bioRxiv preprint doi: https://doi.org/10.1101/460832; this version posted November 4, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Homologous recombination shapes the genetic diversity and drives the evolution of African swine fever viruses

Zhaozhong Zhu[1, #], Zena Cai[1, #], Congyu Lu[1], Zheng Zhang[1], Yunshi Fan[1], Gaihua Zhang[2], Taijiao Jiang[3, 4], Yongjun Tan[1], Chaoting Xiao[1,*], Yousong Peng[1,*]

[1] College of Biology, Hunan University, Changsha, China

[2] College of Life Sciences, Hunan Normal University, Changsha 410081, China

[3] Center of System Medicine, Institute of Basic Medical Sciences, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing, China

[4] Suzhou Institute of Systems Medicine, Suzhou, China

# These authors contributed equally to this work

* To whom correspondence should be addressed. Email: xiaocht@126.com (CX), pys2013@hnu.edu.cn (YP)

Abstract

Recent outbreaks of African swine fever virus (ASFV) in China severely disrupted the swine industry of the country. No vaccine or treatment against ASFV is available in the current. How to effectively control the virus is challenging. Here, by analyzing all ASFV genomes publicly available, we found large genetic diversity among ASFV genomes. Interestingly, they were mainly caused by extensive genomic insertions

bioRxiv preprint doi: https://doi.org/10.1101/460832; this version posted November 4, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

and deletions (indels) instead of genomic mutations. Genomic diversity resulted in proteome diversity, with one-third to half of proteins variable among ASFV proteomes. Besides, nearly 20% of proteins had replications in adjacent positions. Further analysis identified extensive homologous recombination in the ASFV genomes, which is consistent with the occurrence of indels. Repeated elements of 15~50 bp were widely distributed in ASFV genomes, which may facilitate the occurrence of homologous recombination. Moreover, two homologous recombination-related enzymes, the recombinase and DNA topoisomerase, were found to keep conserved in all ASFVs analyzed here. This work highlights the importance of homologous recombination in evolution of the virus, and thus facilitates the prevention and control of it.

Introduction

African swine fever (ASF) virus (ASFV), the causative agent of ASF, is a complex, large, icosahedral multi-enveloped DNA virus. It is classified as the only member of the family Asfarviridae [1] [2]. The genome of the virus belong to double-stranded DNA, the size of which range from 170 kb to 190 kb[3]. ASFV mainly infect suids, which include domestic pigs and wild boars, and soft ticks. The former was reported to be natural host for the virus [4,5]. ASFV was firstly discovered in Kenya in 1921 [6]. It remained restricted to Africa until 1957 when it was reported in Portugal. Until

bioRxiv preprint doi: https://doi.org/10.1101/460832; this version posted November 4, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

now, the virus has caused outbreaks in more than fifty countries in Africa, Europe, Asia and South America [5]. The latest reports show that the virus caused outbreaks in more than ten provinces in China [7,8]. Because of high lethality caused ASFV in domestic pigs, massive culling campaigns and pig movement restrictions are the most commonly used strategies for control of the virus, which often cause huge loss on pig production and people’s livelihoods [4]. Unfortunately, no vaccine or treatment against ASFV is available in the current. Development of more effective method for control of the virus is challenging.

Many efforts have been devoted to develop a vaccine for the ASFV [1,4,9-11], but nearly all these attempts failed. One of the most important reasons is the complex composition of antigen proteins [4,12]. Previous reports show that p72, p30, and p54 were three of the most antigenic proteins during infection, but the immunity against them could only provide partial protection [12,13]. Many other proteins or other factors such as phospholipid composition may influence the antigen of the virus [12]. Therefore, it is necessary to understand the mechanisms of antigen diversity of the virus [1].

Many studies have investigated the genetic diversity of ASFVs. It is reported that the virus genome encodes over 150 proteins, including lots

bioRxiv preprint doi: https://doi.org/10.1101/460832; this version posted November 4, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

of viral enzymes, viral transcription and replication-related proteins, structural proteins and other proteins involved in assembly, proteins involved in modulating host defence, and so on [3,14,15]. For example, the transcription of the virus occurs independently of the host RNA polymerase because the virus contains relevant enzymes and factors [3]. The ASFV genome contains a conserved central region of about 12 kb and two variable ends, which account for the variable size of the genome [3,16,17]. There are significant variations between ASFV genomes due to genomic insertion or deletion, such as the deletion of multigene family (MGF) members[3]. Although much progress have been made on genetic diversity of the virus, the extent and mechanisms are still far from clear. Besides, most of these studies investigate the genetic diversity on only some common genes, such as p72 and p54 [18,19], or only used one or several isolate genomes [3,16,17]. The number of viral genomes has increased rapidly as the development of DNA sequencing technology. A comprehensive study of genetic diversity of ASFVs is necessary.

Homologous recombination has been reported to occur in several groups of viruses [20-23], such as herpesvirus, retroviruses, and coronaviruses, and play an important role in viral evolution [21]. A few studies on several ASFV genes have suggested occurrence of recombination in evolution of ASFVs [3,18]. A comprehensive and genomic scale study of homologous

bioRxiv preprint doi: https://doi.org/10.1101/460832; this version posted November 4, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

recombination in ASFV is lacking, and the role of recombination on genetic diversity of the virus is still unknown. Here, by analyzing all ASFV genomes publicly available, we systematically investigated the genomic diversity and homologous recombination of ASFVs, and found the former could be largely attributed to the latter. This work would help understand the evolution of the virus and thus facilitate prevention and control of it.

Materials and Methods

1 ASFV genome and alignment

All the ASFV genomic sequences with over 160000 bp were obtained from NCBI GenBank database on October 7, 2018 [24]. After removing the genomic sequence derived from a patent, a total of 36 ASFV genomes were kept for further analysis. They were then aligned by MAFFT (version 7.127b) [25]. For robustness of the results, the traditional tool of CLUSTAL (version 2.1) [26] was also used for aligning these genome sequences.

2 ORF prediction

To obtain the proteins encoded by ASFV genomes, each genome was searched against all the ASFV protein sequences obtained from NCBI

bioRxiv preprint doi: https://doi.org/10.1101/460832; this version posted November 4, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

protein database on October 7, 2018, with the help of blastx [27]. All genomic regions with significant hits (e-value < 0.001) were checked with a perl script: overlapping regions in the same coding frame were merged to obtain ORFs as long as possible; regions without start codon or stop codon were extended upstream or downstream to search for the start or stop codon. Then, those genomic regions with significant hit, with both start and stop codons, and with over 90 bps were defined as ORFs. They were then translated into proteins with a perl script, and were used for further analysis.

3 Protein groups

All the proteins encoded by ASFV genomes were grouped based on sequence homology with the help of OrthoFinder (version 2.2.7) [28] with the default parameters. Manual check was conducted to ensure that each protein group contains one kind of protein. Four protein groups were found to contain sequences which are part of those in other groups. Therefore, they were merged to other groups.

4 Alignment of ASFV proteome

A ASFV proteome was defined as all proteins encoded by the ASFV genome. Because both the plus and minus strands could code proteins, proteins in a proteome were separated into plus and minus proteome

bioRxiv preprint doi: https://doi.org/10.1101/460832; this version posted November 4, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

based on their coding directions. To align the ASFV proteomes, the name of proteins in the proteome were replaced by the names of protein groups. Then, the ASFV proteomes were aligned with the dynamic programming algorithm. Manual check was conducted to ensure no mismatch of protein groups in the alignment.

5 Function prediction of ASFV proteins

Functions of ASFV proteins were inferred based on protein group. For each protein group, the longest protein sequence was selected as a representative of this protein group. InterproScan (version 5) [29] was used to infer the function of the representative protein sequences. The TMHMM Server (version 2.0) [30] was used to predict whether a representative protein has a trans-membrane helix. Membrane proteins were defined as those have at least one trans-membrane helixes. Besides, the representative proteins were searched against the NCBI non-redundant protein database with the help of PSI-BLAST [31]. After excluding the hits of ASFV, the function of the best hit was assigned to the representative protein.

6 Functional classification of ASFV proteins

Five antigen-related proteins, i.e., p54, p72, pEP153R, pCP204L, pEP402R, were adapted from Rock's work [32]; four host immunity

bioRxiv preprint doi: https://doi.org/10.1101/460832; this version posted November 4, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

evasion-related proteins, i.e., pA238L, MGF505/530, pEP152R and 4CL were adapted from Fraczyk and Reis's work [18,33]; five virulence-related proteins, i.e., pA240L, pB119L, pK196R, pDP71L, DP96R, were adapted from Rock's work [32].

7 Detection of recombination events

RDP (version 4) [34] was used to infer the recombination events based on the aligned ASFV genomes. Multiple methods in RDP were used. Only those recombination events detected by at least two methods were used for further analysis.

8 Phylogenetic tree inference and visualization

Maximum-likelihood phylogenetic trees were inferred with the help of MEGA (version 5.0) [35] with the default parameters. Bootstrap analysis was conducted using 100 replicates. The phylogenetic tree was visualized with the help of Denscrope (version 2.4) [36].

9 Statistics analysis

All the statistical analysis were conducted in R (version 3.2.5) [37].

Results

bioRxiv preprint doi: https://doi.org/10.1101/460832; this version posted November 4, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

1 ASFV genomes

A total of 36 genome sequences of ASFVs were obtained from NCBI GenBank database, which were listed in Table 1. They were mainly isolated from Africa and Europe during the years from 1950 to 2016. The genome size of them ranged from 170101 bp to 193886 bp, with an average of 185800 bp. The viral isolate Kenya50 had the largest genome, while the isolate BA71V had the smallest genome. No increasing or decreasing trend of the genome size was observed from 1950 to 2017 (Figure 1A), suggesting dynamic changes of the viral genome size.

Table 1: The viruses used in this study.

Isolate name	Accession number	Isolation year	Isolation country	Genome size (bp)
Kenya50	AY261360.1	1950	Kenya	193,886
Portugal60	KM262844.1	1960	Portugal	182,362
Malawi62	AY261364.1	1962	Malawi	185,689
Portugal68	KM262845.1	1968	Portugal	172,051
Spain71	KP055815.1	1971	Spain	180,365
BA71V-nonvirulent	NC_001659.2	1971	Spain	170,101
BA71V	U18466.2	1971	Spain	170,101
Spain75	FN557520.1	1975	Spain	181,187
SouthAfrica79	AY261362.1	1979	South Africa	192,714
Namibia80	AY261366.1	1980	Namibia	186,528
Malawi83	AY261361.1	1983	Malawi	187,612
SouthAfrica87	AY261365.1	1987	South Africa	190,773
Portugal88	AM712240.1	1988	Portugal	171,719
SouthAfrica96	AY261363.1	1996	South Africa	190,324
Benin97	AM712239.1	1997	Benin	182,284
Kenya05	KM111294.1	2005	Kenya	191,058
Kenya06	KM111295.1	2006	Kenya	184,368
Georgia07	FR682468.1	2007	Georgia	189,344

bioRxiv preprint doi: https://doi.org/10.1101/460832; this version posted November 4, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Italy08	KX354450.1	2008	Italy	184,638
Italy10	KM102979.1	2010	Italy	182,906
Russia13	KJ747406.1	2013	Russia	189,387
Estonia14	LS478113.1	2014	Estonia	182,446
Russia14	KP843857.1	2014	Russia	189,333
Poland15	MH681419.1	2015	Poland	189,394
Uganda-TororoDistrict1	MH025919.1	2015	Uganda	188,611
Uganda-R8	MH025916.1	2015	Uganda	188,627
Uganda-R25	MH025918.1	2015	Uganda	188,630
Uganda-R7	MH025917.1	2015	Uganda	188,628
Uganda-TororoDistrict2	MH025920.1	2015	Uganda	188,629
Pol16-o10	MG939585.1	2016-2017	Poland	189,405
Pol16-o7	MG939583.1	2016-2017	Poland	189,401
Pol16-o23	MG939586.1	2016-2017	Poland	189,393
Pol17-C210	MG939588.1	2016-2017	Poland	189,401
Pol17-C220	MG939589.1	2016-2017	Poland	189,393
Pol16-o9	MG939584.1	2016-2017	Poland	189,399
Pol17-C201	MG939587.1	2016-2017	Poland	189,405

2 Genomic diversity of ASFVs

To obtain the genomic differences between ASFV genomes, pairwise comparisons between ASFV genomes were conducted after genome alignment. The average genomic difference between viruses was 24570 bp, more than 10% of the genome. Interestingly, the genomic differences caused by indels were much larger than those caused by genomic mutation (Figure 1B & Figure S1) in most cases. For example, there was 31833 bp differences between virus Mkuzi79 and BA71V, 78% of them were caused by indels.

We next analyzed the size and position of indels in ASFV genomes. As shown in Figure 1C, the occurrence of indels was much more frequent in

bioRxiv preprint doi: https://doi.org/10.1101/460832; this version posted November 4, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

both end of the genome, especially in the 5’ end. Besides, the size of indels in both ends was also much larger than that in other regions. Large indels with over 500 bp happened mostly in the 5’ end. Attention to note is that there was some extent of variation in the conserved central region (marked in black arrow) reported previously.

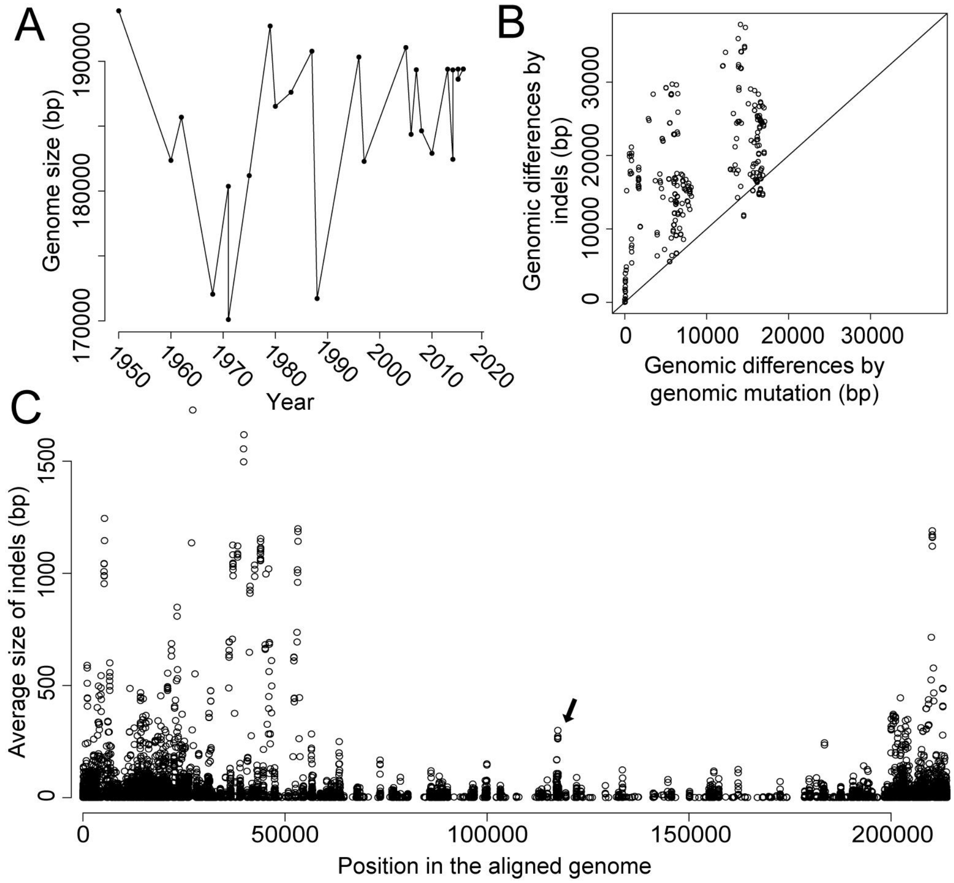


Figure 1. Genomic differences between ASFV genomes. (A) The genome size variation along the isolation time of the virus. (B) Genomic differences caused by genomic mutations and indels. (C) The size and location of indels along the aligned genome. For each position, the average size of indels covering the position was shown.

bioRxiv preprint doi: https://doi.org/10.1101/460832; this version posted November 4, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

3 Proteome diversity of ASFV

We continued to compare the ASFV genomes by genes. The candidate ORFs encoded by ASFV genomes were obtained (Materials and Methods) (Table S1). The plus strand encoded 83~107 proteins, with an average of 96 proteins; the minus strand encoded 76~95 proteins, with an average of 88 proteins. Taken together, the ASFV genome encoded 159~200 proteins, with an average of 184 proteins. The viral isolate Russia13 encoded most proteins, although it was not the largest genome. The ratio of encoding region in each genome ranged from 65% to 78%.

Then, we identified the ortholog or paralog groups based on homology. A total of 214 protein groups plus 24 singletons could be obtained (Table S2). The number of proteins in each group ranged from 2 to 91. Only half of protein groups (105) were observed in all 36 ASFV viruses, which could be considered as core protein sets of the virus. They were mainly coded by central regions in both plus and minus strands (Figure 2). The core protein groups included 30 groups related to replication and transcription (such as pO174L, pD205R and CP80R), 23 groups of membrane protein (such as p54, p12 and p17) (marked with asterisks in Figure 2), 10 groups of enzyme (such as pS273R, pEP424R and pI215L), 3 groups of virulence determinants (pA240L, pB119L and pK196R), 3

bioRxiv preprint doi: https://doi.org/10.1101/460832; this version posted November 4, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

groups related to host immunity evasion (pA238L, pEP152R and MGF505/530), and so on (Figure 3). The remaining 109 protein groups were variable and observed in 2~35 viruses, 77 of which were observed in no less than half of viruses. Compared to the core protein groups, 60% of variable protein groups had unknown functions (colored in black in Figure 3). Interestingly, three antigen-related protein groups (pEP153R, pCP204L and pEP402R), two virulence-related protein groups (pDP71L and DP96R), and eight replication and transcription-related protein groups (such as pG1211R and F334L) were also observed in the variable protein groups (Figure 3). Besides, they included 39 membrane protein groups (marked with asterisks in Figure 2).

Forty-four protein groups were observed to have paralogs in at least one virus (colored in Figure 2). They were mostly located in both ends of the genome in both plus and minus strand, especially in the 5’ end of the plus strand and 3’ end of the minus strand. Besides six MGF protein groups, they were involved in multiple functions mentioned above (Table S2). For example, the antigen-related protein group pEP153R had two paralogs in seven viruses (marked with a black arrow in the middle of plus strand in Figure 2). Most paralogs were clustered in adjacent positions. Exceptions were observed for some protein groups which included proteins encoded by the first one~three thousands nucleotides in

bioRxiv preprint doi: https://doi.org/10.1101/460832; this version posted November 4, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

the plus and minus strands, such as the protein group “p01990-3L” (colored in dark green and located in both the most left-side and right-side). Further analysis shows that for most viral genomes, there were a segment of 200~3000 bp which were exactly the same in the beginning of the plus and minus strands (Table S3).

Extensive insertion and deletions of proteins were observed when comparing the aligned proteome of ASFVs. The number of different proteins encoded by the plus strand between viruses ranged from 0 to 60, with an average of 30; those encoded by minus strand between viruses ranged from 0 to 39, with an average of 19. The total number of different proteins between viruses ranged from 0 to 93, with an average of 49, which was about one-fourth of the proteome. By functional class, there was a median difference of 22 membrane proteins between ASFVs, suggesting large diversity of membrane proteins among ASFVs. Besides, a median difference of 2, 2, 1 and 1 proteins involved in replication and transcription, host immunity evasion, antigen and virulence, respectively, were observed between ASFVs.

bioRxiv preprint doi: https://doi.org/10.1101/460832; this version posted November 4, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

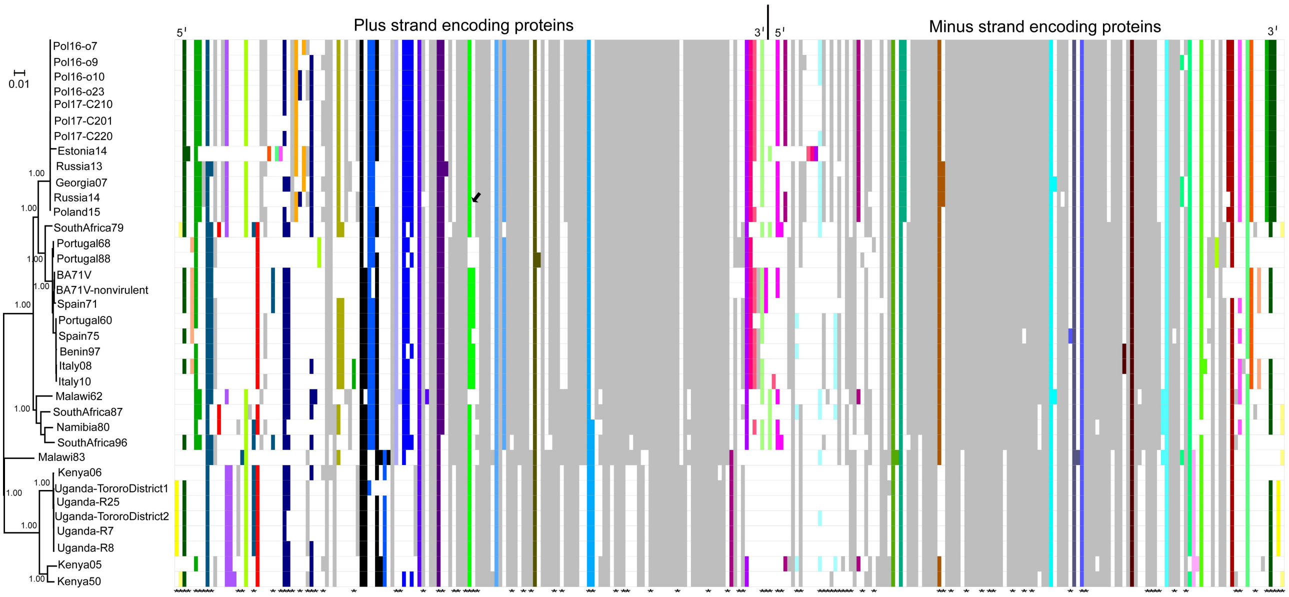


Figure 2. The phylogenetic tree of ASFVs (left side) and the alignment of their proteomes in plus and minus strand (right side). Each row refers to the proteome of the virus in the phylogenetic tree; each colume refers to one protein group. Protein groups with at least two paralogs in at least one virus were colored, while those with only one paralog were colored in gray. “White” refers to no protein group in the virus. Asterisks refer to membrane proteins. For clarity, the singletons were ignored in the alignment.

bioRxiv preprint doi: https://doi.org/10.1101/460832; this version posted November 4, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

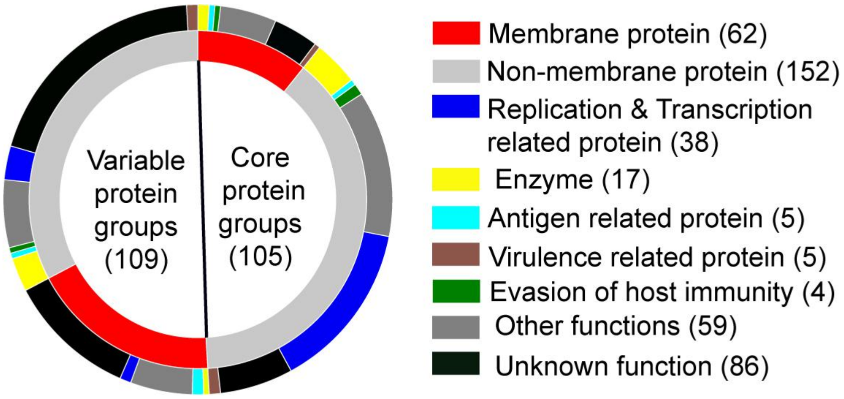


Figure 3. Functional classification of 214 protein groups.

4 Extensive recombination among ASFV genomes

Then, we investigated the mechanisms of indel in ASFV genomes. Recombination analysis showed that all ASFVs experienced 3~22 recombination events (Figure 4 & Table S4). The virus isolate SouthAfrica79 experienced the most recombination events. On average, each virus experienced 11 recombination events. The size of recombination region ranged from 174 to 22628 bp. The ratio of recombination region in the whole genome ranged from 2% to 27%. Most recombination events happened in both ends, especially the 5’ end. Interestingly, when aligning the mapping of recombination events and the ratio of gap in the genome, large consistence was observed. Nearly all recombination events happened in or near the gap-rich regions where indels happened.

bioRxiv preprint doi: https://doi.org/10.1101/460832; this version posted November 4, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

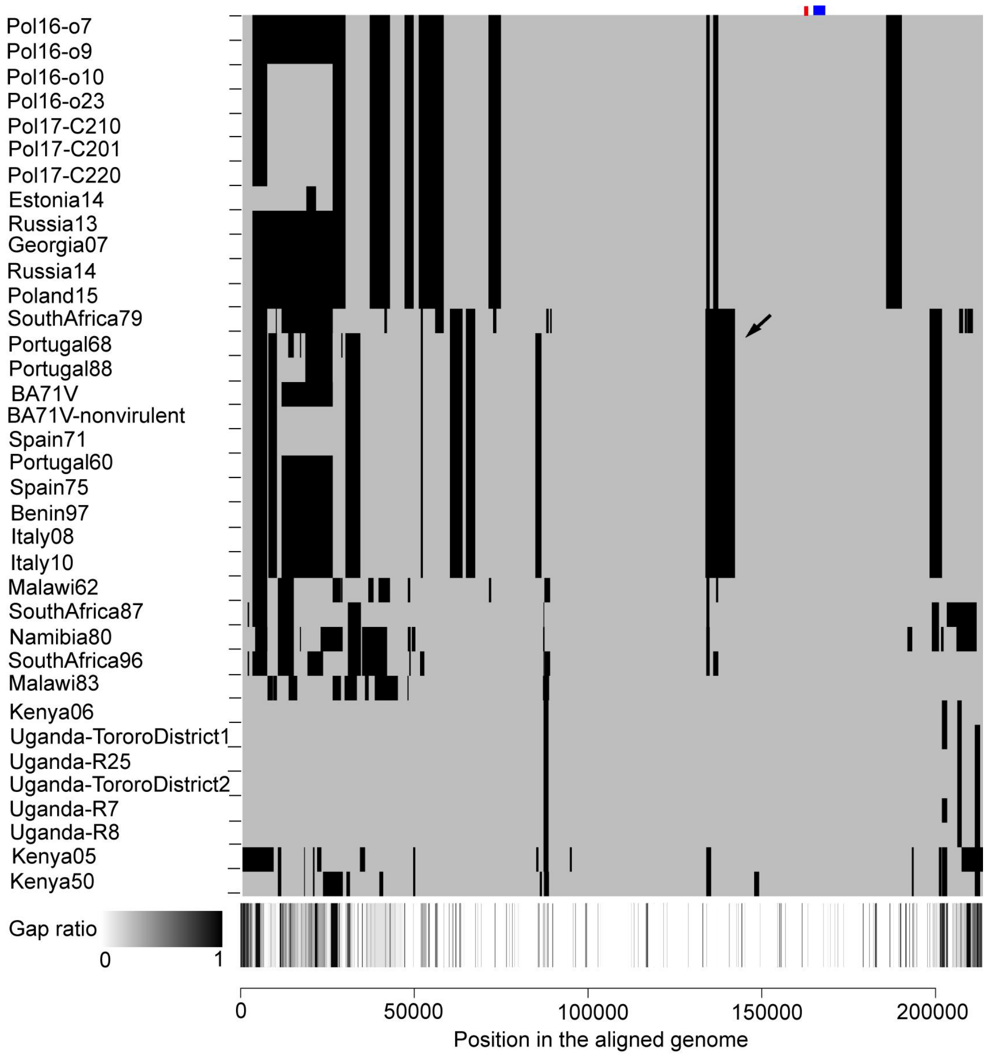
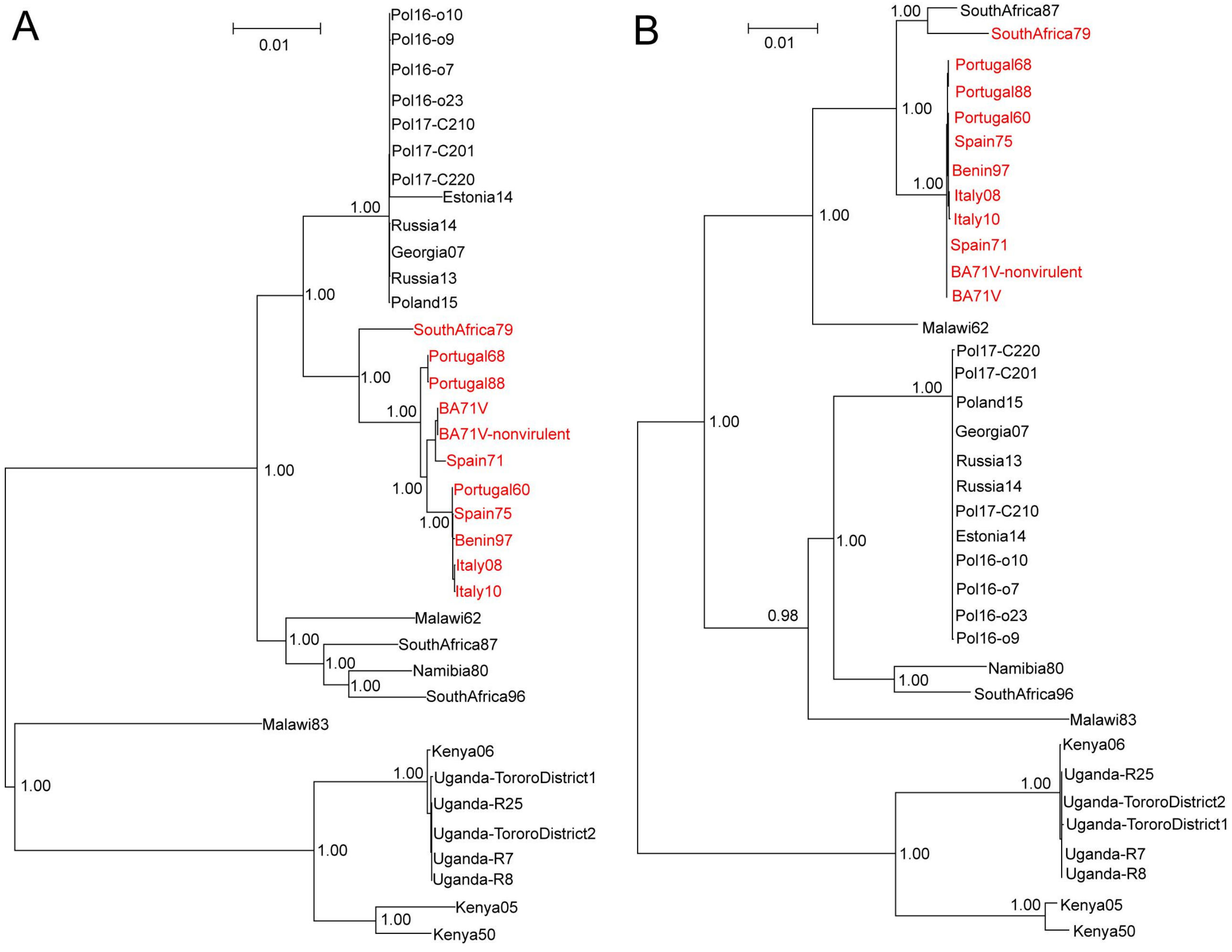


Figure 4. Recombination of ASFV genomes. The black area refer to the recombination region for each genome. The bottom panel shows the ratio of gap, was colored according to the legend, in each position of the aligned genomes. The red and blue rectangles in the top-right refer to the coding region of recombinase and DNA topoisomerase, respectively. The black arrow refers to the recombination event displayed in Figure 5.

Figure 5 illustrated a recombination event happened in 11 viral isolates

bioRxiv preprint doi: https://doi.org/10.1101/460832; this version posted November 4, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(color in red), which included two virus from Africa (SouthAfrica79 and Benin97) and nine viruses from Portugal, Spain and Italy. They formed a separate lineage in the phylogenetic tree. The recombination region ranged from 133683 to 142222 bp, located in the central conserved region of the genome (the black arrow in Figure 4). In the phylogenetic tree built with genomic sequences without this recombination event, the recombinant virus are neighbors of a clade containing viruses from Eastern Europe countries, while in that built with genomic sequences with the recombination event, they are descendants of Malawi62 from Africa.



bioRxiv preprint doi: https://doi.org/10.1101/460832; this version posted November 4, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 5. An example of recombination events happened in twelve ASFVs (colored in red). (A) and (B) refer to the maximum-likelihood phylogenetic trees built with genome sequences without and with the recombination event (133683~142222 in the aligned genome), respectively. The numbers refer to the bootstrap values of nodes in 100 bootstraping test.

Next, we analyzed the proteins involved in the recombination events. A total of 85 protein groups and 7 singletons were involved, including 24 core protein groups and 61 variable protein groups. Among them, 24 protein groups belong to membrane proteins. Three antigen-related protein groups (pEP153R, pEP402R and pCP204L), three host immunity evasion-related protein groups (pEP152R, MGF505/530 and pA238L) and ten replication and transcription-related protein groups were involved in the recombination events. Many of them belong to the core protein sets, such as pA238L and pEP152R.

5 Conserved recombinase and DNA topoisomerase in ASFVs

Interestingly, we found a recombinase existing in all ASFV isolates. It has 345 amino acids and was encoded by the minus strand. It contains the YqaJ-like viral recombinase domain. The recombinase was highly conserved in these viruses, with an average sequence identity of 96.7%

bioRxiv preprint doi: https://doi.org/10.1101/460832; this version posted November 4, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

between ASFVs.

Besides, topoisomerase is also reported to be related to homologous recombination. We also found a DNA topoisomerase existing in all ASFV isolates. It was a protein of 1192 amino acids, and was coded by the plus strand. Although they were encoded by different strands, interestingly, the DNA topoisomerase and recombinase were encoded by genomic sequences in adjacent regions: the former was encoded in 162236～163273 (colored in blue),while the latter was encoded in 164743～168319 (colored in red) (**Figure** 4). The DNA topoisomerase protein is also conserved in these viruses, with an average sequence identity of 98.3% between that of ASFVs. Both the enzymes were not involved in any recombination events. The phylogenetic trees for them were similar to that of the genome (Figure S2).

6 An abundance of repeated elements in ASFV genomes

Repeated elements were reported to be important for homologous recombination. We identified the repeated elements ranging from 5~100 bp. As shown in Figure 6A, the number of repeated elements in each genome increased as the size of elements ranging from 5 to 9. Then, it began to decrease monotonously. On average, a genome had 400~1700 repeated elements with size ranging from 15~30bp. We then investigated the distances between adjacent elements for a given repeated element

bioRxiv preprint doi: https://doi.org/10.1101/460832; this version posted November 4, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(Figure 6B). The change dynamics of the distance for repeated elements in all genomes versus the size of elements was similar to that for the number of repeated element. However, when the distance reach to the shortest (136 bp), it kept constant at size of 23~46. Then, it began to increase as the size of the repeated elements.

Take the elements of 30 bp as an example. For each genome, they had a median of 427 elements which repeated at least two times in the genome. Some elements appeared in over ten times in the genome, such as the element "AGGCGTTAAACATTAAAATTATTACTACTG" in the viral strain BA71V. The region taken by repeated elements account for 9%~47% of the genome. When analyzing the distance between repeated elements, we found they had a median distance of 136 bp, suggesting they tend to cluster in neighboring regions. Figure 6C shows the distribution of repeated elements in the aligned genome. Most repeated elements were located in both ends of the genome. Besides, there were two clusters of repeated elements around 55000 bp and 120000 bp.

bioRxiv preprint doi: https://doi.org/10.1101/460832; this version posted November 4, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

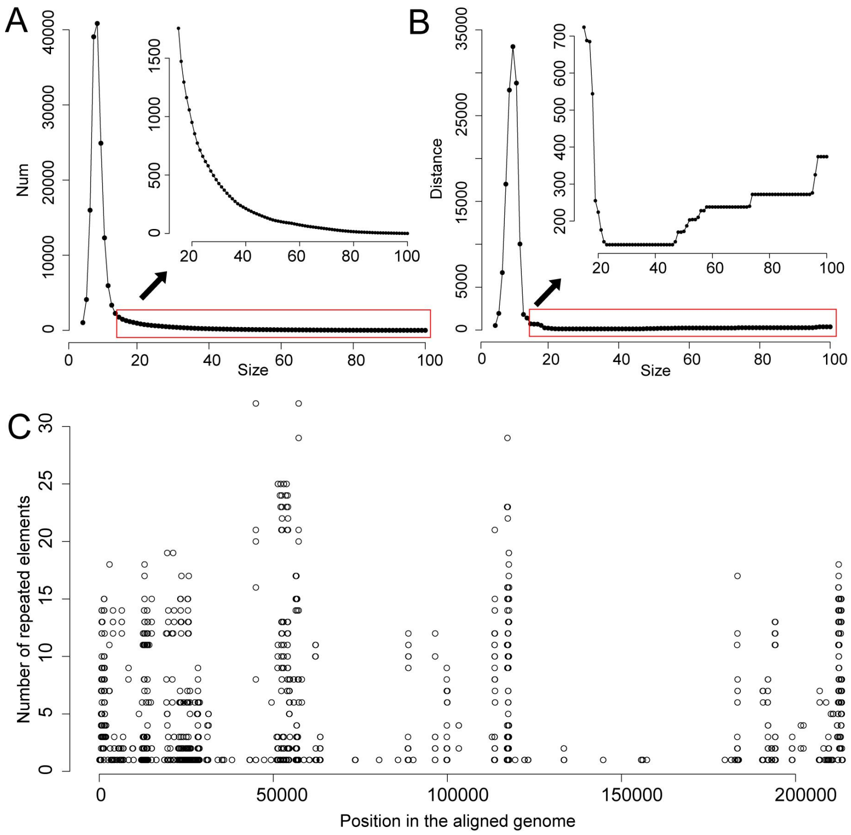


Figure 6. Distribution of repeated elements. (A) The median number of repeated elements in ASFV genomes versus the size of repeated elements. (B) The median distance between repeated elements versus the size of repeated elements. (C) Number of repeated elements of 30 bp happened in each genomic position.

Discussion

This work systematically analyzed the genetic diversity of ASFVs. The large genome enable the virus encode an abundance of proteins. The functions of the virus in its life circle could be accomplished by multiple proteins. For example, 38, 5, and 4 protein groups were respectively involved in DNA replication and transcription, antigen and evasion of

bioRxiv preprint doi: https://doi.org/10.1101/460832; this version posted November 4, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

host immunity. The numbers are very likely to be under-estimated since 40% of proteins had unknown functions. This may facilitate efficient control of host cell and precise regulation of viral activities.

Large differences were observed between the proteomes of ASFVs due to two reasons. On one hand, one-third to one-half of proteins of each ASFV are variable among ASFVs. On the other hand, there were lots of gene replications (or paralogs) in ASFV genomes. Diverse proteome among viruses may lead to diverse phenotype, such as antigen and virulence. This may shape a great challenge for prevention and control of the virus. For example, viruses of diverse antigenicity may need multiple kinds of vaccines because the cross-protection between viruses may be limited.

Although lots of efforts have been devoted to develop vaccines against the ASFV, unfortunately, all of them have been unsuccessful. The failure could be attributed to many factors, such as the absence of neutralizing antibodies, diverse antigen-related proteins, complexity of neutralization, and so on. This work found that besides the core proteins p54 and p72, three other antigen-related proteins (pEP153R, pCP204L and pEP402R) were variable among ASFVs. Besides, a total of 62 membrane proteins were identified, half of which had unknown functions. They are very likely to be antigen-related. More efforts are needed to determine their

bioRxiv preprint doi: https://doi.org/10.1101/460832; this version posted November 4, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

role in stimulating neutralizing antibodies, and the neutralization mechanisms and efficiency. Previous studies showed that immunized pigs with the baculovirus expressed hemagglutinin of ASFV were protected against the viral lethal infection. This suggests that incorporation of more antigen proteins in the vaccine may provide better protection.

Indels were found to contribute much larger to the genetic diversity of ASFVs than genomic mutations did. Compared to genomic mutations, indels could introduce large variation to the genome, and cause a great damage to the genome structures, which may lead to death of the individual organisms. Therefore, few indels were observed for viruses with small genome, such as influenza viruses and HBV. However, for viruses with large genomes, such as ASFV, they are more robust to indels since they have lots of repeated elements and proteins (paralogs). Moreover, indels may provide a more efficient way of survival than genomic mutations under the natural selection pressure, since the virus could rapidly change its phenotype, such as antigen, virulence, or ability of replication and transcription. For example, deletion of some MGF genes could reduce viral replication or virulence, which may help viral infection of soft ticks [3,38].

Homologous recombination could generate indels in the genomes by

bioRxiv preprint doi: https://doi.org/10.1101/460832; this version posted November 4, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

exchanging the segments of unequal length, as is illustrated in Figure 7A. This could be facilitated by frequent occurrence of clustered repeated elements in the ASFV genome (Figure 6). A large concordance was observed between recombination events and indels, suggesting the recombination seems to a very effective way for generating indels in ASFVs. Besides, all ASFVs analyzed in this study contain the recombinase and DNA topoisomerase, which are two of the most commonly observed enzymes responsible for homologous recombination. Both of them were much conserved and didn't experience any recombination, suggesting their important roles for ASFVs. Taken, together, homologous recombination should be taken by ASFV as an effective strategy for generating genetic diversity, which further lead to diverse antigen, virulence, replication and transcription ability, weapons of escaping host immunity of the virus (Figure 7B).

bioRxiv preprint doi: https://doi.org/10.1101/460832; this version posted November 4, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

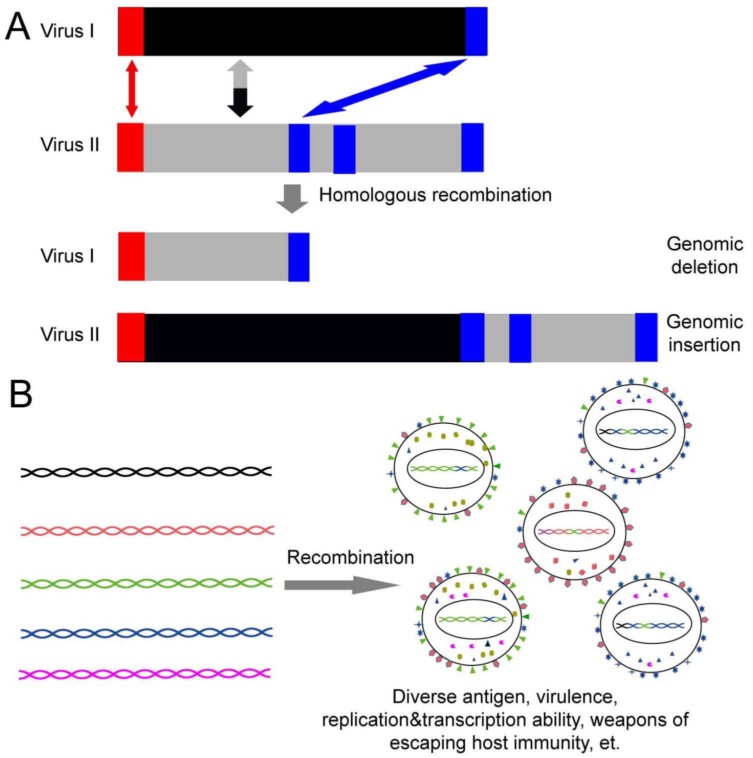


Figure 7. Homologous recombination leads to indels and shape the genetic diversity of ASFVs.

There are some limitations to this study. Firstly, the number of ASFV genomes is limited, which hindered a more comprehensive study of the evolution of ASFV genomes. Fortunately, the isolates covered a long time period from 1950 to 2017, and also covered a large area including Africa and Europe, two major areas of ASFV circulating. They should reflect the genetic diversity of ASFVs to a large extent. Secondly, the location and size of indels observed in ASFV genomes are affected by alignment algorithms. To remove this bias, we have used two different kinds of commonly used methods for generating alignment of ASFV genomes. In both cases, frequent indels were observed and found to play

bioRxiv preprint doi: https://doi.org/10.1101/460832; this version posted November 4, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

a more important role than genomic mutations. Thirdly, the proteome of each ASFV was inferred by computation methods. They may be not incomplete since they only cover about 70% of the genome. Although all of them had significant homology to proteins in NCBI protein database, most of the latter were also predicted without experimental validations. Besides, functions were unknown for 40% of ASFV proteins. Much more efforts were needed to experimentally determine the proteome and their functions of ASFVs [14,15].

Overall, this work provided a systematic view of the genetic diversity of ASFVs.

Extensive homologous recombination were detected and are very likely to cause the widespread indels observed in ASFV genomes, which further lead to large genetic diversity of ASFVs. These would help understand the evolution of the virus and thus facilitate the prevention and control of it.

Acknowledgements

This work was supported by the National Key Plan for Scientific Research and Development of China (2016YFD0500300, 2016YFC1200200 and 2017YFD0500104), the National Natural Science

bioRxiv preprint doi: https://doi.org/10.1101/460832; this version posted November 4, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Foundation of China (31500126 and 31671371) and the Chinese Academy of Medical Sciences (2016-I2M-1-005).

The authors have declared that no competing interests exist.

References

1 Arias, M., Jurado, C., Gallardo, C., Fernandez-Pinero, J. & Sanchez-Vizcaino, J. M. Gaps in African swine fever: Analysis and priorities. *Transboundary and emerging diseases* **65**, 235-247, doi:10.1111/tbed.12695 (2018).

2 Galindo, I. & Alonso, C. African Swine Fever Virus: A Review. *Viruses* **9**, doi:10.3390/v9050103 (2017).

3 Dixon, L. K., Chapman, D. A. G., Netherton, C. L. & Upton, C. African swine fever virus replication and genomics. *Virus research* **173**, 3-14, doi:10.1016/j.virusres.2012.10.020 (2013).

4 Sanchez-Cordon, P. J., Montoya, M., Reis, A. L. & Dixon, L. K. African swine fever: A re-emerging viral disease threatening the global pig industry. *Vet J* **233**, 41-48, doi:10.1016/j.tvjl.2017.12.025 (2018).

5 Costard, S., Mur, L., Lubroth, J., Sanchez-Vizcaino, J. M. & Pfeiffer, D. U. Epidemiology of African swine fever virus. *Virus research* **173**, 191-197, doi:10.1016/j.virusres.2012.10.030 (2013).

6 Arzt, J., White, W. R., Thomsen, B. V. & Brown, C. C. Agricultural Diseases on the Move Early in the Third Millennium. *Veterinary pathology* **47**, 15-27, doi:10.1177/0300985809354350 (2010).

7 Department, W. A. H. I. a. A. *African Swine Fever (ASF) Report N°4: October 5 - 18, 2018*, <http://www.oie.int/en/animal-health-in-the-world/information-on-aquatic-and-terrestrial-animal-diseases/african-swine-fever/reports-on-asf/> (2018).

8 Ge, S. Q. *et al.* Molecular Characterization of African Swine Fever Virus, China, 2018. *Emerg Infect Dis* **24**, 2131-2133, doi:10.3201/eid2411.181274 (2018).

9 SS, S. & WR, H. Antibody response to inactivated preparations of African swine fever virus in pigs. *American journal of veterinary research* **28**, 6 (1967).

10 King, K. *et al.* Protection of European domestic pigs from virulent African isolates of African swine fever virus by experimental immunisation. *Vaccine* **29**, 4593-4600, doi:10.1016/j.vaccine.2011.04.052 (2011).

11 Reis, A. L. *et al.* Deletion of the African Swine Fever Virus Gene DP148R Does Not Reduce Virus Replication in Culture but Reduces Virus Virulence in Pigs and Induces High Levels of Protection against Challenge. *Journal of virology* **91**, doi:UNSP e01428-17 10.1128/JVI.01428-17 (2017).

12 Escribano, J. M., Galindo, I. & Alonso, C. Antibody-mediated neutralization of African swine fever virus: Myths and facts. *Virus research* **173**, 101-109, doi:10.1016/j.virusres.2012.10.012

bioRxiv preprint doi: https://doi.org/10.1101/460832; this version posted November 4, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(2013).

13 P, G.-P. *et al.* Neutralizing antibodies to different proteins of African swine fever virus inhibit both virus attachment and internalization. *Journal of virology* **70**, 6 (1996).

14 Kessler, C. *et al.* The intracellular proteome of African swine fever virus. *Scientific reports* **8**, doi:Artn 14714 10.1038/S41598-018-32985-Z (2018).

15 A, A., T, M., M, G. & G, A. A proteomic atlas of the African swine fever virus particle. *Journal of virology*, doi:10.1128/JVI.01293-18 (2018).

16 Chapman, D. A., Tcherepanov, V., Upton, C. & Dixon, L. K. Comparison of the genome sequences of non-pathogenic and pathogenic African swine fever virus isolates. *The Journal of general virology* **89**, 397-408, doi:10.1099/vir.0.83343-0 (2008).

17 de Villiers, E. P. *et al.* Phylogenomic analysis of 11 complete African swine fever virus genome sequences. *Virology* **400**, 128-136, doi:10.1016/j.virol.2010.01.019 (2010).

18 Fraczyk, M. *et al.* Evolution of African swine fever virus genes related to evasion of host immune response. *Veterinary microbiology* **193**, 133-144, doi:10.1016/j.vetmic.2016.08.018 (2016).

19 Michaud, V., Randriamparany, T. & Albina, E. Comprehensive phylogenetic reconstructions of African swine fever virus: proposal for a new classification and molecular dating of the virus. *PloS one* **8**, e69662, doi:10.1371/journal.pone.0069662 (2013).

20 Wang, Y. *et al.* Origin and Possible Genetic Recombination of the Middle East Respiratory Syndrome Coronavirus from the First Imported Case in China: Phylogenetics and Coalescence Analysis. *mBio* **6**, e01280-01215, doi:10.1128/mBio.01280-15 (2015).

21 Nagy, P. D. & Bujarski, J. J. Homologous RNA recombination in brome mosaic virus: AU-rich sequences decrease the accuracy of crossovers. *Journal of virology* **70**, 415-426 (1996).

22 Roossinck, M. J. Mechanisms of plant virus evolution. *Annual review of phytopathology* **35**, 191-209, doi:10.1146/annurev.phyto.35.1.191 (1997).

23 Wikipedia. *Homologous recombination*, <https://en.wikipedia.org/wiki/Homologous_recombination#In_viruses> (2018).

24 Agarwala, R. *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **44**, D7-D19, doi:10.1093/nar/gkv1290 (2016).

25 Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution* **30**, 772-780, doi:10.1093/molbev/mst010 (2013).

26 Larkin, M. A. *et al.* Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2947-2948, doi:10.1093/bioinformatics/btm404 (2007).

27 Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *Journal of molecular biology* **215**, 403-410 (1990).

28 Emms, D. M. & Kelly, S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome biology* **16**, 157, doi:10.1186/s13059-015-0721-2 (2015).

29 Quevillon, E. *et al.* InterProScan: protein domains identifier. *Nucleic Acids Res* **33**, W116-120, doi:10.1093/nar/gki442 (2005).

30 Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E. L. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *Journal of*

bioRxiv preprint doi: https://doi.org/10.1101/460832; this version posted November 4, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

molecular biology **305**, 567-580, doi:10.1006/jmbi.2000.4315 (2001).

31 Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389-3402 (1997).

32 Rock, D. L. Challenges for African swine fever vaccine development-"... perhaps the end of the beginning.". *Veterinary microbiology* **206**, 52-58, doi:10.1016/j.vetmic.2016.10.003 (2017).

33 Reis, A. L., Netherton, C. & Dixon, L. K. Unraveling the Armor of a Killer: Evasion of Host Defenses by African Swine Fever Virus. *Journal of virology* **91**, doi:10.1128/jvi.02338-16 (2017).

34 Martin, D. & Rybicki, E. RDP: detection of recombination amongst aligned sequences. *Bioinformatics* **16**, 562-563 (2000).

35 Tamura, K., Stecher, G., Peterson, D., Filipski, A. & Kumar, S. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Molecular biology and evolution* **30**, 2725-2729, doi:10.1093/molbev/mst197 (2013).

36 Huson, D. H. *et al.* Dendroscope: An interactive viewer for large phylogenetic trees. *Bmc Bioinformatics* **8**, 460, doi:10.1186/1471-2105-8-460 (2007).

37 Team, R. C. *R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria*, <https://www.R-project.org/> (2018).

38 TG, B., Z, L., JG, N., DL, R. & L, Z. African swine fever virus multigene family 360 genes affect virus replication and generalization of infection in Ornithodoros porcinus ticks. *Journal of virology* **78**, 9 (2004).