Journal of General Virology (2007), **88**, 487–492 DOI 10.1099/vir.0.82555-0

Short Communication

Mosaic structure of foot-and-mouth disease virus genomes

A. L. Jackson,[1]† H. O'Neill,[2]‡ F. Maree,[2] B. Blignaut,[2] C. Carrillo,[3] L. Rodriguez[3] and D. T. Haydon[1]

Correspondence
D. T. Haydon
D.Haydon@bio.gla.ac.uk

[1]Division of Environmental and Evolutionary Biology, University of Glasgow, Glasgow G12 8QQ, UK

[2]Onderstepoort Veterinary Institute, Exotic Diseases Division, Private Bag X05, Onderstepoort 0010, South Africa

[3]Agricultural Research Service, USDA, Plum Island Animal Disease Center, PO Box 848, Greenport, NY 11944, USA

Received 8 September 2006
Accepted 30 October 2006

The results of a simple pairwise-scanning analysis designed to identify inter-serotype recombination fragments, applied to genome data from 156 isolates of *Foot-and-mouth disease virus* (FMDV) representing all seven serotypes, are reported. Large numbers of candidate recombinant fragments were identified from all parts of the FMDV genome, with the exception of the capsid genes, within which such fragments are infrequent. As expected, intertypic fragment exchange is most common between geographically sympatric FMDV serotypes. After accounting for the likelihood of intertypic convergence in highly conserved parts of the FMDV genome, it is concluded that intertypic recombination is probably widespread throughout the non-structural genes, but that recombination over the 2B/C and 3B/C gene boundaries appears to be less frequent than expected, given the large numbers of recombinant gene fragments arising in these genes.

Foot-and-mouth disease virus (FMDV) is a single-stranded, positive-sense RNA virus in the family *Picornaviridae* and the cause of an economically important disease of cloven-hoofed animals. Seven different serotypes of FMDV exist: A, O, C, Asia 1, SAT 1, 2 and 3. The genome of FMDV consists of approximately 8200 nt excluding the poly(C) and (A) tracts and comprises a 5′ untranslated region of about 1150 nt, followed by a single open reading frame (ORF) of approximately 6890 nt and a 3′ untranslated region of approximately 160 nt. The ORF encodes 12 proteins: the leader protease (L^{pro}), four capsid proteins (1A–D) and the non-structural proteins 2A–C and 3A–D.

The geographical distributions of FMDV serotypes have probably undergone a series of expansions and contractions over the last century. Current notions of the distribution of FMDV serotypes are subject to under-reporting, particularly in Africa (Kivaria, 2003). However, it is clear that SAT serotypes have been and remain almost exclusively restricted to Africa (Vosloo *et al.*, 2002) and are thought to have originated in sub-Saharan Africa, where they circulate mainly among African buffalo (*Syncerus caffer*). SAT serotypes are also found in north Africa (Vosloo *et al.*, 2002) and SAT 1 and 2 make occasional incursions into the Middle East (Aidaros, 2002). Serotypes A, O and C used to arise commonly in Europe prior to the introduction of widespread vaccination, and have been reported to circulate in many African countries, the Middle East, southern Asia, the Far East and South America (Aidaros, 2002; Correa Melo *et al.*, 2002; Vosloo *et al.*, 2002), although reports of type C are increasingly infrequent. Asia 1 appears to be restricted to southern and eastern Asia. The historical distributions of FMDV serotypes (i.e. prior to approx. 1900) are not known with any certainty. Indeed, the whole subject of the origin of FMDV serotypes remains unstudied.

Where FMDV circulates among what are thought to be ancestral host populations (for example, SAT serotypes in African buffalo), they cause subclinical infections that can last for several years, and co-infection with multiple serotypes is encountered frequently (Hedger, 1972; Hedger *et al.*, 1972). In domestic livestock, the disease tends to cause more acute and shorter-lived infections, but often the immune system fails to clear virus, which can linger as a persistent infection in so-called 'carrier' animals

†Present address: Trinity Centre for Bioengineering, School of Engineering, Trinity College, Dublin 2, Ireland.

‡Present address: Division of Medical Biochemistry, Institute of Infectious Diseases and Molecular Medicine, University of Cape Town, Observatory 7925, South Africa.

A supplementary table showing FMDV sequence data used in this study is available in JGV Online.

for up to 2 years and from which transmission is thought to be relatively rare (Sutmoller *et al.*, 2003).

The vast majority of infections of livestock will comprise almost-identical genotypes of the same serotype, making the identification of recombination events very difficult. However, in natural hosts, such as African buffalo, and in areas where multiple serotypes co-circulate, opportunities for intertypic recombination will arise more often; recombinants were first identified in the field by Krebs & Marquardt (1992). Analyses of closely related strains of FMDV have demonstrated lower levels of linkage disequilibrium than anticipated assuming no recombination, suggesting that recombination during the course of such infections is likely to be common (Haydon *et al.*, 2004). Recombination sites have been identified previously in the non-structural genes at the 3′ end of the genome (King *et al.*, 1985; McCahon *et al.*, 1977, 1985; McCahon & Slade, 1981; Wilson *et al.*, 1988) and less frequently among the structural genes (Tosh *et al.*, 2002a, b). As expected, phylogenetic relationships among homologous structural genes reflect the serotype classification of FMDV closely, but the phylogenetic relationships among homologous non-structural genes are often quite different and map poorly onto the serotype classification. This observation has led to the suggestion that recombination may have occurred more often among non-structural than structural genes (Carrillo *et al.*, 2005; van Rensburg *et al.*, 2002).

The opportunistic and idiosyncratic nature of the sampling process by which genomic data have accumulated limits the value of enumerating exactly how often such recombination has occurred in existing sequence data, although sophisticated and complex methodologies are available to address this question (Simmonds & Midgley, 2005; Yang *et al.*, 2006). However, the increasing number of full-genome and multi-gene sequences of FMDV subtypes across all seven serotypes now permits a comprehensive description of the extent to which FMDV genomes of each serotype comprise genomic regions related most closely to different serotypes. Our objective here was to map the genome locations of these exchangeable regions and to determine over what genomic length scales they typically occur and between which serotypes exchange has arisen. The existence of such genomic 'mosaics' is evidence either of gene convergence or horizontal genetic transfer – recombination – between serotypes. Here, we applied a simple and transparent algorithm for identifying inter-serotypic convergence or recombinant regions of different lengths, based on pairwise identity measures, to 156 genomic sequences of FMDV. The results reveal widespread mosaic structure across different serotypes among almost all of the non-structural genes of FMDV.

We examined 156 sequences of FMDV genomes from all seven serotypes [33 partial and 123 complete: A, $n=49$ sequences; O, $n=43$; C, $n=9$; Asia 1, $n=8$; SAT 1, $n=21$; SAT 2, $n=18$; and SAT 3, $n=8$; a full list of GenBank accession numbers is provided in Supplementary Table S1 (available in JGV Online)] for recombination events by using a simple method that looks for genomic fragments that have a higher nucleotide identity with sequences from other serotypes than with sequences within the same serotype. The sequences were initially split into non-coding and coding regions and aligned in BioEdit (Hall, 1999) based on nucleotide and amino acid sequences, respectively, before being converted back to nucleotide sequences. The algorithm considered one sequence at a time, designated the 'target' sequence, and searched all other sequences of different serotypes to find potential 'donor' sequences from which contiguous fragments of RNA might have been transferred through recombination.

The algorithm began at the 5′ end of the sequences with a window of 100 nt, chosen in order to avoid searches triggered by commonly used amino acid motifs (however, the results are qualitatively robust to alternative choices of this initial window size), and made a series of pairwise comparisons between the target sequence and all other sequences. Similarity between two sequences across a window was defined simply as the number of positional nucleotide matches in the window. When the alignment showed a gap–gap match between two sequences of the same serotype, it was registered as an identity (1), whereas a gap–gap match between serotypes was registered as a difference (0). The third-nucleotide codon position in the coding regions was excluded from all aspects of the analysis.

The target sequence was compared with each sequence from the same serotype as itself and the nucleotide similarity for each pairwise comparison, S_{within}, was calculated. The highest of these similarities was designated S^*_{within}. A similar comparison was made between the target sequence and all potential donor sequences of different serotype to yield a set of between-serotype similarity scores, S_{between}. The highest of these similarities was designated S^*_{between}. If $S^*_{\text{within}} > S^*_{\text{between}}$, then the comparison window was advanced by one nucleotide position and the test was recalculated. If $S^*_{\text{within}} \leqslant S^*_{\text{between}}$, then a search was initiated to find a potentially wider window that maximized an identity function, C, defined as $C=(\text{window size}) \times (S^*_{\text{between}}+1)/(S^*_{\text{within}}+1)$ conditional on $S^*_{\text{within}} \leqslant S^*_{\text{between}}$, so that longer windows with greatest similarity between serotypes were identified. This test was performed for all pairwise comparisons involving the target and potential donors from other serotypes. This window-maximizing algorithm considered all possible start positions within the initially located window, spanning 100 nt, and considered all possible window lengths up to the end of the genome. Fragments that maximized C were termed candidate recombination fragments (CRFs). When donor sequences were found from more than one serotype, the serotype with the highest proportion of donor sequences was selected. Ties within serotypes were broken through simple equiprobable random selection.

Following identification of a CRF, two further tests were performed. Firstly, a bootstrapping process was performed wherein 500 alignments were constructed by randomly

selecting nucleotide positions with replacement across all 156 sequences from within the identified window until the length of the bootstrapped sequences matched the length of the identified window. The proportion of these alignments that still satisfied the condition of $S^*_{within} \leqslant S^*_{between}$ for target and donor sequences was recorded. Secondly, the mean cross-serotype similarity score was computed across the identified window between all pairwise comparisons of sequences of the target and donor serotype. This index provides a measure of conservation of the genome between the two serotypes over the identified window. Upon completion of these tests, the search for new fragments was resumed, starting with the nucleotide position immediately adjacent to the 3′ end of the identified window. The analyses were performed in MATLAB and code is available from the authors on request.

In total, 7528 CRFs were identified. However, in order to proceed conservatively and to reduce the number of fragments arising simply as a result of cross-serotype sequence conservation or convergence, or those identified on the basis of a very limited number of nucleotide identities, a smaller, 'filtered' subset of CRFs was arrived at by: (i) breaking ties between multiple donors; (ii) discarding any CRF for which the bootstrapping algorithm scored $<90\,\%$; and (iii) discarding any CRF corresponding to a window in which cross-serotype mean pairwise identity was $>90\,\%$. CRFs removed through this filtering process are not necessarily non-recombinant in origin, but do have plausible non-recombinant explanations. This filtered subset numbered 286 CRFs.

The distributions of fragment lengths are shown in Fig. 1, both filtered (Fig. 1a) and unfiltered (Fig. 1b). The mean length of CRFs prior to filtering was 488 nt [median, 279 nt; 95 % confidence intervals (CIs), 109–1828 nt]; post-filtering, the mean length became 296 nt (median, 201 nt; 95 % CIs, 105–600 nt). Estimation of CRF length should be regarded with some caution because determining precise lengths of CRFs is difficult, as there is no clear basis for weighting fragment similarity and fragment length when computing overall fragment scores. Nevertheless, the results are not expected to be qualitatively sensitive to alternative weighting arrangements of length and identity in the identity function *C*. Potential recombination profiles across the genome were identified from the proportion of all sequences at a given position that accommodated a CRF. Profiles were constructed both pre- and post-filtering (Fig. 2). Prior to filtering, CRFs arose in the 5′ UTR, the leader and 1A genes and throughout all non-structural genes. Even prior to filtering, CRFs were rare in the capsid genes. Filtering removed almost all CRFs in the 5′ UTR (as a result of the bootstrapping procedure) and in the leader, 1A and 3C genes (as a result of high pairwise cross-serotype similarity, which renders identification of recombination events ambiguous).

The pattern of intertypic exchange is illustrated in Fig. 3, which maps the filtered CRFs onto the 156 genome sequences. Exchange between SAT types (26/286) and between A–O–C–Asia 1 types (255/286) occurs more frequently than between SAT and non-SAT serotypes (5/286), providing a reasonable reflection of the current geographical distributions of the virus and going some way to validating the effectiveness of the methods. One fragment spanned almost the entire length of the alignment, suggesting that the SAT 1 sequence with GenBank accession no. AY593844 (SAT1/7 ISRL/4/62) resembles SAT 2 with GenBank accession no. AY593849 (SAT2/3 Kenya 11/60) more closely across its length than any of the other SAT 1 sequences, and suggesting that this sequence and its classification should be examined further.

The viability of recombinant viruses clearly depends on whether gene function is affected by recombination, and it is possible that recombination events that affect more than one gene would be less likely than recombination within a gene locus. Therefore, the distribution of filtered CRFs was examined to see whether they were more or less likely to cross gene boundaries than expected by chance. This was accomplished by examining pairs of adjacent genes and computing the observed proportion of occasions on which a CRF spanned the boundary between the two adjacent genes, given that the CRF arose in one gene or the other. Each of the k observed fragments in the two sequential genes (denoted i and $i+1$) has a probability of crossing the boundary given

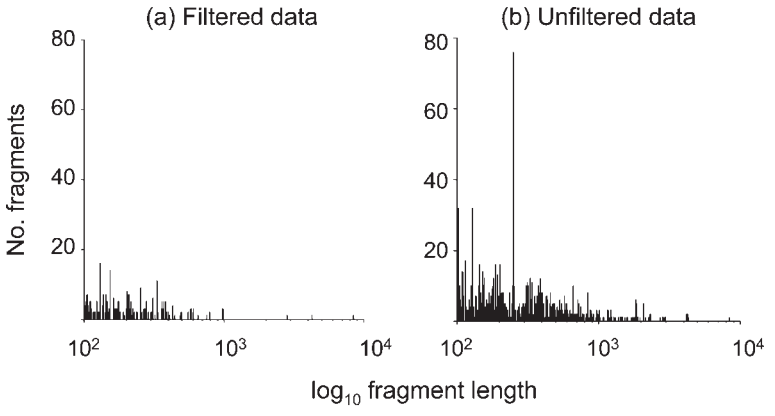

Fig. 1. Frequency distribution of candidate recombination fragment (CRF) lengths. The frequency of fragment lengths (*y* axis) of different lengths (*x* axis; note the logarithmic scale) located throughout the FMDV genome was identified by using the algorithm described in the text, (a) subsequent to filtering ('filtered', 286 fragments) and (b) prior to filtering ('unfiltered', 7528 fragments).

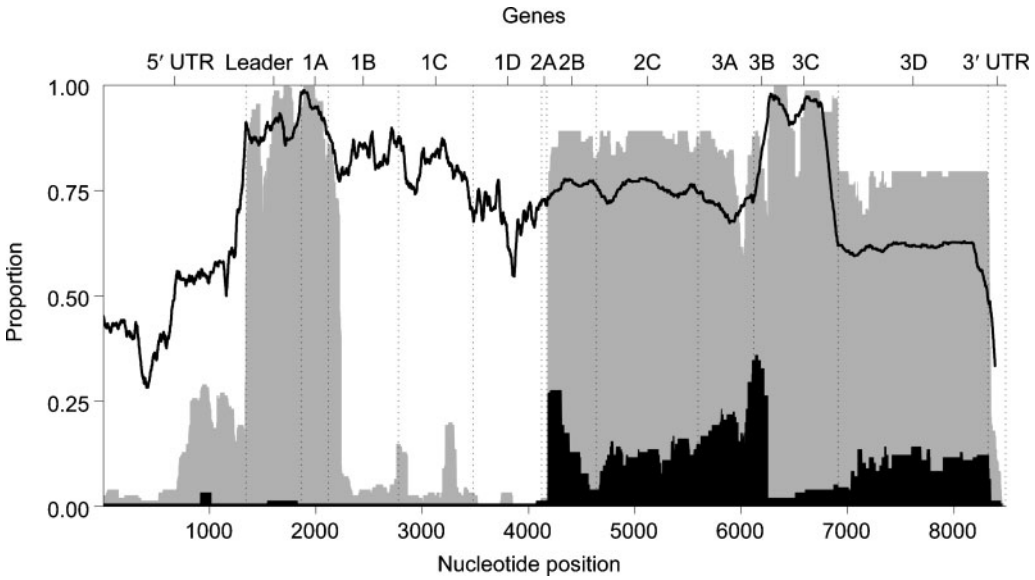

Fig. 2. Occurrence of candidate recombinant fragments (CRFs) throughout the FMDV genome. The black bars represent the recombination profile across the FMDV genome, constructed by using the filtered CRFs, and the grey bars represent the profiles constructed by using the unfiltered data. The bars represent the proportion of sequences (out of 156) at a given locus that contained a CRF. The solid black line indicates how conserved all of the sequences are over a 100 nt window starting at a given locus, as measured by the mean proportion of nucleotide matches across all pairwise comparisons. The black vertical dotted lines are the gene boundaries and are labelled on the upper *x* axis at their centre.

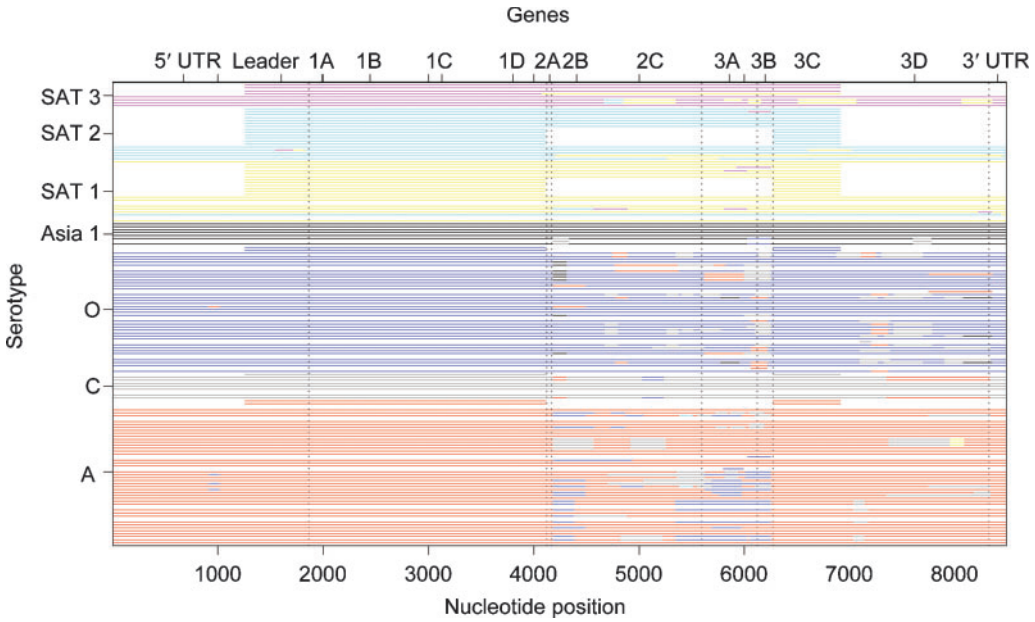

Fig. 3. Mosaic plot showing the location and the most likely serotype of the donor sequence for candidate recombinant fragments (CRFs). The filtered data represented here are the same as those used to generate the black bars in Fig. 2. Gene boundaries are indicated by vertical dotted lines and are labelled at their centre on the upper *x* axis. White regions indicate unsequenced regions of partial genomes. The order of the genomes in the figure (from bottom to top) corresponds to those reported in Supplementary Table S1 (available in JGV Online).

by

$$P_k = 1 - \frac{L_{i+1} + \max(L_i - L_k, 0)}{L_i + L_{i+1}}$$

where L_i is the length of gene i, L_{i+1} is the length of gene $i+1$ and L_k is the length of the kth fragment. The expected number of boundary-crossing events between a pair of sequential genes, given the occurrence of n CRFs in either one gene or the other, is then Poisson-distributed with rate

$$\lambda_{i,i+1} = \sum_{k=1}^{n} P_k$$

which permits straightforward estimation of the significance of the observed number of boundary-crossing events. The only gene boundaries at which a non-random pattern of fragment crossing was identified were the 2B/2C boundary ($P<0.01$, crossed less often than expected), the 3B/3C boundaries ($P<0.05$, crossed less often than expected) and the 3A/3B boundary ($P<0.05$, crossed more often than expected).

Our analysis confirms previous findings that recombination in FMDV appears to be widespread between serotypes, but is mostly restricted to the non-structural genes (Carrillo *et al.*, 2005; van Rensburg *et al.*, 2002). Whilst the exact quantitative findings of this analysis are somewhat sensitive to filtering rules necessitated by the need to distinguish between putative recombination events and chance convergence in highly conserved areas of the FMDV genome, the qualitative pattern is robust.

Previous analyses based on patterns of linkage disequilibrium have suggested that recombination rates between closely related FMDV genomes are likely to be high. The results of the analysis presented here suggest (not surprisingly) that recombination patterns are constrained by the geographical distribution of different serotypes. Intertypic exchange is likely to be constrained further by the infrequency with which individual animals will be infected simultaneously by multiple serotypes. Our analysis would not detect the wholesale exchange of capsid gene-coding regions between two serotypes (as serotype classification is based on the phenotypic expression of capsid genes), but it would be expected to pick up chimeric capsid genotypes. The infrequency of such chimeras suggests that there are severe functional constraints on intertypic recombination between FMDV structural genes, perhaps arising from epistatic interactions between the four capsid subunits. The phenotypic consequences of recombination in the non-structural genes are unlikely to manifest themselves in terms of antigenic shifts (specific immune responses are largely considered to be directed toward the capsid proteins; McCullough *et al.*, 1992); however, other possible effects on viral fitness remain unexplored.

Our analyses, conducted on the most extensive FMDV dataset currently available, confirm the impression hinted at by previous studies on more restricted FMDV datasets and reveal this genus to be no different from others in the family *Picornaviridae* in possessing largely non-recombining sets of structural genes that are linked to a more widely mixed set of non-structural genes (Lukashev, 2005; Oberste *et al.*, 2004; Simmonds & Welch, 2006).

References

Aidaros, H. A. (2002). Regional status and approaches to control and eradication of foot and mouth disease in the Middle East and North Africa. *Rev Sci Tech* **21**, 451–458.

Carrillo, C., Tulman, E. R., Delhon, G., Lu, Z., Carreno, A., Vagnozzi, A., Kutish, G. F. & Rock, D. L. (2005). Comparative genomics of foot-and-mouth disease virus. *J Virol* **79**, 6487–6504.

Correa Melo, E., Saraiva, V. & Astudillo, V. (2002). Review of the status of foot and mouth disease in countries of South America and approaches to control and eradication. *Rev Sci Tech* **21**, 429–436.

Hall, T. A. (1999). BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symp Ser* **41**, 95–98.

Haydon, D. T., Bastos, A. D. S. & Awadalla, P. (2004). Low linkage disequilibrium indicative of recombination in foot-and-mouth disease virus gene sequence alignments. *J Gen Virol* **85**, 1095–1100.

Hedger, R. S. (1972). Foot-and-mouth disease and the African buffalo (*Syncerus caffer*). *J Comp Pathol* **82**, 19–28.

Hedger, R. S., Condy, J. B. & Golding, S. M. (1972). Infection of some species of African wild life with foot-and-mouth disease virus. *J Comp Pathol* **82**, 455–461.

King, A. M. Q., McCahon, D., Saunders, K., Newman, J. W. I. & Slade, W. R. (1985). Multiple sites of recombination within the RNA genome of foot-and-mouth-disease virus. *Virus Res* **3**, 373–384.

Kivaria, F. M. (2003). Foot and mouth disease in Tanzania: an overview of its national status. *Vet Q* **25**, 72–78.

Krebs, O. & Marquardt, O. (1992). Identification and characterization of foot-and-mouth disease virus O_1 Burgwedel/1987 as an intertypic recombinant. *J Gen Virol* **73**, 613–619.

Lukashev, A. N. (2005). Role of recombination in evolution of enteroviruses. *Rev Med Virol* **15**, 157–167.

McCahon, D. & Slade, W. R. (1981). A sensitive method for the detection of recombinants of foot-and-mouth disease virus. *J Gen Virol* **53**, 333–342.

McCahon, D., Slade, W. R., Priston, R. A. J. & Lake, J. R. (1977). An extended genetic recombination map for foot-and-mouth disease virus. *J Gen Virol* **35**, 555–565.

McCahon, D., King, A. M. Q., Roe, D. S., Slade, W. R., Newman, J. W. & Cleary, A. M. (1985). Isolation and biochemical characterization of intertypic recombinants of foot-and-mouth disease virus. *Virus Res* **3**, 87–100.

McCullough, K. C., De Simone, F., Brocchi, E., Capucci, L., Crowther, J. R. & Kihm, U. (1992). Protective immune response against foot-and-mouth disease. *J Virol* **66**, 1835–1840.

Oberste, M. S., Maher, K. & Pallansch, M. A. (2004). Evidence for frequent recombination within species human enterovirus B based on complete genomic sequences of all thirty-seven serotypes. *J Virol* **78**, 855–867.

Simmonds, P. & Midgley, S. (2005). Recombination in the genesis and evolution of hepatitis B virus genotypes. *J Virol* **79**, 15467–15476.

Simmonds, P. & Welch, J. (2006). Frequency and dynamics of recombination within different species of human enteroviruses. *J Virol* **80**, 483–493.

Sutmoller, P., Barteling, S. S., Olascoaga, R. C. & Sumption, K. J. (2003). Control and eradication of foot-and-mouth disease. *Virus Res* **91**, 101–144.

Tosh, C., Hemadri, D. & Sanyal, A. (2002a). Evidence of recombination in the capsid-coding region of type A foot-and-mouth disease virus. *J Gen Virol* **83**, 2455–2460.

Tosh, C., Sanyal, A. & Hemadri, D. (2002b). Genetic and antigenic analysis of a recombinant foot-and-mouth disease virus. *Curr Sci* **83**, 1016–1019.

van Rensburg, H., Haydon, D., Joubert, F., Bastos, A., Heath, L. & Nel, L. (2002). Genetic heterogeneity in the foot-and-mouth disease virus Leader and 3C proteinases. *Gene* **289**, 19–29.

Vosloo, W., Bastos, A. D. S., Sangare, O., Hargreaves, S. K. & Thomson, G. R. (2002). Review of the status and control of foot and mouth disease in sub-Saharan Africa. *Rev Sci Tech* **21**, 437–449.

Wilson, V., Taylor, P. & Desselberger, U. (1988). Crossover regions in foot-and-mouth disease virus (FMDV) recombinants correspond to regions of high local secondary structure. *Arch Virol* **102**, 131–139.

Yang, J., Xing, K., Deng, R., Wang, J. & Wang, X. (2006). Identification of hepatitis B virus putative intergenotype recombinants by using fragment typing. *J Gen Virol* **87**, 2203–2215.