The genome sequence of the virulent Kabete ‘O’ strain of rinderpest virus: comparison with the derived vaccine

Michael D. Baron, Yasue Kamata, Valérie Barras, Lynnette Goatley and Thomas Barrett

Institute for Animal Health, Ash Road, Pirbright, Surrey GU24 0NF, UK

We have compared the complete genome sequences of the vaccine strain of rinderpest virus and the virulent strain from which it was derived. Only 87 bases differed between the two genomes (0.55%). Possibly significant differences in amino acid sequence were found in the N, P, F, H and L proteins. A number of differences were also found in the leader region (3’ end of the genome), whilst the trailer region appears to be more conserved. In addition, the length of the genome was found in both cases to be 15882, an exact multiple of six, fulfilling predictions made earlier based on work with Sendai and measles viruses.

Rinderpest is an economically highly important disease affecting cattle and wild bovids. Widespread in sub-Saharan Africa in the mid-1980s, it has now been restricted to parts of eastern Africa by the efforts of the Pan African Rinderpest Campaign (PARC). The disease remains enzootic in parts of the Indian sub-continent, as well as several countries in the Near and Middle East, with recent outbreaks in Kenya, Ethiopia, Pakistan, Iran, Yemen and Turkey. The causative agent, rinderpest virus (RPV), belongs to the morbilliviruses, a genus of the family Paramyxoviridae, and is thus related to measles virus (MV), canine and phocid distemper viruses (CDV and PDV), the dolphin morbillivirus (DMV) and peste des petits ruminants virus (PPRV). As part of our investigations into virulence factors in RPV we have cloned and sequenced the entire genome of the vaccine (RBOK) strain (hereafter referred to as RPV-R) (Chamberlain, 1992; Baron et al., 1993, 1994; Baron & Barrett, 1995a, b).

The situation with regard to the rinderpest vaccine is unique among morbilliviruses in that the virulent parental strain of the virus, unadapated to tissue culture, is available for comparative studies. This strain, the virulent Kabete ‘O’ strain (hereafter referred to as RPV-K), was originally collected in about 1910 from a sick animal in Kenya, and has been kept by animal-to-animal passage. The vaccine was derived from RPV-K in the late 1950s by multiple passages in primary calf kidney cell cultures (Plowright & Ferris, 1962). The virulent virus stock used in these studies had been passaged less than ten times since the vaccine was derived. Any sequence differences between the two genomes will be the result of the attenuation process, and not geographical or temporal drift, and inspection of the two sequences may be informative. In addition, since we wish eventually to clone both viruses and produce recombinant chimeras, it is necessary for us to have the sequence of both genomes. Just over half the sequence of the virulent strain has already been determined by ourselves and others (Hsu et al., 1988; Yamanaka et al., 1988, 1992; Baron et al., 1994; Ismail et al., 1994; Baron & Barrett, 1995b). In order to complete the comparison of the two sequences, and document the changes that occurred during the derivation of the vaccine strain, we have now sequenced the remaining sections of the virulent strain (primarily the large polymerase L gene). During this process we were able to eliminate several errors in the published RPV-K sequence, as well as one in the RPV-R sequence.

The genome of RPV, as with all the morbilliviruses, consists of six genes (3’ N-P-M-F-H-L 5’ on the genome) with short regions (roughly 50 bases) at each end of the genome which are thought to contain the polymerase binding sites and the encapsidation signals for the virus (Barrett et al., 1991). Each gene transcript terminates with a run of A residues which is extended by the viral polymerase to form the poly(A) tail of the mRNA. Between each gene is an untranscribed trinucleotide, usually CTT (in mRNA sense). The exact sequence in this region is probably important in determining the rate of initiation of transcription of the downstream gene; the H–L intergenic in RPV-R, MV and CDV all differ from the CTT consensus (Crowley et al., 1987; Sidhu et al., 1993; Baron & Barrett, 1995b) and it is probable that this difference is one of the causes of the extremely low levels of expression of the L gene found in infected cells.

The unpublished sections of the RPV-K (virulent) genome sequence comprise the N–P, F–H and H–L intergenic regions and the entire L (RNA polymerase) gene. All the sequences were amplified by RT–PCR from RPV-K genomic RNA using the proof-reading Vent DNA polymerase (New England
Biolabs). Total RNA was extracted from tissue obtained from a cow infected with RPV-K, and reverse transcribed, as previously described (Baron & Barrett, 1995b). The resultant cDNA was amplified using Vent polymerase in a 50 µl reaction containing 5 µl of cDNA, 2·5 units of enzyme, 10 pmol of each primer, 200 µM-dNTPs and 4 mM-Mg²⁺ ions in the buffer supplied by the manufacturer. The reaction was cycled 25 times (94 °C, 1 min; 50 °C, 1 min; 72 °C, 1 min per kb of amplified product). Amplified DNA was purified in a low melting point agarose gel and cloned into pGEM-SZf(+) (Promega). Inserts were sequenced on both strands using nested deletion sets generated with the Erase-a-Base kit (Promega) (L gene fragments) or using internal restriction sites and internal primers (F gene) (the number of A residues in the genomic N–P and F–H intergenic sequences were determined on one strand only). Sequencing was performed by the dideoxy method using unmodified T7 polymerase (Pharmacia). All PCR products were sequenced from two independently produced clones. Analysis of sequence data was carried out using the Staden and Genetics Computer Group packages.

In the case of the intergenic regions, the number of A residues and the intergenic trinucleotides were found to be identical to the RPV-R sequence in all three cases. Alignment of the 3' untranslated regions of the N genes of the two strains revealed a large apparent difference between the RPV-K sequence in the database and that determined by us for RPV-K, and for RPV-R (Fig. 1). Such a difference might have consequences for the expression of the N protein. Examination of the difference, however, suggested that it might be due to a transposition of four bases from gene position 1635 to 1659 in the database RPV-K sequence. If this is taken into account, the three sequences match (Fig. 1).

Fig. 1. Alignment of the N–P intergenic region of RPV-R and those of the RPV-K sequence determined in our laboratory (RPV-Ka) and that in the database (RPV-Kb). The transposed four bases are shown.

When these sequences were combined with the rest of the RPV-K sequences, the full genome sequences of the two virus strains could be compared. The only place where it was necessary to insert gaps to maintain the alignment of the sequences was in the 5' untranslated region of the RPV-K F gene, part of the genome that has an unusually high GC content, and which we had found difficult to sequence accurately in our RPV-R clones. We therefore resequenced this section from the RPV-K genome, using both deaza-dGTP and deaza-dATP in the sequencing reactions, and sequencing both strands. In this way we were able to identify a number of places where the previously published sequence for this region was in error, usually by the transposition of G and C residues, or the 'loss' of one or more residues in runs of C or G residues (Fig. 3). We also found that our own previously published sequence for the RPV-R F gene was missing a G residue at position 102. Interestingly, these insertions bring the total length of the genome for each strain to 15882, an exact multiple of 6, as predicted by the 'rule of six' first proposed for Sendai virus (Calain & Roux, 1993) and found for MV-based minigenomes (Sidhu et al., 1995).

Fig. 2 shows all the differences between the finalized RPV-K genome sequence and that of the avirulent RPV-R strain; for ease of comparison with mRNA sequences, the numbering used is 5' to 3' on the anti-genome sequence (3' to 5' on the genome).

Of the 61 differences that map to the protein coding regions, 26 do not give rise to a change in the encoded amino acid. Two more, in the proximal part of the P open reading frame (ORF), do not change the sequence of the encoded P protein, but give rise to conservative (K → R or R → K) changes in the C protein encoded in an overlapping reading frame.

No protein is without changes to at least one amino acid. Four differences are found between the N protein sequences, of which two are conservative and two non-conservative (Fig. 2). Surprisingly, in the latter two cases it is the RPV-K sequence that differs from the RPV consensus; position 106 is Q in RPV-R and the lapinized and the virulent Kuwait strains of RPV, as well as in MV, PPRV and DMV (CDV and PDV have P). Similarly, all morbillivirus N proteins have L (or other large
In the P ORF there are three amino acid changes: the G → R difference at position 84 is in a non-conserved position, R being found also in PDV; at 339 PDV, CDV, MV and RPV-R have L, only RPV-K having V, while at 370 only RPV-R has the large hydrophobic F, the other known sequences having the smaller V or I. As this is in an otherwise strongly conserved region of the P protein, this must be considered a significant change. The carboxy-terminal half of P has been shown in MV to be necessary for interaction of N and P (Huber et al., 1991) and P with itself (Harty & Palese, 1995). More detailed mapping has been carried out on the Sendai virus P protein; the P proteins of this group of paramyxoviruses show only limited similarity to those of morbilliviruses, but position 370 appears to map roughly to Sendai virus P position 428, which is in the L protein binding domain (Smallwood et al., 1994).

The M protein is the most strongly conserved of all the morbillivirus proteins (Baron & Barrett, 1995a), so it is not surprising that the only differences seen in this gene are three conservative changes. The F protein sequence is also quite well conserved, perhaps because its function as a fusion protein...
Fig. 3. Alignment of the first 450 nucleotides of the F gene of RPV-R with the corresponding region of RPV-K as determined in our laboratory (RPV-Ka) and as previously published (RPV-Kb). The position of the G residue identified in RPV-R in this paper is shown white-on-black, and all other differences are shown black-on-grey.

places constraints on the whole structure of the molecule. Of the three differences found between the F proteins, only that at position 294 is non-conservative (A → R); interestingly, it is again the RPV-K sequence that differs from all the other morbillivirus sequences available.

The H protein shows a large number of changes, most of them non-conservative. However, this is not surprising as the H protein is the least conserved of all the morbillivirus proteins (Baron & Barrett, 1995 a), perhaps reflecting the fact that this viral protein must interact with one or more host cell surface proteins, and these targets may be different for each virus-host combination. One or more of the unconserved regions in H presumably forms part of the receptor-binding region, and changes in these regions may have critical effects on the efficiency and phenotype of the interaction of the virus with its receptor. It has recently been shown that changes in just two amino acids in the sequence of MV H can add or remove the ability of the protein to cause down-regulation of the viral receptor and haemadsorption (Lecouturier et al., 1996). Almost all the changes shown in Fig. 2 are in such unconserved regions, except for the V → G at 269, and that immediately follows a conserved proline, indicating it is part of a turn, which the G residue is unlikely to affect. The first five changes are all in a region rich in potential N-linked glycosylation sites and predicted loops which probably forms part of the surface of the H protein molecule. Functional investigation of the RPV-R and RPV-K proteins in isolation will require the identification of in vitro phenotypic markers of vaccine and virulent strains of RPV. In any event, studies on the H protein alone may not be sufficient: even limited passage of field isolates of MV in Vero cells has been shown to generate a phenotype of haemadsorption, in some cases without any change in the amino acid sequence of the H protein (Shibahara et al., 1994), indicating that other viral proteins must be involved in the adaptation to Vero cells.

Of the 29 differences found in the L ORF, 18 give rise to no change and a further five to conservative changes. Positions 96, 1653 and 2023, where there are major changes, are not conserved between morbilliviruses (Baron & Barrett, 1995 b). Two of the three remaining changes are from a conserved positively charged amino acid (R/K) to glutamine (positions 1233 and 2065); these may have significant effects on the
polymerses function, but too little is known at present of the distribution of binding domains and enzymatically active sites in the L protein to draw any firm conclusions. The L protein is one of the most conserved of the morbillivirus structural proteins (Baron & Barrett, 1995a; Komase et al., 1995), so any changes may be significant.

Although only a minority of differences lie in untranslated regions (26 out of 87), the rate of substitution in these regions (1.49 per 100 bases) is 34 times that in coding regions (0.43 per 100 bases), as the untranslated regions make up only 11% of the total genome. Notable among the differences in the UTRs are four within the first 55 bases of antigenome, the so-called 'leader' region. The first two of these occur within an otherwise highly conserved region (Baron & Barrett, 1995b) which is thought to contain the viral polymerase binding site. These changes may have a significant effect on the rate of transcription of viral mRNA or full-length anti-genomes. Interestingly, the corresponding region at the other end of the genome (the 'trailer'), where genome transcription starts, is more highly conserved, having only one difference between the two strains, and that in the less-well-conserved distal part of the sequence.

Over the whole RPV genome, therefore, between virus strains that differ completely in their pathogenic phenotype, only a few sequence changes have been found which either significantly change the protein coding sequence or occur in untranslated regions known to be of importance in virus function. Non-conservative changes that may affect function are seen in the N, P, F, H and L proteins, and these may severally or jointly be responsible for the change in virus virulence, no one protein being obviously more altered than any other. The sequence differences seen in the 'leader' region may also be responsible for, or contribute to, the altered phenotype of RPV-R.

These differences have arisen during the repeated passage in tissue culture, and the gross phenotypic change observed was the loss of virulence. In the case of MV, lesser phenotypic changes are observable: repeated passage, such as is used to generate attenuated strains, has been shown to increase the binding of the virus to its receptor (Shibahara et al., 1994), but does not appear to affect the phenotype of receptor down-regulation (Schneider-Schaulies et al., 1995). Unfortunately, the original MV isolate, from which most virus strains have been derived, is no longer available, the oldest available sample having already been adapted to two different cell lines (Rota et al., 1994a), which was almost certainly enough to change the growth phenotype of the virus, and hence, presumably, its sequence. Recently, it has proved possible to isolate wild-type MV strains without repeated passage using the marmoset lymphocyte cell line B95a (Kobune et al., 1990). Modern isolates of MV show considerable genetic drift from the original Edmonston isolate (Rota et al., 1992, 1994a, b), and it will not be sufficient to simply compare new isolates with the original vaccine. In the case of RPV, we now have the sequences of the vaccine and the non-adapted virulent parent. Elucidation of the role of these sequences in the virus life-cycle and the effects of changes at different positions within them will depend on the construction of chimeric genomes and replicons, currently under way in our laboratory. Care will have to be taken not to rescue such recombinant viruses in such a way that mutations are produced by the rescue system. The observation that field isolates of RPV, including RPV-K, grow without adaptation in B95a cells may be of assistance in this task since it appears that passage in this cell line does not alter the pathogenic phenotype of MV (Kobune et al., 1990).

References

Baron, M. D. & Barrett, T. (1995a). The sequence of the N and L genes of rinderpest virus and the 5' and 3' extra-genic sequences: the completion of the genome sequence of the virus. <i>Veterinary Microbiology</i> 44, 175-186.


Received 3 June 1996; Accepted 29 July 1996