Duplicated genes within the variable right end of the genome of a pathogenic isolate of African swine fever virus

Soopayah Vydelingum, Sally A. Baylis, Chrissy Bristow, Geoffrey L. Smith and Linda K. Dixon

1 AFRC Institute for Animal Health, Pirbright Laboratory, Ash Road, Pirbright, Woking, Surrey GU24 0NF
and 2 Sir William Dunn School of Pathology, University of Oxford, South Parks Road, Oxford OX1 3RE, U.K.

The right variable region of the genome of a pathogenic strain of African swine fever virus (ASFV), Malawi LIL20/1, has been sequenced and 15 open reading frames (ORFs) identified by computer analysis. Eight of these ORFs were found to be similar to previously described ASFV ORFs and three of these belong to two previously described multiple gene families (MGF), 360 and 110. Four of the remaining five ORFs belong to a novel MGF, designated MGF 100, and the last ORF encodes a protein that is similar to the virus structural protein, p22. Copies of MGF 110 and the gene coding for p22 have previously been characterized only at the left end of the ASFV genome. The organization of these genes suggests evolution by duplications, deletions and sequence transposition from one end of the genome to the other. Sequence comparisons of members of MGF 360 suggest that the Malawi LIL20/1 genome has undergone separate DNA rearrangements compared to the Ba71V genome. Lastly, one ORF was found to be similar to the myeloid differentiation primary response protein, MyD116 and to the herpes simplex virus neurovirulence-associated factor ICP34.5.

Introduction

African swine fever virus (ASFV) is a large, icosahedral, cytoplasmic DNA virus which is responsible for an economically important disease of swine. African swine fever is currently enzootic only within sub-Saharan Africa, the Iberian peninsula and Sardinia (Wilkinson, 1984). The disease varies from hyperacute to acute and chronic to inapparent depending on the virus isolate involved (Hess, 1981). ASFV infects wild and domestic pigs as well as warthogs, bushpigs and ticks of the family Ornithodoros.

The virus infects cells of the reticuloendothelial system in vivo and primarily monocytes and macrophages in vitro (Casal et al., 1984). The genome structure and replication strategy of ASFV are similar to those of poxviruses although the virus particle is structurally different (for reviews, see Vifiuela, 1985 and Costa, 1990). The genome consists of a linear double-stranded DNA molecule of 170 to 190 kbp with terminal inverted repetitions and terminal hairpin loops (Sogo et al., 1984; González et al., 1986). The central genome region of 125 kb is conserved in length but regions near both the left and the right termini are variable, the latter varying in size from 11 to 16 kb and the former from 38 to 47 kb (Blasco et al., 1989). In addition, a variable region located between 93 and 97 kb from the left end has been reported (Sumption et al., 1990). Two families of homologous genes have been identified at the variable ends of the genome (González et al., 1990; Almendral et al., 1990). Copies of multiple gene family (MGF) 360, so designated because the genes belonging to that family code for proteins of approximately 360 amino acids, are present at both ends of the genome. Copies of a second MGF, MGF 110, have so far been found only near the left end of the genome (Almendral et al., 1990).

The similarities between the genomes of ASFV and poxviruses and the presence of several genes encoding virulence factors at the right end of the poxvirus genome (Goebel et al., 1990) led us to sequence an 11.3 kb fragment from the right end of the genome of a virulent ASFV isolate from Malawi (LIL20/1) which has a different restriction site map (Dixon, 1988) compared to those isolates (Ba71V, LIS57) from which sequences of MGF 360 and MGF 110 have previously been determined. Here the presence, near the right end of the LIL20/1 genome, of additional members of MGF 360 and MGF 110 is reported with an unrelated and novel MGF (MGF 100). One ORF with similarity to a virus
structural protein encoded near the left end of the Ba71V genome is also described. Another ORF has similarities over a carboxy-terminal domain with both a myeloid differentiation primary response protein (MyD116) (Lord et al., 1990) and a herpesvirus-encoded neurovirulence-associated protein (McGeoch & Barnett, 1991).

**Methods**

ASVF DNA was isolated from clone LMw23, which is part of a bacteriophage λ library containing inserts from the genome of the ASVF Malawi LIL20/1 isolate (Dixon, 1988). The isolated ASVF DNA was sonicated, end-repaired and cloned into M13 vectors. Single-stranded DNA was sequenced by the chain termination method (Sanger et al., 1977), using [α-32P]dATP. Sequences were read using a sonic digitizer and assembled by the computer programs DBUTIL and DBAUTO (Staden, 1982). ORFs were identified using the program ORFFILE (kindly provided by M. E. G. Boursnell, Cantab Pharmaceuticals, Cambridge, U.K.). Database searches were carried out using FASTA (Pearson & Lipman, 1988). Protein sequence comparisons, secondary structure and potential signal peptide predictions were carried out using the University of Wisconsin Genetics Computer Group package of programs (Devereux et al., 1984).

**Results**

Fig. 1 shows the position of the 11.3 kbp portion from the SalI I fragment contained in clone LMw23 (Dixon, 1988), which has been sequenced. Restriction enzyme site mapping indicated that about 2.5 kbp from the right end of the genome terminus, consisting of repeated sequences present within the terminal inverted repeats (K. J. Sumption, L. K. Dixon & P. J. Wilkinson, unpublished), is not contained within the insert in clone LMw23 (Dixon, 1988) and has therefore not been sequenced. ORFs coding for 65 amino acids or more were identified by computer analysis and the deduced sequences compared with each other and with protein sequence databases. ORFs were designated 1 to indicate the SalI fragment from which they are derived and numbered sequentially from the left. Those read towards the left end of the genome are designated L and those to the right R. This analysis identified duplicated genes and ORFs encoding proteins similar to those in the databases and these are described here. A sequence of seven thymidines has been defined as the transcription termination signal for some early and late ASVF genes (Almazán et al., 1992, 1993). This sequence has been found downstream within 140 bp of five ORFs. Other ASVF transcription signals have yet to be defined.

**Protein p22**

The ORF coding for protein p22, an external virion protein which is also expressed on the cell surface, was identified at the left end of the genome of the attenuated Spanish isolate Ba71V (Camacho & Vifueira, 1991). The protein predicted to be encoded by ORF ll0L of the Malawi isolate has 58% amino acid similarity and 40.8% identity with p22 of Ba71V. An alignment of these sequences is shown in Fig. 2. The Malawi ll0L ORF codes for a protein of 167 amino acids, 10 amino acids shorter than the p22 of Ba71V which has nine extra amino acids at the N terminus. Despite this the hydrophobic character of the N-terminal sequence is conserved with stretches of 24 or 15 amino acids at the N terminus. Despite this the hydrophobic character of the N-terminal sequence is conserved with stretches of 24 or 15 amino acids at the N terminus. In contrast to the p22 of Ba71V, the Malawi-encoded protein does not contain a predicted cleavage site downstream of the N-terminal hydrophobic region.

Protein p22 of the Ba71V isolate has 10 cysteine residues, all of which are retained within the ll0L protein described here. Furthermore, five of the cysteines retain
Duplicated genes within ASFV genome

Table 1. Amino acid identities* between members of MGF 360 from two different ASFV isolates and ORFs 13R and 115R

<table>
<thead>
<tr>
<th></th>
<th>Ba71V K'360</th>
<th>K'362 L356</th>
<th>J319 D'311</th>
<th>D'363</th>
<th>LIS57 LIS375 LIS382</th>
</tr>
</thead>
<tbody>
<tr>
<td>13R</td>
<td>33 38 38</td>
<td>37 89 36</td>
<td>31 30</td>
<td></td>
<td></td>
</tr>
<tr>
<td>115R</td>
<td>74 61 53</td>
<td>47 37 52</td>
<td>30 29</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* The percentage identity was determined by pairwise alignment of amino acid sequences using GAP. Sequences of MGF 360 copies, i.e. K'360, K'362, L356, J319, D'311, D'363 from the Ba71V isolate and LIS375, LIS382 from the LIS57 isolate, were compared with the 13R and 115R products.

The motif CX_sCX_sCX_4CX_4C described before for the p22 molecule encoded by Ba71V. There are three potential Asn glycosylation sites present within the 110L product of the Malawi isolate genome whereas none was identified within the p22 of Ba71V. The absence of a cleavage site downstream of the N-terminal hydrophobic region within the 110L-encoded molecule suggests that it is expressed on the surface of infected cells rather than secreted. The 110L protein may also be incorporated into the external envelope of the virion.

Multiple gene family 360

Two ORFs (13R and 115R) encode proteins similar to those belonging to the MGF 360 described earlier (de la Vega et al., 1990; González et al., 1990). These code for proteins of 352 (13R) and 310 (115R) amino acids. Table 1 shows the percentage of identical amino acids when the products of ORFs 13R and 115R are compared with each of the MGF 360 copies on the Ba71V and LIS57 genomes. The product of ORF 13R has between 33 % and 89 % identity compared to different products of MGF 360 encoded by Ba71V. The most similarity is observed between the 13R product and the Ba71V D'311 product and these two ORFs are located at similar positions on their respective genomes. In contrast, the 115R ORF, which is the final ORF sequenced at the right end of the LIL20/1 genome, encodes an MGF 360 product that is more similar to that of the MGF 360 furthest to the left of the Ba71V genome (74 % identity within the Ba71V K'360 product) than to products of MGF 360 copies at the right end of the Ba71V genome. All the amino acids that are identical within all MGF 360 products previously described are present in the products of ORFs 13R and 115R.

Multiple gene family 110

The amino acid identity between ORF 112R and members of MGF 110 described earlier (Almendral et al., 1990) varies from 42 to 22.6 %, and the degree of similarity varies from 55 % to 47 % (Table 2). Eight cysteine residues were shown to be conserved within proteins of MGF 110 in Ba71V but only three of these are conserved in the product of ORF 112R (Fig. 3). Copies of MGF 110 have not been identified near the right termini of genomes of other ASFV isolates, but so far have been found near the left end of all isolates except one virus clone (de la Vega et al., 1990; Aguero et al., 1990).

Multiple gene family 100

Four other ORFs (15L, 16L, 17L, 18L) (Fig. 1), are similar to each other and therefore form a new MGF, designated MGF 100. The proteins encoded by these ORFs are aligned in Fig. 4 and range from 77 amino acids (ORF 16L) to 141 amino acids (ORF 15L) with ORFs 17L and 18L coding for proteins of 102 and 103 amino acids respectively. The amino acid identity between the products of the four ORFs varies from 52 % to as low as 18 % (Table 3). However, there are many conservative amino acid changes giving rise to a correspondingly higher percentage similarity (Table 3). These ORFs are contiguous and all read in the same direction towards the left end of the virus genome supporting a model for their evolution by gene duplication. Although the products of

Table 2. Amino acid similarity* and identity between members of MGF 110 from two different isolates and ORF 112R of Malawi LIL 20/1

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Identity (%)</td>
<td>22.6 27.5 42 27</td>
<td>23 26 28 28 28 26 23</td>
</tr>
<tr>
<td>Similarity (%)</td>
<td>50 55 52 47</td>
<td>48 46 50 46 43 46</td>
</tr>
</tbody>
</table>

* Percentage amino acid identity and similarity was determined by pairwise alignment of sequences using GAP (Devereux et al., 1984).
these ORFs are not related to other ASFV-encoded ORFs that have been described, we have identified an ORF within a published sequence from the left end of the LIS57 isolate genome (de la Vega et al., 1990) which is similar to products of MGF 100. This LIS57 ORF (LIS 100) is located between MGF 110 copies 137 and 290 near the left end of the genome and encodes a product of 124 amino acids. The LIS 100 product has the highest homology (44 % identity) and is most similar in length to the 15L product. Alignments between each of these MGF 100 members were found to be statistically significant using FASTACHECK, whereas when the sequences were randomly shuffled using SHUFFLE the alignments were not statistically significant.

Myeloid differentiation primary response protein

Another ORF (ORF 114L) has some similarity with a myeloid differentiation primary response antigen (MyD116) (Lord et al., 1990; Sussman et al., 1992). MyD116 is expressed in myeloid cells only. ORF 114L encodes a polypeptide of 184 amino acids with an Mr of 21.3K which has 51% identity and 61% similarity over a 40 amino acid carboxy-terminal domain with MyD116.
A striking feature of the ASFV genome is the large number of duplicated genes encoded. Variation in the number of members of two MGFs, MGF 110 and MGF 360, accounts for at least some of the large length variations which occur close to the left end of the virus genome (de la Vega et al., 1990; Agüero et al., 1990).

Copies of MGF 360 have previously been described at both ends of the Ba71V genome (González et al., 1990). We have now identified three other types of ORFs at the right end of the Malawi LIL20/1 genome that have previously been found only at the left end of ASFV genomes. These include one ORF which is similar to the gene for p22, a virus structural protein, one ORF which is similar to copies of MGF 110 and four ORFs which constitute a novel MGF, MGF 100.

González et al. (1990) proposed a model to explain the presence of members of MGF 360 at both ends of the virus genome. The basic tenets of the model are that gene duplications and sequence divergence of duplicated members within one end of the genome were followed by translocation of sequences from one genome end to the other. Disparity in the number of members of the MGF at opposite ends may be explained either by deletion of genes or further duplications following translocation. This model explains the presence of related genes at opposite ends of the genome and the observation that individual members of MGF 360 at opposite genome ends are more similar than individual members within one end of the genome. Such sequence transpositions from one end of the genome to the other also occur in poxviruses (Moyer & Graves, 1981; Pickup et al., 1984), a group of viruses with many similarities with ASFV. The direction of transcription of copies of MGF 360, MGF 110, MGF 100 and p22 either towards or away from the genome terminus is the same at the right end of the Malawi LIL20/1 genome as at the left end of the Ba71V or LIS57 genomes. This is as expected from models proposed to explain poxvirus terminal sequence transpositions. Although one of the two (13R) MGF 360 genes present near the right end of the Malawi LIL20/1 genome is most closely related to the Ba71V MGF 360 gene, which has a similar genome location, the second (115R) is most closely related to the Ba71V MGF 360 gene that is furthest to the left on this genome and is not very closely related to MGF 360 members located near the right terminus of the Ba71V genome. This suggests that the right terminal regions of the Ba71V and Malawi LIL20/1 genomes may have undergone deletions at different locations within the genome, leading to loss of different genes and thus the direct homologue of 115R and possibly other genes may have been deleted from the Ba71V genome. Based on sequence comparisons between MGF 360 members on the Ba71V genome, González et al. (1990) also suggested that the two right terminal MGF 360 members have been deleted from the Ba71V genome. Our results support their prediction. ORF 114L, which encodes a product that has homology to MyD116, is adjacent to 115R on the LIL20/1 genome and thus is one ORF which may also have been deleted from the Ba71V genome. The 114L product may be involved in the interaction of ASFV with its host. The protein might function by manipulating the differentiation process of myeloid cells, which are targets for ASFV replication, since the myeloid differentiation primary response genes are believed to be involved in the conversion of undifferentiated myeloid precursors into non-proliferating macrophages and granulocytes (Lord et al., 1990). A similar gene in herpes simplex virus types 1 and 2 has been shown to be a neurovirulence factor (McGeoch & Barnett, 1991; Chou & Roizman, 1990) that may interfere with the regulation of programmed

Discussion

A striking feature of the ASFV genome is the large number of duplicated genes encoded. Variation in the number of members of two MGFs, MGF 110 and MGF 360, accounts for at least some of the large length variations which occur close to the left end of the virus genome (de la Vega et al., 1990; Agüero et al., 1990). Identical amino acids are boxed in black.

Fig. 5. Sequence alignment of a conserved carboxy-terminal region between the protein encoded by ORF 114L and the mouse myeloid primary response differentiation antigen, MyD116. An alignment between the amino acid residues encoded by ORF l14L and those between residues 567 and 607 of the 622 amino acid long MyD116 product was produced using PILEUP and output was generated using PRETTYBOX (Devereux et al., 1984). Identical amino acids are boxed in black.
cell death in neuroblastoma cells (Chou & Roizman, 1992). The presence of MGFs within the three virus isolates for which sequence information is available clearly points to some important functions. It has earlier been shown that MGF 110 is not involved in either virulence or pathogenicity in domestic pigs (Aguero et al., 1990) and it has been proposed that the products of this MGF may be part of the virus adaptation for growth in ticks (Dixon & Wilkinson, 1988). If that is true and given the genetic diversity within a given genus of ticks, such as Ornithodoros moubata (Walton, 1964), it is not surprising that the genes are present in multiple copies and are slightly different from each other.

We would like to acknowledge financial support from MAFF and AFRC.

References


(Received 7 April 1993; Accepted 1 June 1993)