Virology xxx (2010) xxx–xxx

Contents lists available at ScienceDirect

Virology

journal homepage: www.elsevier.com/locate/yviro

Phylogenomic analysis of 11 complete African swine fever virus genome sequences

Etienne P. de Villiers [a,*], Carmina Gallardo [b], Marisa Arias [b], Melissa da Silva [c], Chris Upton [c], Raquel Martin [b], Richard P. Bishop [a]

[a] International Livestock Research Institute, PO Box 30709, Nairobi 00100, Kenya
[b] EU reference Laboratory for ASF, CISA-INIA, Crta Algete el Cesar s/n 28130 Valdeolmos, Madrid, Spain
[c] Biochemistry and Microbiology, University of Victoria, Victoria, BC, Canada V8W 3P6

ARTICLE INFO

Article history:
Received 4 November 2009
Returned to author for revision
22 November 2009
Accepted 13 January 2010
Available online xxxx

Keywords:
African swine fever virus
Comparison
Phylogenetics
Full-genome sequences

ABSTRACT

Viral molecular epidemiology has traditionally analyzed variation in single genes. Whole genome phylogenetic analysis of 123 concatenated genes from 11 ASFV genomes, including E75, a newly sequenced virulent isolate from Spain, identified two clusters. One contained South African isolates from ticks and warthog, suggesting derivation from a sylvatic transmission cycle. The second contained isolates from West Africa and the Iberian Peninsula. Two isolates, from Kenya and Malawi, were outliers. Of the nine genomes within the clusters, seven were within p72 genotype 1. The 11 genomes sequenced comprised only 5 of the 22 p72 genotypes. Comparison of synonymous and non-synonymous mutations at the genome level identified 20 genes subject to selection pressure for diversification. A novel gene of the E75 virus evolved by the fusion of two genes within the 360 multicopy family. Comparative genomics reveals high diversity within a limited sample of the ASFV viral gene pool.

Introduction

African swine fever (ASF) is an acute, highly contagious and often fatal disease of domestic pigs (Hess, 1981; Ley et al., 1984) caused by African swine fever virus (ICTVdB Type Species 00.002.0.01.001 [ASFV]). This is a large cytoplasmic virus and is the only currently recognized member of the family Asfarviridae (Dixon 1988; Dixon et al. 2004). The genome is a single molecule of linear double-stranded DNA (dsDNA) and is between approximately 170 and 190 kbp in size, depending on the isolate. The primary reservoir of the virus is probably soft ticks of the genus *Ornithodoros* and ASFV is the only known arbovirus with a DNA genome. Wild African suids, most importantly warthogs and bush pigs can be infected but do not exhibit clinical symptoms (reviewed by Penrith et al. 2004). The epidemiology of ASF is complex and varies according to location, both a sylvatic cycle involving ticks associated with wild suids and direct pig to pig transmission are important in different regions. Acute disease caused by the virus is characterized by high fever, hemorrhages in the reticuloendothelial system, and a high mortality rate. Infectious virus can survive for several months in fresh and salted dried meat products. Montgomery first formally described the ASFV from Kenya in 1921 (Montgomery 1921). Recent studies indicate that the molecular diversity of the virus, as defined by partial sequencing of the major capsid protein p72 (Bastos et al. 2003), appears to be highest in central and eastern Africa (Lubisi et al. 2005). The disease spread from West Africa to the Iberian Peninsula, initially to Portugal in 1957 and 1960, and subsequently to several other countries in Europe and Latin America (reviewed by Penrith et al. 2004). In 2007 there was also a serious outbreak in Georgia, subsequently spreading to adjacent countries including Russia, which appears to have originated in south east Africa (Rowlands et al. 2008). These outbreaks have been extremely costly, since there is no current control measure, other than slaughter of infected pig herds. In Europe the disease remains endemic in Sardinia. ASFV is also endemic in most countries of Sub-Saharan Africa, where it is widely believed to constrain the development of the smallholder pig industry (Penrith et al. 2004).

The first complete ASFV genome was generated from the avirulent VERO cell culture adapted isolate BA71V (Yanez et al. 1995). The genome sequences of a virulent isolate from Benin and an avirulent tick isolate from Portugal were recently determined and a detailed comparison of the virulent Benin and the two avirulent isolates, derived from the p72 gene sequence group I was presented (Chapman et al. 2008). Chapman et al. (2008) also briefly describe automated annotation of seven additional complete genomes of southern and eastern African origin that had previously been deposited in GenBank. These seven isolates are described in Table 1, and the preliminary annotation is available on a publicly accessible website (http://athena.bioc.uvic.ca/database.php?db=asfarviridae).

We have assembled and annotated the complete genome sequence of E75, a second virulent isolate classified within p72 genotype I, originating from Spain. This was achieved using raw

* Corresponding author. Fax: +254 20 4223001.
E-mail address: e.villiers@cgiar.org (E.P. de Villiers).

0042-6822/$ – see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.virol.2010.01.019

Table 1
Summary of ASF virus genomes used in the study.

Abbreviation	GenBank accession	Strain name	Genome name	Genome size (bp)	No. ORFs	Notes[a]
ASFV-BA71V	NC_001659	BA71V	African swine fever virus strain BA71V	170101	160	Yanez et al. 1995; Country: "Spain"; Tissue culture adapted
ASFV-Benin97/1	AM712239	Benin97	African swine fever virus strain ASFV-Benin97	182284	156	Chapman et al. 2008; Country: "Benin"; Host: "Domestic pig"; Virulence High
ASFV-Ken	AY261360	Kenya 1950	African swine fever virus strain Kenya 1950	193886	161	Zsak et al. 2005; Country: "Kenya"; Host: "Domestic pig"; Virulence High
ASFV-Mal	AY261361	Malawi Lil-20-1 1983	African swine fever virus strain Malawi Lil-20-1 1983	187612	160	Haresnape and Wilkinson 1989; Country: "Malawi"; Host: "Tick"; Virulence High
ASFV-Mku	AY261362	Mkuzi 1979	African swine fever virus strain Mkuzi 1979	192714	167	Zsak et al. 2005; Country: "Zululand"; Host: "Tick"; Virulence Unknown
ASFV-OurT88/3	AM712240	OurT88_3	African swine fever virus strain ASFV-OurT88_3	171719	157	Boinas et al. 2004; Country: "Portugal"; Host: "Tick"; Virulence Low
ASFV-Pret	AY261363	Pretorisuskop-96-4	African swine fever virus strain Pretorisuskop-96-4	190324	167	Zsak et al. 2005; Country: "Republic of South Africa"; Host: "Tick"; Virulence High
ASFV-Teng	AY261364	Tengani62	African swine fever virus strain Tengani62	185689	162	Pan 1992; Country: "Malawi"; Host: "Domestic pig"; Virulence High
ASFV-War	AY261366	Warthog	African swine fever virus strain Warthog	186528	164	Zsak et al. 2005; Country: "Namibia"; Host: "Warthog"; Virulence Unknown
ASFV-Warm	AY261365	Warmbaths	African swine fever virus strain Warmbaths	190244	167	Zsak et al. 2005; Country: "Republic of South Africa"; Host: "Tick"; Virulence Unknown
ASFV-E75	FN557520	E75	African swine fever virus strain Spanish isolate	181187	166	This study; Country: "Spain"; Host: "Domestic pig"; Virulence High

[a] Notes indicate reference for virus, geographical origin, host and virulence of virus isolates.

sequence data generated by a commercial company and represents the first ASFV genome to be determined using 'next generation' pyrosequencing. We provide a phylogenetic analysis of the 11 publicly available complete ASFV genomes and contrast the results with genotyping based on sequence data from three single copy ASF genes, p72, p54 and central variable region (CVR) within the *B602L* gene.

Results

Sequencing of the ASFV E75 genome

The ASF Spanish isolate E75 (ASFV-E75) is a virulent and highly infective hemadsorbing virus, isolated from domestic pigs during outbreaks that occurred in Lerida (Spain) in 1975. The genome was sequenced using a 454 Life Sciences GS-20 sequencer and assembly produced a genome with 64× coverage and following PCR to fill in small gaps, the complete assembly resulted in a 181,187 bp genome. The genome was annotated and compared to the other sequenced ASFV isolates that were available in GenBank (see Table 1 for the origin of these sequences).

Identification of a core set of common genes among ASFV isolates

Concatenating large multigene datasets to improve the accuracy of phylogenetic inference is an accepted technique (Gontcharov et al. 2004; Rokas et al. 2003; Sanderson et al. 2003). We determined the core set of orthologous genes for each of the 11 ASF virus genomes using OrthoMCL (Li et al. 2003). A set of 123 orthologous ORFs comprising all conserved single copy genes and members of paralogous gene families common to all 11 genomes were identified. There was a high degree of conservation among these 123 genes, with an average BlastP similarity of 92%. The core set of 123 genes were regarded as orthologous and concatenated for each genome to create an input for phylogenetic analysis. The VOCs database identified 118 orthologous genes from the 11 genomes analyzed, compared to 120 genes when only 2 genomes are used. Chapman et al. (2008) reported 109 conserved genes in their study with 10 genomes. VOCs groups orthologous genes into families based on BLASTP scores set by a human database curator and this might explain the discrepancy in the number of conserved genes identified in the two studies. The larger number of orthologous genes predicted by OrthoMCL compared to VOCs are due to the fact that it was designed to identify both orthologous and paralogous genes which is not the case with VOCs.

Phylogenetic analyses of ASFV isolates

In order to determine the genetic relationship at the whole genome level between the ASFV isolates, we performed multiple sequence alignments of the concatenated core conserved set of genes from the 11 ASFV isolates. Several studies have shown that a concatenated multi gene approach can resolve ambiguities in phylogenetic reconstructions based on single genes (Gontcharov et al. 2004; Rokas et al. 2003). The amino acid sequences of the core set of orthologous genes from each of the 11 ASFV isolates were therefore concatenated into a single pseudo-sequence, and a neighbor-joining phylogenetic tree was constructed from a multiple amino acid sequence alignment of the concatenated sequences (Fig. 1A). This tree topology was used in subsequent analyses of the genetic relatedness of the isolates.

The phylogenetic tree analysis derived from the 123 concatenated genes separated the viruses into two major clusters that correlate with their geographical distribution. One cluster consists of four closely related isolates from West Africa and the Iberian Peninsula classified within p72 genotype group I (Bastos et al. 2003). This clade is in not visually obvious as a discrete cluster due to the geometry of the tree in which the four isolates form two sub-clusters located at the top and bottom, respectively (Fig. 1A). However according to the genetic distance these four isolates are very close to one another when compared to the other seven genomes. ASFV-Benin97/1 and ASFV-E75, whose sequence was determined in this study, form one sub-cluster while the culture adapted Spanish laboratory strain ASFV-BA17V and the non-pathogenic *Ornithodoros erraticus*-tick derived isolate ASFV-OURT88/3 from Portugal represented a second sub-cluster. The second major cluster consists of several Southern African ASFV isolates. Two South African tick-derived isolates (ASFV-Warm and ASFV-Pret) together with a warthog isolate from Namibia (ASFV-War) and the Tengani porcine isolate (ASFV-Ten) from Malawi form a single sub-cluster. A domestic pig isolate from Kenya (ASFV-Ken) and a tick-derived isolate from Malawi (ASFV-Mal) appear to be outliers and are not very closely related in terms of the overall phylogenomic analysis. Bootstrap support for the 123 gene concatenated tree was

Please cite this article as: de Villiers, E.P., et al., Phylogenomic analysis of 11 complete African swine fever virus genome sequences, Virology (2010), doi:10.1016/j.virol.2010.01.019

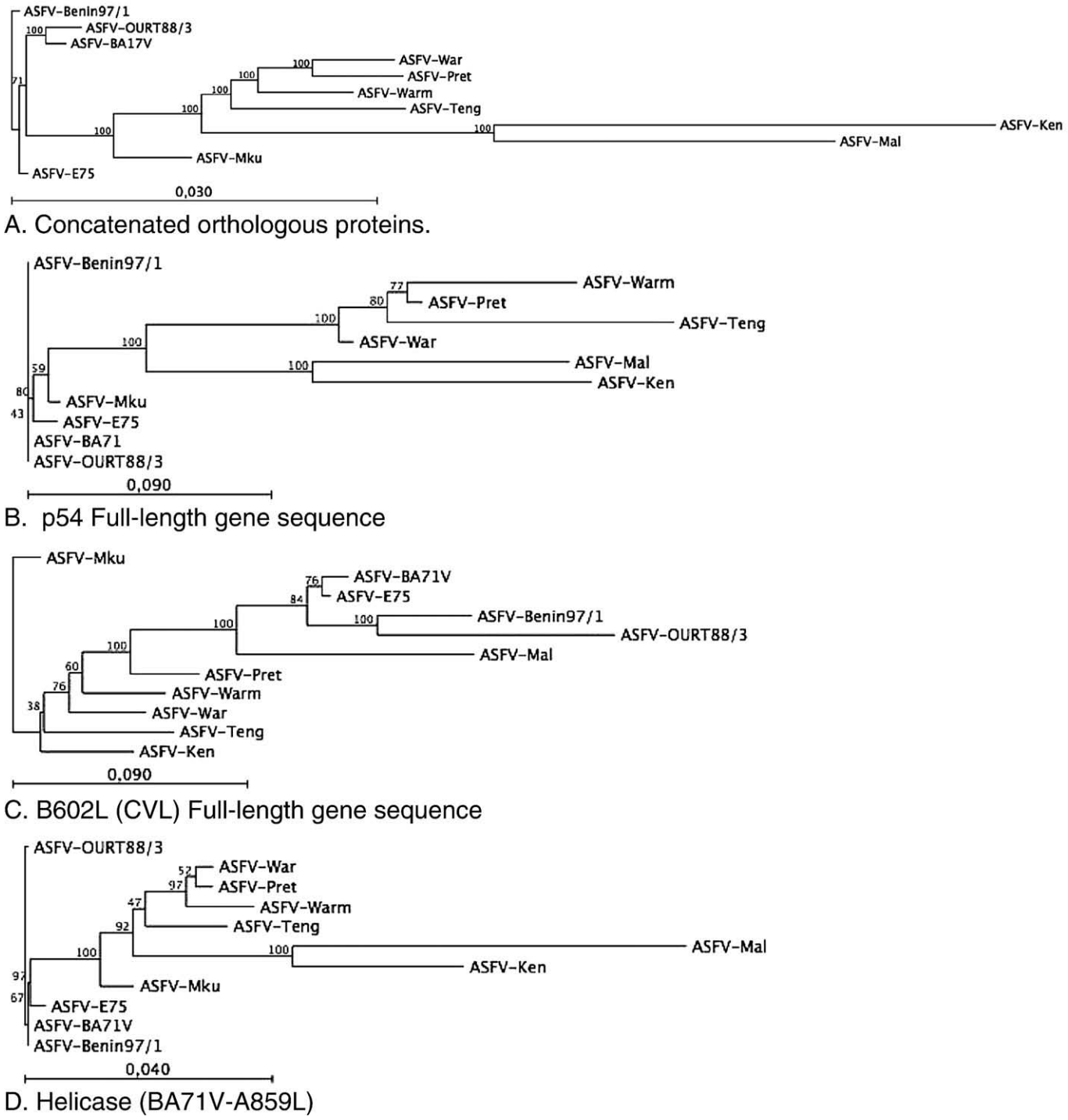

Fig. 1. (A) Neighbour-joining phylogenetic tree constructed from a multiple amino acid sequence alignment of 123 core-concatenated sequences. (B) Neighbour-joining phylogenetic tree constructed from a multiple amino acid sequence alignment of p54 sequences. (C) Neighbour-joining phylogenetic tree constructed from a multiple amino acid sequence alignment of B602L sequences. (D). Neighbour-joining phylogenetic tree constructed from a multiple amino acid sequence alignment of Helicase (BA71V-A859L) sequences. All trees are unrooted.

100% at all nodes except for that separating the ASFV BA71V/OURT88 sub-cluster from the other seven isolates, which had a value of 71. Using a similar neighbor-joining algorithm the p54 gene sequence (Fig. 1B) generated similar clusters for the West African/European p72 group I. The p54-based neighbor joining tree also supported the Southern African cluster, with the exception of the position of the Mkuzi tick isolate (ASFV-Mku) (Fig. 1B). However using the CVR within B602L (Fig. 1C), ASFV-OURT88/3 exhibited a substantial genetic difference from ASFV-BA71V and the same was also true when the virulent ASFV-Benin 97/1 and ASFV-E75 were compared. The genetic relationships defined by the variation in the CVR were therefore not concordant with analyses based on the entire genome, or other individual single copy genes. Phylogenetic tree constructed from amino-acid alignments of a predicted ASFV helicase protein (BA71V-A859L), gave a similar tree to that obtained from the core set of concatenated orthologous proteins (Fig. 1D).

A phylogenetic tree was generated from a 478 bp C-terminal sequence of p72 using a maximum likelihood algorithm. The

Please cite this article as: de Villiers, E.P., et al., Phylogenomic analysis of 11 complete African swine fever virus genome sequences, Virology (2010), doi:10.1016/j.virol.2010.01.019

maximum likelihood algorithm was employed, instead of the neighbor-joining algorithm used for the concatenated genes and other single copy loci, for consistency with previous molecular epidemiological analyses using the p72 gene (Bastos et al. 2003; Lubisi et al. 2005). This revealed that seven of the isolates with complete genome sequences were located within the p72 group I (Fig. 2). In addition to the four West African and European isolates this also included three members of the Southern African cluster, the exceptions being the Namibian warthog isolate (ASFV-War) which was classified within group IV and the Mkuzi tick-derived isolate (ASFV-Mku) which was within group VII (Fig. 2). The two outlying virulent isolates, specifically the Malawi tick-derived isolate (ASFV-Mal) and the Kenya 1950 porcine isolate (ASFV-Ken) fell into groups VIII (which comprised two sub-clades one with six isolates and the other containing only a single isolate) and IX, respectively.

Overall sequence identity at the genome level within the two clusters is shown in Table 2. Within the Southern African cluster identity varies between 91.0% and 94.8% (Table 2A), whereas the variation is between 88.5% and 99.1% in the West African-European cluster (Table 2B). Thus the range of variation is wider in the West

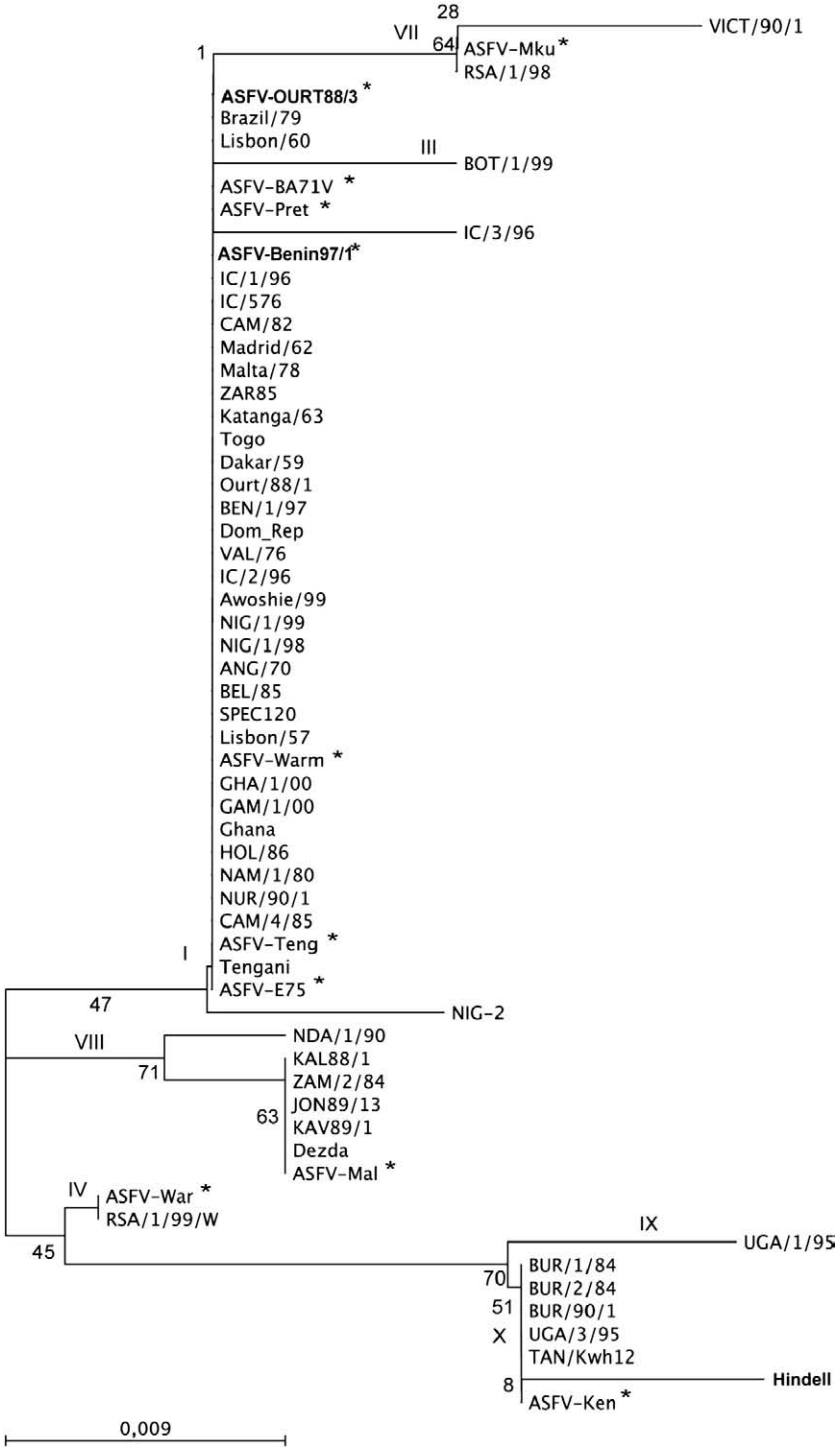

Fig. 2. Minimum evolution tree depicting the evolutionary relationships of 61 ASFV isolates based on alignment of a 478 bp C-terminal sequence of p72 genes. Genotypes are labeled (I – X) based on data from Bastos et al. 2003. Isolates falling in the Southern African cluster are marked with *. See Materials and methods for further details.

Please cite this article as: de Villiers, E.P., et al., Phylogenomic analysis of 11 complete African swine fever virus genome sequences, Virology (2010), doi:10.1016/j.virol.2010.01.019

Table 2

(A) Summary of percent identity between all pairs of sequences in a global nucleotide alignment of Southern African cluster genomes.

	AFSV-Mku	ASFV-War	ASFV-Warm	ASFV-Pret	AFSV-Ten
AFSV-Mku	100	92.2	94.6	92.7	91.0
AASFV-War	92.2	100	94.8	94.3	93.1
ASFV-Warm	94.6	94.8	100	94.3	92.5
ASFV-Pret	92.7	94.3	94.3	100	92.3
AFSV-Ten	91.0	93.1	92.5	92.3	100

(B) Summary of percent identity between all pairs of sequences in a global nucleotide alignment of West African and European cluster genomes.

	ASFV-OURT-88/3	ASFV-Benin97/1	ASFV-E75	ASFV-BA71V
ASFV-OURT-88/3	100	91.7	91.2	92.1
ASFV-Benin97/1	91.7	100	99.1	88.9
ASFV-E75	91.2	99.1	100	88.5
ASFV-BA71V	92.1	88.9	88.5	100

(C) Summary of percent identity of a global amino acid alignment of genes from all 11 genomes that have putative functions assigned to them or has been used for phylogenetic analysis.

Gene	% Identity
p72 (BA71V-B646L)	99.2
p54 (BA71V-E183L)	92.1
Helicase (BA71V-F1055L)	97.2
CD2v homolog (BA71V-EP402R)	72.1
Chaperone capsid folding (BA71V-B602L (9RL))	96.3
IkB-like protein (BA71V-A238L (5EL))	89.9

African-European clade. A diagrammatic representation of the two major clusters developed using whole genome alignments is presented in (Fig. 3). Fig. 3A shows the Southern African cluster and Fig. 3B the west African-European cluster. This analysis confirms the relative similarity of the genomes of the Southern African cluster (Fig. 3B) and indicates a lower occurrence of insertion and deletions in the left hand variable region than in The West-African European group. By contrast there appears to be more variation within the left hand variable section (Fig. 3B) in the West-African European group (note higher incidence of gray shaded areas in this region of the genome).

Ratio of synonymous and non-synonymous substitutions in ASFV proteins

To test for positive selection at the individual amino acid sites, four models of codon substitution were investigated (M1, M2, M7 and M8). Likelihood ratio tests (LRTs) were calculated and both models M2 and M8 significantly favored selection in comparison to models M1 and M7 ($P<0.001$). Model M2 identified 14 genes as being under positive selection (Table 3). These included multigene family proteins in the 360 and 505 families, several hypothetical proteins, the CD2 homologue, several enzymes and B602L, a chaperone that ensures the correct folding of the major capsid protein p72. The most stringent model for positive selection, M8, identified eighteen genes under positive selection (Table 3). Eight of the 18 genes under positive selection are genes that may be involved in modulating host cell function (Dixon et al. 2004) (highlighted in bold in Table 3). The major capsid protein of p72 does not have any sites that are under positive selection pressure and virtually all mutations are synonymous indicating strong stabilizing selection. A codon alignment of the p72 gene is provided under supplementary information (S2).

A novel open reading frame created by gene fusion between two members of the gene 360 family within the left hand variable region of the E75 genome

A deletion was detected within the left hand variable region of ASFV-E75 and presence of this was confirmed by PCR amplification using primers from the sequences flanking this region. This deletion result in the creation of an ORF that is unique among all 11 ASFV genomes, in which the 5′ end of the ASFV-BA71V-003 was fused in frame with the 3′end of ASFV-BA71V-004 (Fig. 4). The novel ORF comprised 360 amino acids and is a member of the 360 multigene family located very close to the left hand of the virus.

Comparison of the genomes of two virulent and two avirulent ASFV isolates classified within p72 genotype I

ASFV isolates have left and right variable genomic regions. These comprise a 38–47 kbp left variable genomic region and a 1–16 kb right hand variable genomic region (Blasco et al. 1989a; Blasco et al. 1989b; Irusta et al. 1996; Sumption et al. 1990). Additions or deletions in the inverted terminal repetitions ranging from approximately 1 to 7 kbp have also been observed (Blasco et al. 1989b; Tabares et al. 1987; Yozawa et al. 1994).

We compared the genome sequences of four ASF viruses in p72 group 1; ASFV-Benin97/1, and ASFV-E75, that are virulent isolates (from West Africa and Spain, respectively) and ASFV-BA17V and ASFV-OURT88/3 that are avirulent. Overall our results regarding the molecular basis of the differences between virulent and non-virulent isolates, including a second virulent isolate (ASFV-E75), supported those of Chapman et al. (2008), based on a single virulent isolate (Benin 97/1). A visual overview of similarities between each of the four aligned ASFV genome sequences in Base-by-Base (Brodie et al. 2004) is shown in Fig. 3A.

Chapman et al. (2008) demonstrated that both ASFV-BA71V and ASFV-OURT88/3 isolates are missing a sequence ~10 kbp, from a region 20 kbp from the left end of the genomes that is present in the virulent ASFV-Benin97/1. We found that in the Spanish E75 isolate (ASFV-E75), genome this region is also present and exhibits a high level of similarity to that in ASFV-Benin 97/1. This section of the genome contains six genes classified within the MGF 360 gene family. The ASFV-E75 and ASFV-Benin97/1 genome, each have one unique MGF 360 gene among a total of 16 in the family, MGF 360-1L in ASFV-Benin97/1 and MGF 360-20R in ASFV-E75. Both virulent isolates ASFV-E75 and ASFV-Benin97/1 share an additional MGF 360 gene, MGF 360-6L when compared to the avirulent isolates.

Discussion

We describe a phylogenetic analysis of 11 complete ASFV genomes using concatenation of a set of 123 orthologous proteins. Such phylogenies based on a method that combine multiple gene analysis have advantages relative to single gene analyses and may be more robust (Fitzpatrick et al. 2006; Rokas et al. 2003). Application of the concatenation technique could be compromised if extensive recombination has occurred between different virus isolates. However whole genome alignment of several ASFV genomes (see Fig. 3) does not suggest the occurrence of such inter-viral recombination, based on the lack of inversion or translocation of genomic regions.

Application of this technique identifies two clusters containing four and five isolates, respectively, with strong bootstrap support, plus two outlying genomes from Malawi and Kenya. One group comprises three different South African isolates from ticks and a Namibian warthog isolate, suggesting that it is ancestrally derived from a South African sylvatic transmission cycle. However it also contains the virulent porcine Tengani isolate from Malawi, suggesting that

Please cite this article as: de Villiers, E.P., et al., Phylogenomic analysis of 11 complete African swine fever virus genome sequences, Virology (2010), doi:10.1016/j.virol.2010.01.019

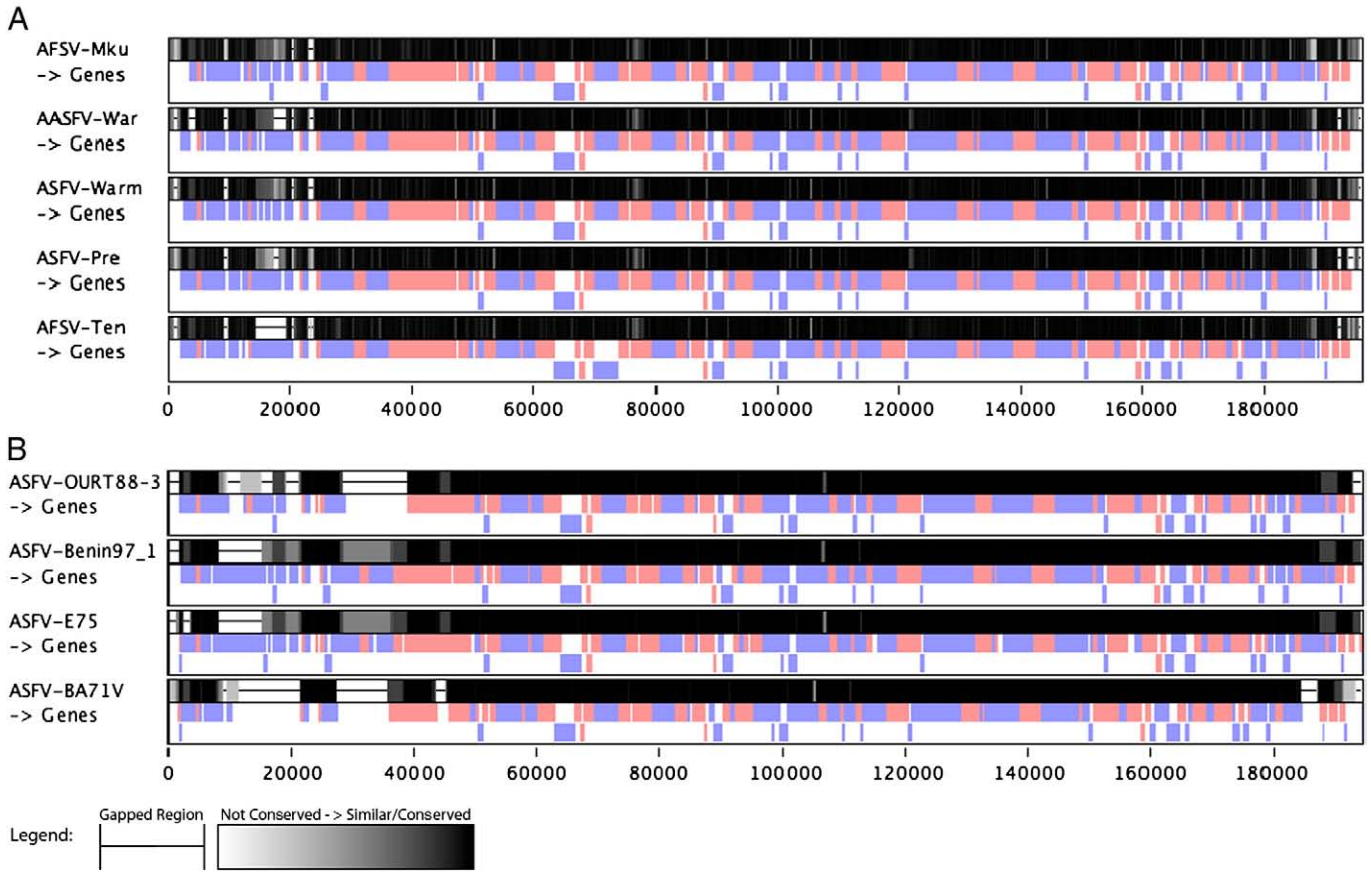

Fig. 3. (A) Scored similarity visual summary of complete virus genome alignments of five Southern African ASF virus genomes and (B) of four West African and European ASF virus genomes. The top grey scale track indicates similarity (black-perfect or 100% match; white-low or 0% match) and gaps are shown as dashes. The track showing red and blue boxes indicate top– and bottom strand open reading frames respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

outbreaks in domestic swine in northern South Africa may have historically originated from the warthog-tick sylvatic cycle. It is interesting to note that Tengani is classified within p72 genotype I along with two of the South African *Ornithodoros* tick isolates (Fig. 2). This is the same p72 genotype as West African-Iberian group, which is not thought to be tick-transmitted in West Africa. This suggests that p72 is subject to stabilizing selection (see supplementary data Table S1) and that while p72 C-terminal sequencing represents a useful technique for initial ASFV discrimination, based primarily on neutral genes, other gene sequences should also be analyzed to increase the value of the information from genotyping. To emphasize this point the Namibian warthog isolate and the Mkuzi tick isolate fall into different

Table 3
Log-likelihood values and parameter estimates for ASFV virus genes under positive selection determined under maximum-likelihood models M1, M2, M7 and M8. M1vsM2 likelihood ratio test statistic for model M1 versus M2; M7vsM8 likelihood ratio test statistic for model M7 versus M8. Parameter estimates: number of sites in $d_N/d_S>1$ category and estimated d_N/d_S parameter under model M8. *Significance with $P>0.05$. **Significance with $P>0.01$. ***Significance with $P>0.001$.

Locus	Gene	Function	LRT statistics		Parameter estimates		
			M1vsM2	M7vsM8	M2	M8	d_N/d_S
ASFV-E75-023	MGF 360-8L	MGF 360	6.18*	6.77*	11	11	2.61
ASFV-E75-033	MGF 505-4R	MGF 505	28.60***	29.04***	4	4	6.12
ASFV-E75-036	MGF 505-7R	MGF 505	16.19***	21.51***	21	21	1.83
ASFV-E75-037	MGF 505-9R	MGF 505	19.29***	19.20***	3	3	6.25
ASFV-E75-047	BA71V-A859L	Helicase superfamily II	5.87	6.78*	1	1	4.29
ASFV-E75-055	BA71V-K205R	Unknown	8.46*	8.80*	3	4	6.03
ASFV-E75-064	BA71V-EP153R	Lectin like protein	9.31**	10.50**	8	10	2.55
ASFV-E75-065	BA71V-EP402R	CD2 homologue	27.90***	42.12***	9	11	2.85
ASFV-E75-071	BA71V-C717R	Unknown	4.34	6.65*	2	2	2.98
ASFV-E75-087	BA71V-B602L	Chaperone	15.64***	17.68***	2	3	5.67
ASFV-E75-099	BA71V-CP2475L	Putative DNA primase	15.13***	20.54***	1	1	3.84
ASFV-E75-106	BA71V-NP1450L	RNApol 1	0.73	9.87**	3	3	1.43
ASFV-E75-121	BA71V-H171R	Unknown	6.39*	6.39*	3	3	5.93
ASFV-E75-124	BA71V-H108R	Putative signal peptide	8.87*	9.18*	8	8	6.56
ASFV-E75-129	BA71V-Q706L	Helicase superfamily II	7.32*	10.19*	0.2	0.2	9.75
ASFV-E75-134	BA71V-E423R	Unknown	5.62	6.45*	0.3	0.3	16.83
ASFV-E75-151	BA71V-I215L	Ubiquitin-conjugation enzyme	22.80***	23.44***	6	6	5.89
ASFV-E75-155	MGF 360-16R	MGF 360	27.14***	28.10***	1	1	19.29

Please cite this article as: de Villiers, E.P., et al., Phylogenomic analysis of 11 complete African swine fever virus genome sequences, Virology (2010), doi:10.1016/j.virol.2010.01.019

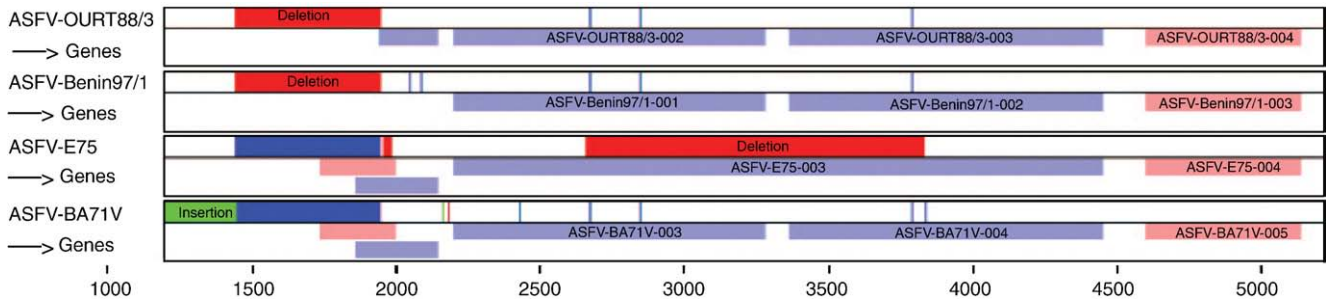

Fig. 4. Visual summary of region showing fused gene E75-004. CLUSTALW was used to align complete virus genomes and the alignments manually optimised using Base-by-Base (Brodie et al. 2004). A visual overview of similarities between each of the four aligned ASF virus genome sequences in Base-by-Base (Brodie et al. 2004) is shown in Fig. 3. The blue and pink boxes indicate top– and bottom strand open reading frames respectively, red boxes show deletions and green boxes insertions in E75 relative to other genomes. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

p72 genotypes indicating a lack of congruence between whole genome and single gene analyses based on p72 in certain instances.

Although phylogenetic discrimination between ASFV isolates using concatenation of a set of 123 orthologous proteins appears more robust than single gene analyses, it is impractical as a routine technique since it would require the whole genome sequencing of all newly isolated ASFV viruses. Analysis of other ASFV proteins that are under positive selection revealed that the helicase gene, BA71V-A859L appears to be a good phylogenetic marker and trees based on this sequence reproduce the genetic relationships indicated by analysis of the concatenated set of orthologous proteins.

Four isolates are classified within p72 genotype I (Bastos et al. 2003) that is currently widespread in West Africa and formerly the Iberian Peninsula and contains the largest group of ASFV isolates so far characterized. However it is interesting to note that, although the Benin virulent isolate originates from West Africa and the E75 isolate from Spain they are significantly more similar to one another than the two avirulent isolates from the same lineage. In particular the region deleted from BA71 and OURT88/3 (Chapman et al. 2008) that contains members of the 360 and 505 multicopy gene families is conserved in the virulent E75 isolate. These regions have previously been determined using functional analysis as important determinants of virulence in macrophages and tick host-range (Burrage et al. 2004; Ziheng et al. 2005). Thus the conclusions reached by Chapman et al. (2008) based on comparison of a single virulent isolate, Benin/71 from West Africa with same two avirulent viruses are reinforced by the newly determined genome sequence of E75. The BA71 virus is adapted to heterologous VERO cell cultures and can no longer replicate in porcine macrophages (Wesley and Tuthill 1984). By contrast OURT88/3 was isolated from *O. erraticus* ticks (Chapman et al. 2008), which likely represents a recent tick host, acquired subsequent to the introduction of the virus to Portugal in 1957. It therefore seems possible that the loss of virulence is associated with deletion of similar regions within the ASFV genome that have occurred independently. This may be attributable to lack of normal selective pressure from the unusual mammalian host cell in the case of BA71 and the novel *O. erraticus* vector in the case of OURT88/3. While all four West African-Iberian Peninsula viruses are similar in respect of their p72 (Bastos et al. 2003) and also p54 (Gallardo et al. 2009) sequences, the OURT88/3 B602L CVR sequence has diverged radically from the other Iberian Peninsula isolates in the approximately 50 years since introduction of the virus (Fig. 1C). This emphasizes the value of this gene for high-resolution discrimination of viral genotypes.

We report for the first time the use of 454 pyrosequencing (Margulies et al. 2005) to determine the genome sequence of an ASFV (E75) that was isolated from an infected pig exhibiting clinical symptoms. This technique allows determination of the sequence of ORFs located at ends of the virus genome, which are not possible using primer-based methods (Chapman et al. 2008). Shotgun 454 sequencing will in future allow determination of the sequence of more divergent ASF viruses, since primers designed from existing genome sequences are not required.

We have identified a novel ORF in E75 resulting from fusion of two ORFs present in the left hand variable region. This adds an additional mechanism to the ASFV diversification mechanism repertoire. In addition to ORF fusion this already includes deletions and copy number variations in the five multicopy gene families, and also frame shifts leading to the creation of non-functional genes, such as the CDv2 locus in the tick-derived OUR88/3. In addition single nucleotide polymorphisms occur within generally conserved coding sequences. These may prove to result in important changes in viral phenotype, but their significance is currently difficult to investigate functionally.

The analysis of the ratio of non-synonymous to synonymous mutations at the genome wide level reveals several genes that appear to be under selective pressure. A number of these are involved in interactions with the porcine host and tick vector. These include the Cdv2 gene that is responsible for red cell adherence and interference with the function of bystander host T cells (Rowlands et al. 2009). The 360 and 505 multicopy genes families have been demonstrated through functional deletion and replacement studies (Burrage et al. 2004; Zsak et al. 2001) to be involved in determination of virulence and host range. The most striking examples of positive selection are one of the members of the 360 multicopy gene family. Surprisingly a large number of ORFs, annotated as hypothetical proteins, are included among the group of genes most strongly subject to positive selection, although many additional genes of unknown function are thought to be involved in virus host interaction (Dixon et al. 1999; Dixon et al. 2004).

It can be concluded that phylogenetic analysis of the limited set of 11 complete ASFV genome sequences that are currently publicly available, representing only 5 of the 22 currently defined p72 genotypes reveals considerable genetic diversity at the genome level. The set of sequenced ASFV genomes originating from Southern and Western Africa and the Iberian Peninsula, together with one Kenyan virus would benefit from rapid expansion using the high throughput genome sequencing technologies whose use is now becoming widespread.

Materials and methods

Viruses and cells

The ASFV-E75 virus was isolated from the spleens of infected pigs and then passaged three times in swine buffy coat culture (E75L3). The titer of virus was determined as the amount of virus causing hemadsorption in 50% of infected cultures (HAD50/ml) as previously described (Carrascosa et al. 1982; Malmquist and Hay, 1960). Briefly, cells were seeded in 96-well tissue culture grade micro titer plates (200 μl; 300,000 cells/well) in homologous swine serum, and incubated in a humidified atmosphere containing 5% CO_2 at 37 °C.

Please cite this article as: de Villiers, E.P., et al., Phylogenomic analysis of 11 complete African swine fever virus genome sequences, Virology (2010), doi:10.1016/j.virol.2010.01.019

Three-day-old cultures were infected with 10-fold dilutions of the sample supplemented with 5 μg/ml gentamicin sulfate (BioWhittaker) and incubated for 24 h at 37°. After inoculation, a preparation of 1% homologous red blood cells in buffered saline was then added to each well. The plates were examined for hemadsorption during a 6-day period and the virus titer was estimated.

Purification of viral DNA

In order to provide a virus stock for extraction of virus DNA an animal was intramuscularly inoculated with 10^5 50% hemadsorbing doses (HAD50) of the pathogenic ASFV isolate E75. Virus was purified from the red blood cell fraction of infected pig blood essentially as described by Wesley and Tuthill (1984) except that 50% rather than 60% sucrose was used in step gradients. In addition, virus preparations were treated with DNase (50 pg/ml) followed by 1% Tween 80 in order to remove contaminating cellular DNA before loading onto sucrose gradients. DNA was prepared from isolated virus by phenol extraction following lysis of virus with SDS and proteinase K.

Sequence determination and DNA sequence analysis

The ASFV-E75 genome sequence was generated at Inqaba Biotechnical Industries (Pty) Ltd using a 454 Life Sciences GS-20 sequencer using standard protocols (Margulies et al. 2005). A total of 11.6 Mb of sequence data was generated from a single run of the GS20 sequencer and assembled using the Newbler assembler software. There were nine large contigs with a total size of 180 kb with the largest contig comprising 47 kb. PCR primers flanking the eight gaps were designed and PCR products sequenced. Reassembly resulted in a final ASFV-E75 genomic sequence that was 181,187 bp in length. Apparent errors in base copy number, due to ambiguities in the length of polynucleotide tracts that are an inherent feature of the pyrosequencing technique, were resolved by comparative mapping of the ASFV-E75 against reference ASFV sequences that had previously been annotated. This was performed using Artemis Comparison Tool (Carver et al. 2005).

The ASFV-E75 genome was annotated using the Genome Annotation Transfer Utility (GATU) (Tcherepanov et al. 2006) and the data entered into an ASFV specific Virus Orthologous Clusters (VOCs) database (Ehlers et al. 2002). Open reading frames identified had a minimum length of 180 bp and did not substantially overlap with other larger ORFs. Analysis of genome sequences and orthologous protein families was carried out using various programs available at Viral Bioinformatics—Canada (Brodie et al. 2004). This analysis included searches against the most recent databases using BLAST (Altschul et al. 1997) and against protein motif databases including Pfam. GenBank accession numbers for the genome sequence is FN557520. The nomenclature of ORFs is based on that used in the paper of Chapman et al. (2008).

Data generated in this study are also available at the online bioinformatics resource Viral Bioinformatics—Canada (http://athena.bioc.uvic.ca/database.php?db=asfarviridae).

Identification of orthologous genes

OrthoMCL (Li et al. 2003) was used to identify orthologous groups (families) in the 11 sequenced ASFV genomes deposited in the Asfarviridae Bioinformatics Resource (http://athena.bioc.uvic.ca/database.php?db=asfarviridae). It takes all-against-all BLASTP results from a set of protein sequences as input and defines putative pairs of orthologs or recent paralogs based on the reciprocal best BLASTP hit. Paralogs are identified as genes within the same genome that are more similar to each other than any sequence from another genome. OrthoMCL then converts the reciprocal BLASTP values to a normalized similarity matrix that is analyzed by a Markov Cluster algorithm (MCL). This yields a set of clusters each containing a set of orthologs and/or recent paralogs. OrthoMCL was run with default values for BLAST E-value cutoff of $1e^{-6}$ and an inflation parameter of 1.5.

Sequence alignment and phylogenetic analysis

The nucleotide sequences of ASFV genomes were aligned using CLUSTALW and substitutions, insertions and deletions observed between the aligned genomes were visualized using base-by-base (Brodie et al. 2004). Amino acid sequences of concatenated core set of orthologous genes were aligned with T-COFFEE (Notredame et al. 2000) and extensively hand edited in Jalview (Clamp et al. 2004) to remove gaps and optimize the alignment. Other protein sequences were aligned using CLUSTALW. Phylogenetic relationships among sequences were determined using neighbor-joining trees from DNA distances, each with 100 bootstrap replicates using the software MEGA3 (Kumar et al. 2004).

Validation of a novel genomic deletion specific to the ASFV E75 isolate

The initial sequence assembly indicated the presence of a potential gap located within the left hand variable region of ASFV E75. The existence of this missing sequence was validated by sequencing the PCR product generated from genomic DNA using primers (GapBA2660F: CTCTTCAAACGCATCAGCTCCT and GapBA3840R: CGGAGCATACTTGGAATTTCTG) that flanked the missing sequence.

Analysis of the ratio of nonsynonymous and synonymous mutations

Calculation of the ratio (ω) of nonsynonymous (d_N) and synonymous (d_S) distances relative to the consensus sequences is widely considered to provide an estimate of the selective pressure on the encoded proteins. A maximum-likelihood (ML) method that utilizes models of sequence evolution that can calculate ω ratios to identify amino acid sites that are conserved, neutral, or positively selected, were calculated using maximum likelihood ratio method implemented in the CODEML program from the Maximum Likelihood program (PAML) package (Yang 1997; Yang et al. 2000). Instead of assuming that all sites are under the same selection pressure with the same underlying d_N/d_S ratio, this program allows for variable selection intensity among amino acid sites. The package contains several models that account for a range of different statistical distributions of ω ratios between codons. In this study we compared four different models of codon substitution: M1 (neutral) versus M2 (selection) and M7 (beta) versus M8 (beta + ω). Null models M1 and M7 do not allow for the existence of positively selected sites since ω ratios are fixed or estimated between 0 and 1, whereas models M2 and M8 account for positively selected sites by using parameters that estimate as $\omega > 1$. Significance of positively selection is confirmed by a likelihood ratio test (LRT) between the null models and those that account for positive selection.

The LRT statistic approximately follows a chi-square distribution and the number of degrees of freedom is equal to the number of additional parameters in the more complex model. Sites with an excess of non-synonymous substitutions over synonymous substitutions ($\omega > 1$) are good candidates for being subject to positive natural selection at the molecular level. Bayesian probabilities were calculated for all of the sites and only those sites with $\omega > 1$ and a posterior probability ≥ 0.95 were regarded as likely to be under positive, diversifying selection. For each data set the phylogenetic trees required as input were generated using the DNA distance method implemented in PHYLIP (Felsenstein, 2002). Both M1–M2 and M7–M8 comparisons were performed with two degrees of freedom. Model M0 was included to provide an initial value of estimated branch lengths to increase the speed of implementation of the likelihood iterations of the other models.

Please cite this article as: de Villiers, E.P., et al., Phylogenomic analysis of 11 complete African swine fever virus genome sequences, Virology (2010), doi:10.1016/j.virol.2010.01.019

Acknowledgments

We are grateful to INIA-Spain, for the funding of this research and Dr. Oliver Preisig at Inqaba Biotec, South Africa for technical assistance with E75 genome assembly and closure. This is ILRI publication number IL-200908.

The work was funded by the Spanish MCINN-INIA grant RTA 2005-00017. The donor INIA did not play any role in study design. The study was developed together with collaborators at CISA-INIA, Valdeolmos.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.virol.2010.01.019.

References

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402.

Brodie, R., Smith, A.J., Roper, R.L., Tcherepanov, V., Upton, C., 2004. Base-by-base: single nucleotide-level analysis of whole viral genome alignments. BMC Bioinformatics 5, 96.

Bastos, A.D.S., Penrith, M.L., Cruciere, C., Edrich, J.L., Hutchings, G., Roger, F., Couacy-Hymann, E., Thompson, G.R., 2003. Genotyping field strains of African swine fever virus by partial p72 gene characterisation. Arch. Virol. 148, 693–706.

Blasco, R., Aguero, M., Almendral, J.M., Vinuela, E., 1989a. Variable and constant regions in African swine fever virus DNA. Virology 168, 330–338.

Blasco, R., de la Vega, I., Almazan, F., Aguero, M., Vinuela, E., 1989b. Genetic variation of African swine fever virus: variable regions near the ends of the viral DNA. Virology 173, 251–257.

Boinas, F.S., Hutchings, G.H., Dixon, L.K., Wilkinson, P.J., 2004. Characterization of pathogenic and non-pathogenic African swine fever virus isolates from Ornithodoros erraticus inhabiting pig premises in Portugal. J. Gen. Virol. 85, 2177–2187.

Burrage, T.G., Lu, Z., Neila, J.G., Rock, D.L., Zsak, L., 2004. African swine fever virus multigene family 360 genes affect virus replication and generalization of infection in Ornithodoros porcinus ticks. J. Virol. 78, 2445–2453.

Carver, T.J., Rutherford, K.M., Berriman, M., Rajandream, M.A., Barrell, B.G., Parkhill, J., 2005. ACT: the Artemis comparison tool. Bioinformatics 21, 3422–3423.

Carrascosa, A.L., Santarén, J.F., Viñuela, E., 1982. Production and titration of African swine fever virus in porcine alveolar macrophages. J. Virol. Methods 3, 303–310.

Chapman, D.A.G., Tcherepanov, V., Upton, C., Dixon, L.K., 2008. Comparison of the genome sequences of apathogenic and pathogenic African swine fever virus isolates. J. Gen. Virol. 89, 397–408.

Clamp, M., Cuff, J., Searle, S.M., Barton, G.J., 2004. The Jalview Java Alignment Editor. Bioinformatics 20, 426–427.

Dixon, L.K., 1988. Molecular cloning and restriction enzyme mapping of an African swine fever virus isolate from Malawi. J. Gen. Virol. 69, 1683–1694.

Dixon, L.K., Abram, A.C., Miskin, J.E., Parkhouse, M.E., 1999. African swine fever virus: can current research lead to vaccine development? Outlook Agric. 28, -187.

Dixon, L.K., Abrams, C.C., Bowick, G., Goatley, L.C., Kay-Jackson, P.C., Chapman, D., Liverani, E., Nix, R., Silk, R., Zhang, F., 2004. African swine fever virus proteins involved in evading host defence systems. Vet. Immunol. Immunopathol. 100, 117–134.

Ehlers, A., Osborne, J., Slack, S., Roper, R.L., Upton, C., 2002. Poxvirus orthologous clusters (POCs). Bioinformatics 18, 1544–1545.

Felsenstein, J., 2002. Phylogenetic Inference Package (PHYLIP), Version 3.6. University of Washington, Seattle.

Fitzpatrick, D.A., Logue, M.E., Stajich, M.E., Butler, G.A., 2006. A fungal phylogeny based on 42 complete genomes derived from supertree and combined gene analysis. BMC Evol. Biol. 6, 99 doi: 10.118/1471.

Gallardo, C., Mwaengo, D.M., Macharia, J.M., Arias, M., Taracha, E.A., Soler, A., Okoth, E., Martín, E., Kasiti, J., Bishop, R.P., 2009. Enhanced discrimination of African swine fever virus isolates through nucleotide sequencing of the p54, p72, and pB602L (CVR) genes. Virus Genes 38, 85–95.

Gontcharov, A.A., Marin, B., Melkonian, M., 2004. Are combined analyses better than single gene phylogenies? A case study using SSU rDNA and rbcL sequence comparisons in the Zygnematophyceae (Streptophyta). Mol. Biol. Evol. 21, 612–624.

Haresnape, J.M., Wilkinson, P.J., 1989. A study of African swine fever virus infected ticks (Ornithodoros moubata) collected from three villages in the African swine fever enzootic area of Malawi following an outbreak of the disease in domestic pigs. Epidemiol. Infect. 102, 507–522.

Hess, W.R., 1981. African swine fever: a reassessment. Adv. Vet. Sci. 25, 39–69.

Irusta, P.M., Borca, M.V., Kutish, G.F., Lu, Z., Caler, E., Carrillo, C., Rock, D.L., 1996. Amino acid tandem repeats within a late viral gene define the central variable region of African swine fever virus. Virology 220, 20–27.

Kumar, S., Tamura, K., Nei, M., 2004. MEGA3: Integrated Software for Molecular Evolutionary Genetics Analysis and Sequence Alignment Briefings in Bioinformatics. 5, 150–163.

Ley, V., Almendral, J.M., Carbonero, P., Beloso, A., Vinuela, E., Talavera, A., 1984. Molecular cloning of African swine fever virus DNA. Virology 133, 249–257.

Li, L., Stoeckert, C.J., Roos, D.S., 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 13, 2178–2189.

Lubisi, B.A., Bastos, A.D.S., Dwarka, R.M., Vosloo, W., 2005. Molecular epidemiology of African swine fever in East Africa. Arch. Virol. 150, 2439–2452.

Malmquist, W.A., Hay, D., 1960. Hemadsorption and cytopathic effect produced by African swine fever virus in swine bone marrow and buffy coat cultures. Am. J. Vet. Res. 21, 104–108.

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.J., Chen, Z., et al., 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature 437, 376–380.

Montgomery, R.E., 1921. A form of swine fever occurring in British East Africa (Kenya Colony). J. Comp. Pathol. 34, 159–191.

Notredame, C., Higgins, D.G., Heringa, J., 2000. T-coffee: a novel method for fast and accurate multiple sequence alignment J. Mol. Biol. 302, 205–217.

Pan, I.C., 1992. African swine fever virus—generation of subpopulations with altered immunogenicity and virulence following passage in cell-cultures. J. Vet. Med. Sci. 54, 43–52.

Penrith, M.L., Thompson, G.R., Bastos, A.D.S., 2004. African swine fever, In: Coetzer, J.A.W., Tustin, R.C. (Eds.), Infectious Diseases of Livestock with Special Reference to Southern Africa, 2nd ed. Oxford Univ. Press, Cape Town, London, New York, pp. 1087–1119.

Rokas, A., Williams, B.L., King, N., Carroll, S.B., 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425, 798–804.

Rowlands, R.J., Michaud, H.L., Hutchings, G., Oura, C., Vosloo, W., Dwarka, R., Onashvili, T., Albina, E., Dixon, L.K., 2008. African swine fever virus isolate, Georgia, 2007. Emerg. Infect. Dis. 14, 1870–1874.

Rowlands, R.J., Duarte, M.M., Boinas, F., Hutchings, G., Dixon, L.K., 2009. The CD2v protein enhances African swine fever virus replication in the tick vector, Ornithodoros erraticus. Virology 393, 319–328.

Sanderson, M.J., Driskell, A.C., Ree, R.H., Eulenstein, O., Langley, S., 2003. Obtaining maximal concatenated phylogenetic data sets from large sequence databases. Mol. Biol. Evol. 20, 1036–1042.

Sumption, K.J., Hutchings, G.H., Wilkinson, P.J., Dixon, L.K., 1990. Variable regions on the genome of Malawi isolates of African swine fever virus. J. Gen. Virol. 71, 2331–2340.

Tabares, E., Olivares, I., Santurde, G., Garcia, M.J., Martin, E., Carnero, M.E., 1987. African swine fever virus DNA: deletions and additions during adaptation to growth in monkey kidney cells. Arch. Virol. 97, 333–346.

Tcherepanov, V., Ehlers, A., Upton, C., 2006. Genome Annotation Transfer Utility (GATU): rapid annotation of viral genomes using a closely related reference genome. BMC Genomics 7, 150.

Wesley, R.D., Tuthill, A.E., 1984. Genome relatedness among African swine fever field isolates by restriction endonuclease analysis. Prev. Vet. Med. 2, 53–62.

Yanez, R.J., Rodriguez, J.M., Nogal, M.L., Yuste, L., Enriquez, C., Rodriguez, J.F., Vinuela, E., 1995. Analysis of the complete nucleotide-sequence of African swine fever virus. Virology 208, 249–278.

Yang, Z., 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. 13, 555–556.

Yang, Z., Nielsen, R., Goldman, N., Pedersen, A.M.K., 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics 155, 431–449.

Yozawa, T., Kutish, G.F., Afonso, C.L., Lu, Z., Rock, D.L., 1994. Two novel multigene families, 530 and 300, in the terminal variable regions of African swine fever virus genome. Virology 202, 997–1002.

Ziheng, Y., Wong, S.W., Nielsen, R., 2005. Bayes empirical Bayes inference of amino acid sites under positive selection. Mol. Biol. Evol. 22, 1107–1118.

Zsak, L., Lu, Z., Burrage, T.G., Neilan, J.G., Kutish, G.F., Moore, D.M., Rock, D.L., 2001. African Swine fever virus multigene family 360 and 530 genes are novel macrophage host range determinants. J. Virol. 75, 3066–3076.

Zsak, L., Borca, M.V., Risatti, G.R., Zsak, A., French, R.A., Kutish, G.F., Neilan, J.G., Callahan, J.D., Nelson, W.M., Rock, D.L., 2005. Preclinical diagnosis of African swine fever in contact-exposed swine by a real-time PCR assay. J. Clin. Microbiol. 43, 112–119.

Please cite this article as: de Villiers, E.P., et al., Phylogenomic analysis of 11 complete African swine fever virus genome sequences, Virology (2010), doi:10.1016/j.virol.2010.01.019