables of Eck and Dayhoff, 1966) followed by FITCH (using the 'global' option and shuffling of input order; Fitch and Margoliash, 1967) for the distance-based analysis and PROTPARS for the maximum parsimony analysis. The topologies for the parsimony method were randomly resampled (bootstrapped using SEQBOOT) and a consensus tree (using CONSENSE) determined from a 100 datasets of the aligned sequences. Any bias, potentially resulting from the order in which the sequences were read into the programs, was removed by randomising the input order for each dataset. All the trees produced were unrooted and drawn using DRAWTREE. The branches in A show the percentage of amino acid substitutions (based on 1449 amino acids) between sequences or groups of sequences. The total number of amino acid substitutions between viruses can be calculated from the percentages along the branches separating the sequences. The percentages at the forks in B indicate the number of times the group(s) beyond the fork, occurred amongst the trees from the 100 datasets. The sequences were Miller (Wesley, 1990), Nebraska 72 (NEB72), TOY56, Purdue 115 (Madrid) and PRCV 87 (Sánchez et al., 1992), Purdue 115 (France; Rasschaert and Laude, 1987), FS772/70 (Britton and Page, 1990), PRCV 137004 (Britton et al., 1991) and PRCV RM4 (Rasschaert et al., 1990). The places and year of isolation of the viruses are indicated in brackets. Branches involving the Δ2 and Δ224 amino acid deletions are indicated.

## Acknowledgements

## References

Britton, P. and Page, K.W. (1990) Sequence of the S gene from a virulent British field isolate of transmissible gastroenteritis virus. Virus Res. 18, 71–80.

Britton, P., Mawditt, K.L. and Page, K.W. (1991) The cloning and sequencing of the virion protein genes from a British isolate of porcine respiratory coronavirus: comparison with transmissible gastroenteritis virus genes. Virus Res. 21, 181–198.

Britton, P., Kottier, S., Chen, C.-M., Pocock, D.H., Salmon, H. and Aynaud, J.M. (1993) The use of PCR genome mapping for the characterisation of TGEV strains. In: H. Laude and J.F. Vautherot (Eds.), Coronaviruses: Molecular Biology and Virus–Host Interactions, Adv. Exp. Med. Biol. Vol. 342, Plenum Press, New York, pp. 29–34.

Chen, C.M., Tien, S.C., Wu, J.L. and Chu, R.M. (1989) The phenotype of Taiwan field isolated transmissible gastroenteritis virus (TGEV) is different from the other TGEV. J. Chin. Soc. Vet. Sci. 15, 11–18.

Correa, I., Gebauer, F., Bullido, M.J., Sune, C., Baay, M.F.D., Zwaagstra, K.A., Posthumus, W.P.A., Lenstra, J.A. and Enjuanes, L. (1990) Localisation of antigenic sites of the E2 glycoprotein of transmissible gastroenteritis virus. J. Gen. Virol. 71, 271–279.

Delmas, B., Rasschaert, D., Godet, M., Gelfi, J. and Laude, H. (1990) Four Major Antigenic Sites of the Coronavirus Transmissible Gastroenteritis Virus Are Located on the Amino-Terminal Half of Spike Glycoprotein S. J. Gen. Virol. 71, 1313–1323.

Devereux, J., Haeberli, P. and Smithies, O. (1984) A comprehensive set of sequence analysis programs for the VAX. Nucleic Acids Res. 12, 387–395.

Eck, R.V. and Dayhoff, M.O. (1966) Atlas of Protein Sequence and Structure 1966. National Biomedical Research Foundation, Maryland, Silver Spring.

Felsenstein, J. (1988) Phylogenies from molecular sequences: inferences and reliability. Am. Rev. Genet. 22, 512–565.

Fitch, W.M. and Margoliash, E. (1967) Construction of phylogenetic trees. Science 155, 279–284.

Higgins, D.G. and Sharp, P.M. (1988) CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. Gene 73, 237–244.

Page, K.W., Mawditt, K.L. and Britton, P. (1991) Sequence comparison of the 5' end of mRNA 3 from transmissible gastroenteritis virus and porcine respiratory coronavirus. J. Gen. Virol. 72, 579–587.

Posthumus, W.P.A., Lenstra, J.A., Schaaper, W.M.M., van Nieuwstadt, A.P., Enjuanes, L. Meloen, R.H. (1990) Analysis and simulation of a neutralising epitope of transmissible gastroenteritis virus. J. Virol. 64, 3304–3309.

Rasschaert, D. and Laude, H. (1987) The predicted primary structure of the peplomer protein E2 of the porcine coronavirus transmissible gastroenteritis virus. J. Gen. Virol. 68, 1883–1890.

Rasschaert, D., Duarte, M. and Laude, H. (1990) Porcine respiratory coronavirus differs from transmissible gastroenteritis virus by a few genomic deletions. J. Gen. Virol. 71, 2599–2607.

Sánchez, C.M., Gebauer, F., Suñé, C., Mendez, A., Dopazo, J. and Enjuanes, L. (1992) Genetic evolution and tropism of transmissible gastroenteritis coronavirus. Virology 190, 92–105.

Staden, R. (1982) Automation of the computer handling of gel reading data produced by the shotgun method of DNA sequencing. Nucleic Acids Res. 10, 4731–4751.

Wang, L., Junker, D., Hock, L., Ebiary, E., Collisson, E.W. (1994) Evolutionary implications of genetic variations in the S1 gene of infectious bronchitis virus. Virus Res. 34, 327–338.

Wesley, R.D. (1990) Nucleotide sequence of the E2-peplomer protein gene and partial nucleotide sequence of the upstream polymerase gene of transmissible gastroenteritis virus (Miller strain). In: D. Cavanagh and T.D.K. Brown (Eds.), Coronaviruses and their Diseases. Adv. Exp. Med. Biol, Vol. 276, Plenum Press, New York, pp. 301–306.

Wesley, R.D., Woods, R.D. and Cheung, A.K. (1990) Genetic-basis for the pathogenesis of transmissible gastroenteritis virus. J. Virol. 64, 4761–4766.

Wesley, R.D., Woods, R.D. and Cheung, A.K. (1991) Genetic analysis of porcine respiratory coronavirus, an attenuated variant of transmissible gastroenteritis virus. J. Virol. 65, 3369–3373.