ANASFV: a workflow for ASFV whole genome sequencing, assembly, and evaluation

Ke Li [a, †] , Xu Han [b, †] , Yanwen Shao [a] , Xiaomin Zhao [c] , Eager Wayne Johnson [d] , Runsheng Li [a, e, f, *]

[a] Department of Infectious Diseases and Public Health, Jockey Club College of Veterinary Medicine and Life Sciences, City University of Hong Kong, Hong Kong, China.

[b] College of Veterinary Medicine, China Agricultural University, Beijing, China.

[c] College of Veterinary Medicine, Northwest A&F University, Yangling, Shaanxi, China.

[d] Enable Agricultural Technology Consulting, Ltd, Beijing, China.

[e] Department of Precision Diagnostic and Therapeutic Technology, City University of Hong Kong Shenzhen Futian Research Institute, Shenzhen, China

[f] Tung Biomedical Sciences Centre, City University of Hong Kong, Hong Kong, China.

[†] Authors contributed equally.

[*] To whom correspondence should be addressed. Tel: +852 3442-7207; Fax: +852 3442-0589; Email: runsheng.li@cityu.edu.hk

Abstract:

African Swine Fever Virus (ASFV) poses an ongoing threat with widespread outbreaks affecting both domestic and wild pig populations globally. Effective management of ASFV outbreaks necessitates a deep understanding of its genetic diversity and evolutionary dynamics. Despite the advantages of nanopore sequencing for genome analysis, its application to ASFV genomes encounters specific challenges, such as high host DNA contamination that reduces viral genome coverage and an inherently higher error rate producing small insertions and deletions (indels).

To overcome these obstacles, we developed the ANASFV (analysis of an ASFV) pipeline, which enhances both laboratory and bioinformatics workflows for ASFV genome sequencing. This study introduces an amplicon sequencing approach that significantly improves genomic coverage, enabling reliable genome assembly. We also implemented reference-aided polishing techniques to correct small indels caused by nanopore sequencing errors. Additionally, we established a genome quality assessment system, akin to the Benchmarking Universal Single-Copy Orthologue (BUSCO), to evaluate the completeness and accuracy of the assembled genomes. A comprehensive phylogenetic analysis based on coding sequences (CDS) was conducted, allowing us to generate a refined phylogenetic tree that includes all known ASFV genomes. The ANASFV pipeline will facilitate ASFV full genome sequencing using the nanopore platform. It will also support robust downstream bioinformatics analyses to assess genome integrity and conduct detailed phylogenetic studies based on whole-genome data.

Introduction

The African Swine Fever Virus (ASFV) is the causative agent of African Swine Fever (ASF), a severe and often fatal viral disease that affects domestic and wild pigs(1). Since the initial identification of ASF in Kenya in 1921, the disease has emerged in countries throughout sub-Saharan Africa. The first outbreak of ASF outside of Africa occurred in the late 1950s, affecting both Europe and South America. In 2007, a new wave of ASF emerged in Georgia and rapidly expanded to other European countries, resulting in significant economic consequences for the global domestic pig industry(2). In August 2018, ASFV was identified in the northeastern region of China(3). The ASF outbreak in China had a significant impact due to the country's high share in the global pork market, as it is the largest pork producer. By 2019, ASFV had spread rapidly throughout every region of China, causing widespread outbreaks and considerable economic losses in the domestic pig industry(4). The virus also extended its reach to other parts of East and Southeast Asia(5, 6). This demonstrates the ongoing and far-reaching impact of ASFV, highlighting the urgent need for a global understanding of this virus.

Whole genome sequencing is crucial in the context of the current ASFV outbreak. The whole genome assembly provides a complete map of the ASFV genome, allowing researchers to gain a comprehensive understanding of its genetic makeup. This information is vital for studying the structure, function, and evolutionary history of the virus. In practical applications, studying the complete genome of ASFV enables researchers to develop effective vaccines against ASFV and trace the source and transmission patterns of ASFV. The utilization of Oxford Nanopore Technologies (ONT) for ASFV whole-genome sequencing is of vital importance because of its remarkable turnover speed. As of December 2023, 12 of the 312 ASFV isolates deposited in Genbank were assembled with the help of ONT sequencing. ONT-based sequencing methods offer rapid turnaround times compared to traditional sequencing techniques(7). The streamlined workflow, simplified library preparation, and real-time data analysis allow for faster generation of whole genome sequences. This speed is especially important during disease outbreaks or urgent research situations, where timely information is crucial for effective response and control measures.

Currently, there are limitations in the ONT sequencing of ASFV. First, the presence of substantial host DNA contamination poses a challenge in ONT sequencing, resulting in inadequate coverage of ASFV. Insufficient coverage leads to a fragmented and incomplete genome assembly. Second, ONT sequencing, especially with the R9.4.1 flow cell, exhibits a higher error rate, leading to the generation of numerous small indels(8). The indels generated are often found within consecutive identical bases (homopolymers) because of the challenge of accurate length identification of homopolymers by ONT. The presence of these indels can lead to fragmented genes, which can impact subsequent analyzes. In order to solve the error rate problem, researchers have used NGS-assistant polish(9), but this requires additional NGS reads.

Using whole CDS to build a phylogenetic tree is an ideal method to study the evolutionary relationships of ASFV isolates. ASFV has been classified into 24 different genotypes (I to XIV) based on sequence variations observed in p72 (B646L gene)(10). For the currently prevalent genotype II isolate in Eurasia, p72 exhibits high similarity, making it insufficient to differentiate between different genotype II isolates solely based on the p72 gene tree(11, 12). The use of a full

genome to build a phylogenetic tree also presents challenges due to the presence of numerous repetitive elements in the ASFV genome and difficulties in genome assembly for some isolates(13-15). Using only coding regions to build trees can avoid problems such as alignment problems in repeated regions and missing parts of the genome. Whole-proteome trees are more suitable for cross-species comparisons and are obviously not suitable when studying ASFV isolates. Therefore, the whole-CDS tree is a relatively ideal solution. Using whole CDS to build a phylogenetic tree can finely distinguish similar isolates and does not require highly intact genome assembly. As a result, the utilization of whole-CDS facilitates the incorporation of a greater number of isolates into the tree.

In early 2021, an outbreak of ASF occurred on a farm in Shandong, China, leading to significant livestock mortality. To investigate the outbreak, we developed a pipeline called ANASFV (analysis of an ASFV) that incorporates sequencing, assembly, polishing, completeness assessment, and phylogenetic analysis of ASFV genomes. We incorporate PCR amplification with custom-designed primers to ensure adequate coverage. Moreover, we employed reference-guided polishing techniques to effectively rectify small indels introduced by nanopore reads, improving the accuracy of our analysis. These strategies overcome the limitations posed by the low amount of template DNA and the potential sequencing errors associated with ONT sequencing. And we developed a system for completeness evaluation of the ASFV genome. We construct a phylogenetic tree of 297 ASFV isolates using 188 CDS. We tested this pipeline using the isolate ShanDong2021. As a result, we obtained a high-quality ShanDong2021 genome and a refined ASFV phylogenetic tree.

Methods

Sample Collection and Sequencing

The sample was collected from the lung tissue of sick pigs. Genomic DNA was extracted using the QIAamp MinElute Virus Spin Kit (Qiagen, Germany). Two sets of primers were designed with amplification product lengths of 20 Kb and 5 Kb (Table S1, Table S2). The samples were subjected to PCR amplification using the two primer sets, respectively. Subsequently, the amplified DNA products were prepared for sequencing following the nanopore sequencing library preparation protocol (LSK-110, ONT). The library was sequenced in R9.4.1 flow cells (FLO-MIN106D, ONT).

Basecalling and Trimming

The raw nanopore sequencing data was basecalled by Guppy (v5.0.11) with the Super-accurate basecalling model. Subsequently, the reads were filtered and trimmed using NanoFilt(16) (v2.8.0), with a minimum average read quality score of 10 and a minimum read length of 1000 base pairs. Additionally, the first 50 bases were trimmed using the headcrop option.

Genome Assembly

The genome exhibiting the highest read mapping coverage among all available ASFV genomes from NCBI is regarded as a reference genome (ANASFV: find_near_ref.py). This reference genome was then utilized for mapping assembly of the sequenced data. The alignment was performed using the minimap2(17) (v2.17) aligner with the -a option, and high quality trimmed reads as input. The alignment results were saved in SAM format. From the SAM file, a consensus sequence was generated using samtools(18) (v1.17). First, the SAM file was converted to the BAM format,

excluding unmapped reads (-F 4). The resulting BAM file was then sorted using the samtools sort command. Finally, the consensus sequence was extracted from the sorted BAM file using the consensus command of samtools.

Polishing the Consensus Sequence

The consensus sequence was further polished using two rounds of error correction. First, medaka (v1.11.3) (https://github.com/nanoporetech/medaka) was used. The high-quality trimmed reads and the initial consensus sequence were provided as input to the medaka_consensus program. Next, homopolish(19) (v0.4.1) was utilized to perform additional error correction. We performed nBLAST on the consensus sequence against the NCBI database, and the closest isolate we obtained was ASFV/pig/China/CAS19-01/2019 (GenBank: MN172368.1). We use this genome as a reference for homopolish. The homopolish algorithm used the R9.4.pkl model for correction.

Phylogenetic analysis

The ASFV genomes used to build a phylogenetic tree were downloaded from NCBI, covering data up until December 2023 (Table S3). 188 CDS from NC_044959.2 (ASFV Georgia 2007/1) were used as a reference and employed Exonerate(20) (v2.4.0) for pairwise sequence comparison. This facilitated the acquisition of the corresponding CDS for each ASFV isolate. These obtained results were then utilized as input for the tree construction of "denovo" mode using uDance(21) (v1.6.3). Subsequently, the generated tree served as a backbone for a second iteration of tree construction by adjusting the backbone option to 'tree' in uDance, allowing the replacement of unplaceable sequences. In the end, 14 isolates were excluded because they were CDS duplicate sequences, and an additional 2 isolates were excluded because their optimal placement could not be reliably determined. A total of 296 isolates were used to construct a phylogenetic tree with ShanDong2021. Tree visualization was performed using iTOL(22) (v6.9).

Results

To overcome the difficulties of nanopore sequencing of ASFV and conduct a more refined phylogenetic analysis of ASFV, we have developed a comprehensive pipeline called 'ANASFV' that includes sequencing, assembly, polishing, quality evaluation, and phylogenetic analysis of the ASFV genome (Figure 1). We used the DNA sample from the ShanDong2021 isolate to test our pipeline. After performing multiple PCRs, we obtained the amplicon DNA from the ASFV DNA template. Subsequently, we conducted nanopore sequencing using the amplified DNA. These long reads were then subjected to mapping assembly to obtain a draft genome. To improve accuracy, we performed two rounds of polishing using medaka and homopolish. Finally, we obtained the polished genome, which was subjected to quality assessment, resulting in a BUSCO(23) (Benchmarking Universal Single-Copy Orthologs) -like evaluation. Additionally, we obtained sequence files for each CDS across this genome and genomes of other isolates by employing exonerate. Finally, we utilize uDance to construct a phylogenetic tree based on the obtained CDS data. For details on the code, see GitHub: https://github.com/lrslab/anasfv . Researchers conducting whole genome analysis of ASFV using nanopore sequencing can refer to this workflow. For researchers utilizing other sequencing platforms, the downstream analysis of this project, such as completeness evaluation and phylogenetic analysis, can be utilized.

An amplicon system to amplify ASFV whole genome

Nanopore sequencing technology has good adaptability to processing long DNA fragments. We designed a set of primers with an amplification product of 20 Kb (Table S1), which is expected to make better use of the long-read characteristics of Nanopore sequencing technology and obtain more complete long-fragment sequencing data, but its amplification effect may not be good. Therefore, we also designed a set of primers with an amplification product of 5 Kb (Table S2). The 5 Kb set consists of 40 pairs of primers, with an overlap length between amplified products of approximately 100-500. The 20 Kb set consists of 10 pairs of primers, with an overlap length between amplified products of approximately 500-900. In this ShanDong2021 sample, the 5kb primer set designed for 5 Kb amplification product lengths demonstrated superior performance compared to the 20 Kb primer set in terms of amplification efficiency (Figure 2). There were 4 pairs of primers in the 20 Kb primer set that exhibited a suboptimal amplification efficiency. These exceptions may have been influenced by factors such as primer design, template quality, or other experimental variables. The yield of reads for 20 Kb and 5 Kb was 748 Mb and 1.6 Gb, respectively (Table S4).

An evaluation system to evaluate the quality of the ASFV genome assembly

BUSCO is a computational tool used to assess the completeness and quality of genome assemblies(23). ASFV is not included in the BUSCO's database (OrthoDB). Therefore, we developed a completeness evaluation system to generate a BUSCO-like notation. We generated consensus sequences for 148 CDS from genotype II isolates, with 33 MGF genes (MGF_100 : 2, MGF_110 : 3, MGF_300: 2, MGF_360 : 16, MGF_505 : 10), and consensus sequences for 149 CDS from genotype I isolates, with 34 MGF genes (MGF_100 : 2, MGF_110 : 4, MGF_300 : 2, MGF_360 : 16, MGF_505 : 10). The list of CDS used can be found in Table S5. CDS prediction of ShanDong2021 was performed using Prodigal(24) (v2.6.3), and the predicted CDS were compared to consensus sequences using BLASTN (e-value 1e-5). This comparison yielded a BUSCO(25) like genome completeness evaluation, with C:complete [D:duplicated], F:fragmented, M:missing, n:number of genes used. If a consensus CDS term can find a mapping from the predicted gene sequences with an identity larger than 90%, together with a unique mapping length longer than 90%, it can be considered "complete." A consensus CDS term that cannot find a valid mapping (with an identity greater than 30% and a mapping length greater than 30%) in the predicted genes is considered "missing". The other consensus CDS with partial hit is termed "fragmented". The "duplicated" term means that there is more than one 'complete' hit in the predicted gene sequences.

Reference-aided polish helps to resolve the small indels generated by nanopore

The reads provided sufficient coverage for accurate genome assembly, ensuring that no regions of the reference genomes were left unaddressed. We performed BLASTN on the draft assembly sequence against the NCBI database, and the closest isolate we obtained was ASFV/pig/China/CAS19-01/2019 (MN172368.1). We use this genome as a reference for downstream analysis. An evaluation of quality and accuracy was performed using QUAST (v5.2.0, default parameters) to compare our draft assembly with the reference genome. The evaluation revealed a mismatch rate of 7.98 per 100 Kb and an indel rate of 121.75 per 100 Kb. After performing reference-aided polishing using homopolish, the indel rate dropped to 1.6 per 100 Kb. Upon comparing the draft assembly with the consensus sequence of genotype II, we obtained the

BUSCO-like notation of C: 52.7% [D: 0.0%], F: 45.95%, M: 1.35%, n: 148. After excluding the MGF gene, the notation was C: 54.78% [D: 0.0%], F: 43.48%, M: 1.74%, n: 115. Following the application of reference-aided polish using homopolish, we obtained a BUSCO-like notation of C: 99.32% [D: 0.0%], F: 0.68%, M: 0.0%, n: 148. And without MGF genes: C: 99.13% [D: 0.0%], F: 0.87%, M: 0.0%, n: 115.

Completeness evaluation for all ASFV genomes in Genbank

We performed the BUSCO-like evaluation of all available isolates in NCBI as of December 2023 (Table S3). Among the poor-quality genomes, three isolates sequenced by nanopore were clearly due to insufficient polish: CN/GD/2022 (Accession ID: OR290104.2), Philippines A4 2021 (Accession ID: ON963982.1) and ASFV/Kyiv/2016/131 (Accession ID: MN194591.1), as shown in Table 1. These genomes demonstrate a high gene count, significant gene fragmentation, and a long total genome length. We applied reference-aided polish to them using the same method mentioned above. The number of genes and BUSCO fragmentation have been reduced after polishing (Figure 3A).

Correction of homopolymers in the ShanDong2021 genome

The ASFV genome contains an excessive number of homopolymers. We compared the number of homopolymers in the ASFV genome and human genome, together with the simulated random genome sequences with the same GC content of the ASFV (38.38%) or human genome (41%). The results show that the frequency of homopolymers with lengths greater than 3 in the ASFV genome is higher than in the human genome, and both are higher than in the simulated genomes (Figure 3B). Although the total count of homopolymers in the ASFV genome is higher than in the human genome, homopolymers composed of "A" and "T" bases occur more frequently in the human genome compared to those of the ASFV. On the other hand, homopolymers composed of "G" and "C" bases occur with a higher frequency in the ASFV genome (Figure S1). There are 4866 homopolymers with lengths greater than 3 in the ShanDong2021 genome. After homopolish was applied, 229 homopolymers underwent changes. Among these changes, 22 were lengthened, while the majority, specifically 207, were shortened (Figure 3C). In addition to the 229 homopolymer changes after homopolish, there was also one instance where "AA" was changed to "A" after homopolish. Among all corrections made during homopolish, the majority involved single-base changes, with 200 being single-base deletions and 18 being single-base insertions (Table S6). 208 indels (90.4 %) generated by nanopore sequencing are recognized as longer than they actually are, so they are shortened after homopolish. Most of these errors (198 out of 208) involve homopolymers with lengths ranging from 3 to 6. The other 22 (9.6%) indels generated by nanopore sequencing can be recognized as shorter than they actually are, and all of the 22 involve homopolymers with lengths ranging from 6 to 9 (Figure 3D).

Description of Shandong2021

At this stage, we finally obtained an 189 Kb ASFV genome of ShanDong2021 (Accession ID: OR660089) with 72 genes in the forward strand and 97 genes in the reverse strand (Figure 4A). In the comparative analysis of the whole genome, ShanDong2021 exhibited an insertion of approximately 250 bp at the end of MGF_360-21R (Figure 4B), causing the gene length to increase from 1071 bp to 1092 bp. It can be seen from the dotplot that the repeat sequences (mainly MGF

genes) are located at the beginning and the end of the genome, with more at the beginning than at the end.

Phylogenetic analysis

The whole-CDS tree used a combination of 188 CDS to build the tree, which can finely distinguish the similar isolates. In this way, we can include more genomes on the tree (Figure 5). Six distinct clades can be identified. The ShanDong2021 isolate was clustered with LYG18 (OM105586) and ASFV-wbBS01 (MK645909), both of which were sampled from China. ShanDong2021 and ASFV-wbBS01 formed a small clade, indicating the closest evolutionary distance between the two. We identified three genes that showed significant divergence between the two genomes: I196L (Ka/Ks: 0.24238), M1249L (Ka/Ks: 0.335183), MGF_360-21R (Ka/Ks: 0.436954).

Discussion

The read length of NGS is usually less than 500bp, which brings difficulties to genome assembly, especially repetitive regions. There are such repetitive regions at both ends of the ASFV genome, and genome assembly based on NGS reads cannot solve this problem well. The reads of ONT can reach ultra-long (longest >4 Mb), which can directly cover most repetitive regions, making the assembled genome more consistent with the real situation. The feasibility and effectiveness of amplicon sequencing for the ASFV genome, using primers targeting 10 Kb amplicons, have been supported by a previous study(26). In our study, the primer set of 5 Kb amplicon outperformed the primer set of 20 Kb amplicon. However, this preference is not absolute, as the amplification efficiency depends on the quality of the DNA. For specific needs, researchers can utilize primerdiffer(27) to design their own primer set accordingly.

The development of the completeness evaluation system for the ASFV genome has provided a valuable tool to assess the quality of genome assemblies. When comparing genes in the genome to known consensus gene sets, it is possible to assess whether the genes are complete and whether there are missing, duplicated, or fragmented genes. The occurrence of these missing, duplication, or fragmentation issues generally implies poor genome assembly. Since there are relatively few available genomes for other ASFV genotypes, we only established consensus gene sets for genotype I and genotype II. Therefore, this evaluation system is more effective for the currently prevalent genotype I and genotype II. This system fills a gap in the tools available, allowing researchers to conveniently assess the quality of the ASFV genome assembly. The user-friendly nature of this completeness evaluation system allows researchers to focus on their core work of understanding and analyzing the ASFV genome, without being burdened by intricate and trivial details.

In the genome polishing process, we employed reference-aided polish, which has shown remarkable effectiveness in reducing errors introduced during ONT sequencing. The results obtained from applying the completeness evaluation system to the nanopore-based ASFV genomes demonstrated the effectiveness of reference-aided polish in improving ONT sequencing errors in the absence of NGS reads as a reference for polishing. For researchers who only use ONT sequencing to obtain the ASFV genome, reference-aided polish is a good method to solve indel errors in ONT reads.

In our study, uDance was used to construct the whole-CDS tree(21). The uDance implements a divide-and-conquer strategy that simplifies the phylogenetic tree construction process. By utilizing parallel processing, uDance significantly reduces the time needed for tree construction. The refined phylogenetic tree provided valuable information on the genetic diversity and relatedness of different ASFV isolates, contributing to a better understanding of the epidemiology and evolution of ASFV. The ANASFV pipeline automates the process of downloading all available ASFV genomes from the NCBI database, performing CDS prediction and alignment. In case starting from scratch is not desired, we also provide our tree file (newick format), which can serve as a backbone tree to add new isolates by adjusting the backbone option to "tree" in uDance.

In general, our study demonstrates the successful application of ONT sequencing coupled with custom-designed PCR primers for complete genome sequencing of the ShanDong2021 ASFV isolate. The use of homopolish and the construction of a phylogenetic tree further enhanced accuracy and provided valuable insights into the genetic characteristics and evolutionary relationships of ASFV isolates. These findings contribute to the field of ASFV research and have implications for epidemiological studies and control measures.

Data Availability

Nanopore sequencing reads included in this study have been submitted to the Sequence Read Archive (SRA) under accession numbers SRR28789321 and SRR28789322. The whole genome assembly for our isolate has been deposited in GenBank under accession number OR660089.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the Hetao Shenzhen-Hong Kong Science and Technology Innovation Cooperation Zone Shenzhen Park Project (HZQB-KCZYZ-2021017); the Early Career Scheme from the Hong Kong Research Grant Council (project number CityU 21100521); the Hong Kong Health and Medical Research Fund (project number 08194126), and the New Research Initiatives Support from City University of Hong Kong (project number 9610497) to R.L.

We thank Xudong Liu from City University of Hong Kong for the homopolymer occurrence data of human genome.

Supplementary Materials

Supplementary_figure.docx :

Figure S1: Frequency of homopolymer occurrence in ASFV and human genomes.

Supplementary_tables.xlsx :

Table S1: 20kb primer set. Table S2: 5kb primer set. Table S3: Completeness evaluation of all available ASFV genomes. Table S4: Statistics of reads from 20kb and 5kb primer set. Table S5: Consensus gene list. Table S6: Homopolished sites of ShanDong2021.

References

1. Blome S, Franzke K, Beer M. African swine fever–A review of current knowledge. Virus research. 2020;287:198099.
2. Rowlands RJ, Michaud V, Heath L, Hutchings G, Oura C, Vosloo W, et al. African swine fever virus isolate, Georgia, 2007. Emerging infectious diseases. 2008;14(12):1870.
3. Ge S, Li J, Fan X, Liu F, Li L, Wang Q, et al. Molecular characterization of African swine fever virus, China, 2018. Emerging infectious diseases. 2018;24(11):2131.
4. Ito S, Bosch J, Martínez-Avilés M, Sánchez-Vizcaíno JM. The evolution of African swine fever in China: a global threat? Frontiers in veterinary science. 2022;9:828498.
5. Jo YS, Gortázar C. African swine fever in wild boar, South Korea, 2019. Transboundary and Emerging Diseases. 2020;67(5):1776-80.
6. Nga BTT, Tran Anh Dao B, Nguyen Thi L, Osaki M, Kawashima K, Song D, et al. Clinical and pathological study of the first outbreak cases of African swine fever in Vietnam, 2019. Frontiers in Veterinary Science. 2020;7:392.
7. Petersen LM, Martin IW, Moschetti WE, Kershaw CM, Tsongalis GJ. Third-generation sequencing in the clinical laboratory: exploring the advantages and challenges of nanopore sequencing. Journal of clinical microbiology. 2019;58(1):10.1128/jcm. 01315-19.
8. Ni Y, Liu X, Simeneh ZM, Yang M, Li R. Benchmarking of Nanopore R10. 4 and R9. 4.1 flow cells in single-cell whole-genome amplification and whole-genome shotgun sequencing. Computational and Structural Biotechnology Journal. 2023;21:2352-64.
9. Chen Z, Erickson DL, Meng J. Polishing the Oxford Nanopore long-read assemblies of bacterial pathogens with Illumina short reads to improve genomic analyses. Genomics. 2021;113(3):1366-77.
10. Qu H, Ge S, Zhang Y, Wu X, Wang Z. A systematic review of genotypes and serogroups of African swine fever virus. Virus Genes. 2022;58(2):77-87.
11. Ankhanbaatar U, Auer A, Ulziibat G, Settypalli TB, Gombo-Ochir D, Basan G, et al. Comparison of the Whole-Genome Sequence of the African Swine Fever Virus from a Mongolian Wild Boar with Genotype II Viruses from Asia and Europe. Pathogens. 2023;12(9):1143.
12. Wang X, Wang X, Zhang X, He S, Chen Y, Liu X, et al. Genetic characterization and variation of African swine fever virus China/GD/2019 strain in domestic pigs. Pathogens. 2022;11(1):97.
13. Hakizimana JN, Yona C, Makange MR, Kasisi EA, Netherton CL, Nauwynck H, et al. Complete genome analysis of African swine fever virus genotypes II, IX and XV from domestic pigs in Tanzania. Scientific Reports. 2023;13(1):5318.
14. Mazloum A, van Schalkwyk A, Shotin A, Igolkin A, Shevchenko I, Gruzdev KN, et al. Comparative analysis of full genome sequences of African swine fever virus isolates taken from wild boars in Russia in 2019. Pathogens. 2021;10(5):521.
15. Zhao D, Sun E, Huang L, Ding L, Zhu Y, Zhang J, et al. Highly lethal genotype I and II recombinant African swine fever viruses detected in pigs. Nature Communications. 2023;14(1):3096.
16. De Coster W, D'Hert S, Schultz DT, Cruts M, Van Broeckhoven C. NanoPack: visualizing and processing long-read sequencing data. Bioinformatics. 2018;34(15):2666-9.
17. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34(18):3094-100.
18. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078-9.

19. Huang Y-T, Liu P-Y, Shih P-W. Homopolish: a method for the removal of systematic errors in nanopore sequencing by homologous polishing. Genome biology. 2021;22(1):1-17.
20. Slater GSC, Birney E. Automated generation of heuristics for biological sequence comparison. BMC bioinformatics. 2005;6:1-11.
21. Balaban M, Jiang Y, Zhu Q, McDonald D, Knight R, Mirarab S. Generation of accurate, expandable phylogenomic trees with uDance. Nature biotechnology. 2023:1-10.
22. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. Nucleic acids research. 2021;49(W1):W293-W6.
23. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31(19):3210-2.
24. Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC bioinformatics. 2010;11:1-11.
25. Seppey M, Manni M, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness. Gene prediction: methods and protocols. 2019:227-45.
26. Johnston CM, Olesen AS, Lohse L, le Maire Madsen A, Bøtner A, Belsham GJ, et al. A Deep Sequencing Strategy for Investigation of Virus Variants Within African Swine Fever Virus-infected Pigs. Pathogens. 2024;13(2):154.
27. Ren X, Shao Y, Zhang Y, Ni Y, Bi Y, Li R. Primerdiffer: a python command-line module for large-scale primer design in haplotype genotyping. Bioinformatics. 2023;39(4):btad188.

Table 1: The three Nanopore-based Genbank deposited ASFV whole genome sequences before and after polish.

Accession ID	OR290104.2	ON963982.1	MN194591.1
Isolate	CN/GD/2022	Philippines A4 2021	ASFV/Kyiv/2016/131
Genome size (before/after)	189416/**189443**	192377/**192471**	191911/**191943**
Prodigal gene number (before/after)	174/**169**	195/**167**	242/**169**
Completeness(with MGF) (before/after)	C:93.92%[D:0.0%],F:6.08%,M:0.0%,n:148 C:99.32%[D:0.0%],F:0.68%,M:0.0%,n:148	C:82.43%[D:0.0%],F:16.89%,M:0.68%,n:148 C:98.65%[D:0.0%],F:1.35%,M:0.0%,n:148	C:57.43%[D:0.0%],F:39.19%,M:3.38%,n:148 C:99.32%[D:0.0%],F:0.68%,M:0.0%,n:148
Completeness(without MGF) (before/after)	C:93.91%[D:0.0%],F:6.09%,M:0.0%,n:115 C:99.13%[D:0.0%],F:0.87%,M:0.0%,n:115	C:81.74%[D:0.0%],F:17.39%,M:0.87%,n:115 C:99.13%[D:0.0%],F:0.87%,M:0.0%,n:115	C:51.3%[D:0.0%],F:44.35%,M:4.35%,n:115 C:100.0%[D:0.0%],F:0.0%,M:0.0%,n:115

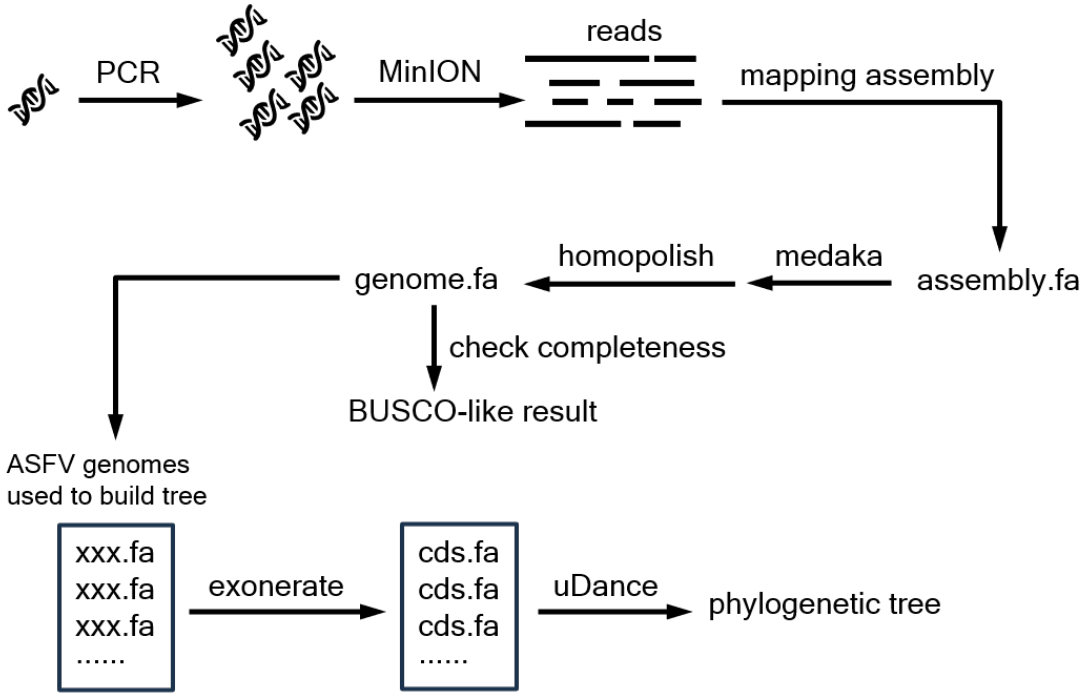


Figure 1. Workflow for ANASFV pipeline. Among all ASFV genomes from NCBI, the genome that ONT reads can map to the most will be used as the reference sequence for mapping assembly using homopolish.

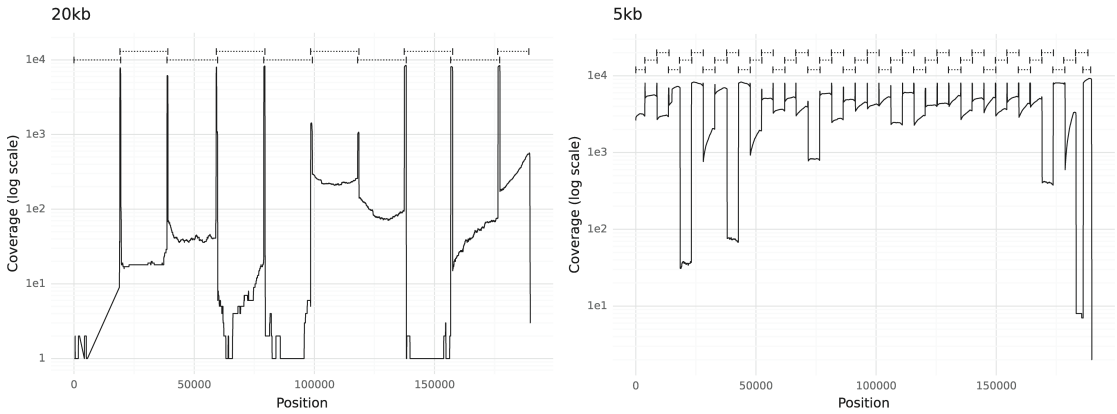


Figure 2. Coverage plot for ASFV amplicons. The logarithmic scale representation of the coverage depth across genomic positions of ShanDong2021, representing the results of amplification products sequencing from primer sets of 5 Kb (right) and 20 Kb (left), respectively. The positions of primer pairs used are indicated at the top of each figure.

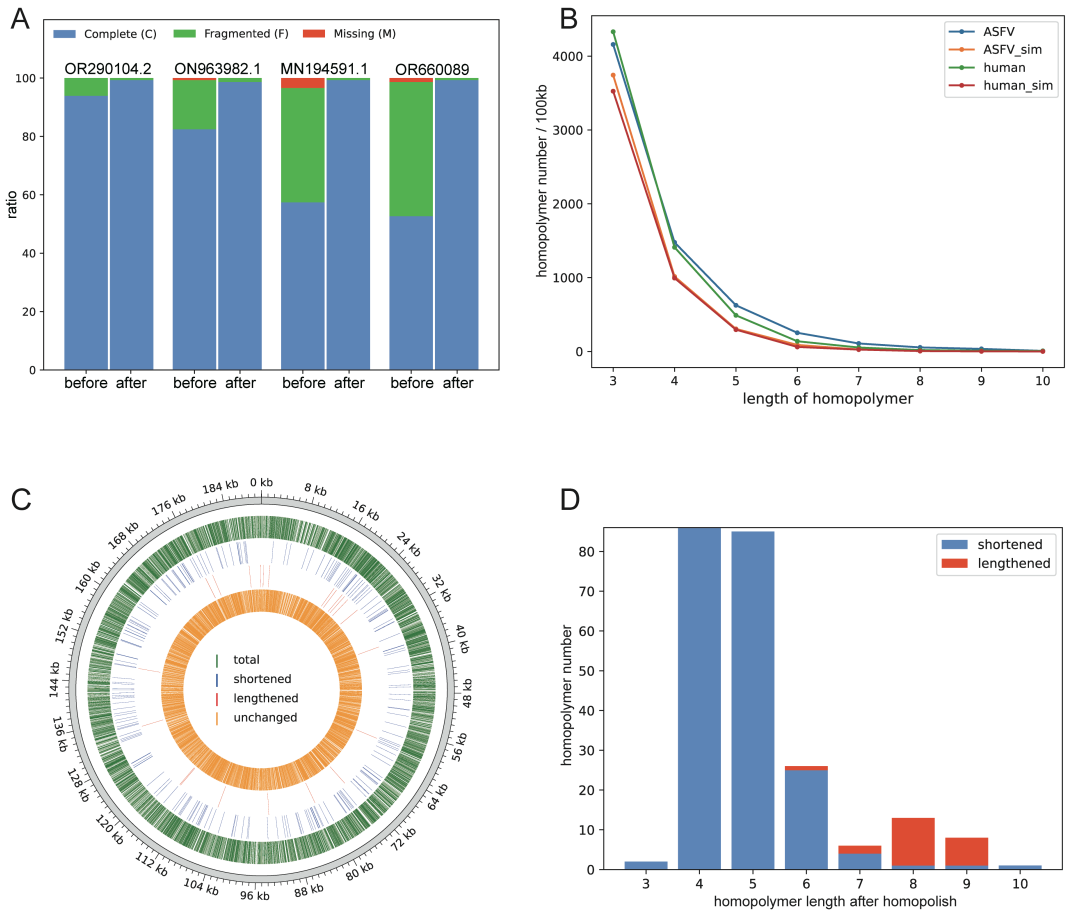


Figure 3. Changes in ASFV genome before and after homopolish. (A) BUSCO-like result of four ASFV genomes that were sequenced by nanopore before and after homopolish: CN/GD/2022 (OR290104.2), Philippines A4 2021 (ON963982.1), ASFV/Kyiv/2016/131 (MN194591.1) and ShanDong2021 (OR660089, from this study). (B) Frequency of homopolymer occurrence in ASFV and human genomes. “sim”, simulated genome using the same GC content. (C) The position for revised homopolymers on the ShanDong2021 genome. Only homopolymers with a length greater than three are shown. (D) The length distributions for revised homopolymers on ShanDong2021. Only homopolymers with lengths 3-10 are shown. Shortened, the homopolymer length becomes shorter after homopolish; Lengthened, the homopolymer length becomes longer after homopolish.

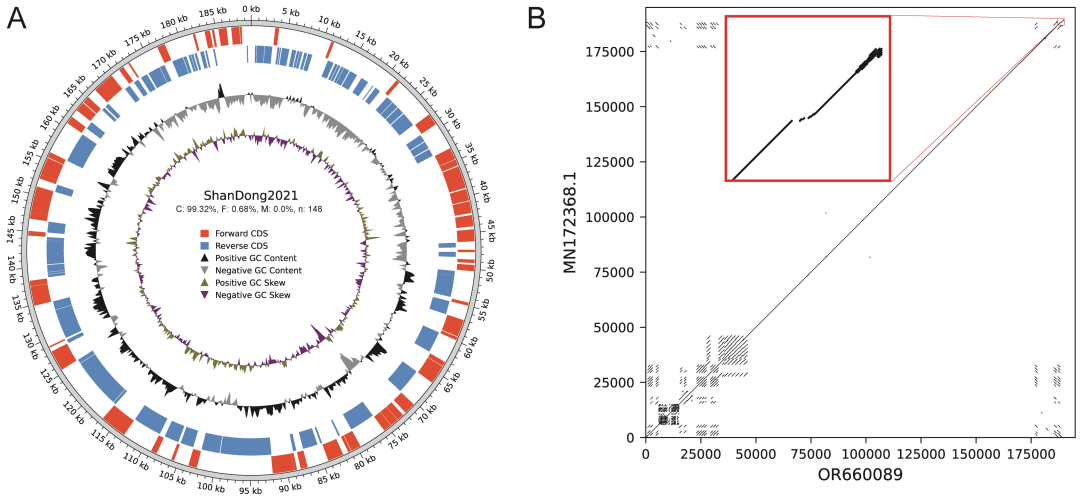


Figure 4. Genomic analysis of ShanDong2021. (A) Circular map of ShanDong2021 genome. (B) Dotplot of ShanDong2021 (OR660089) to ASFV/pig/China/CAS19-01/2019 (MN172368.1). The red box shows a zoomed-in view of the 187 Kb to 190 Kb region.

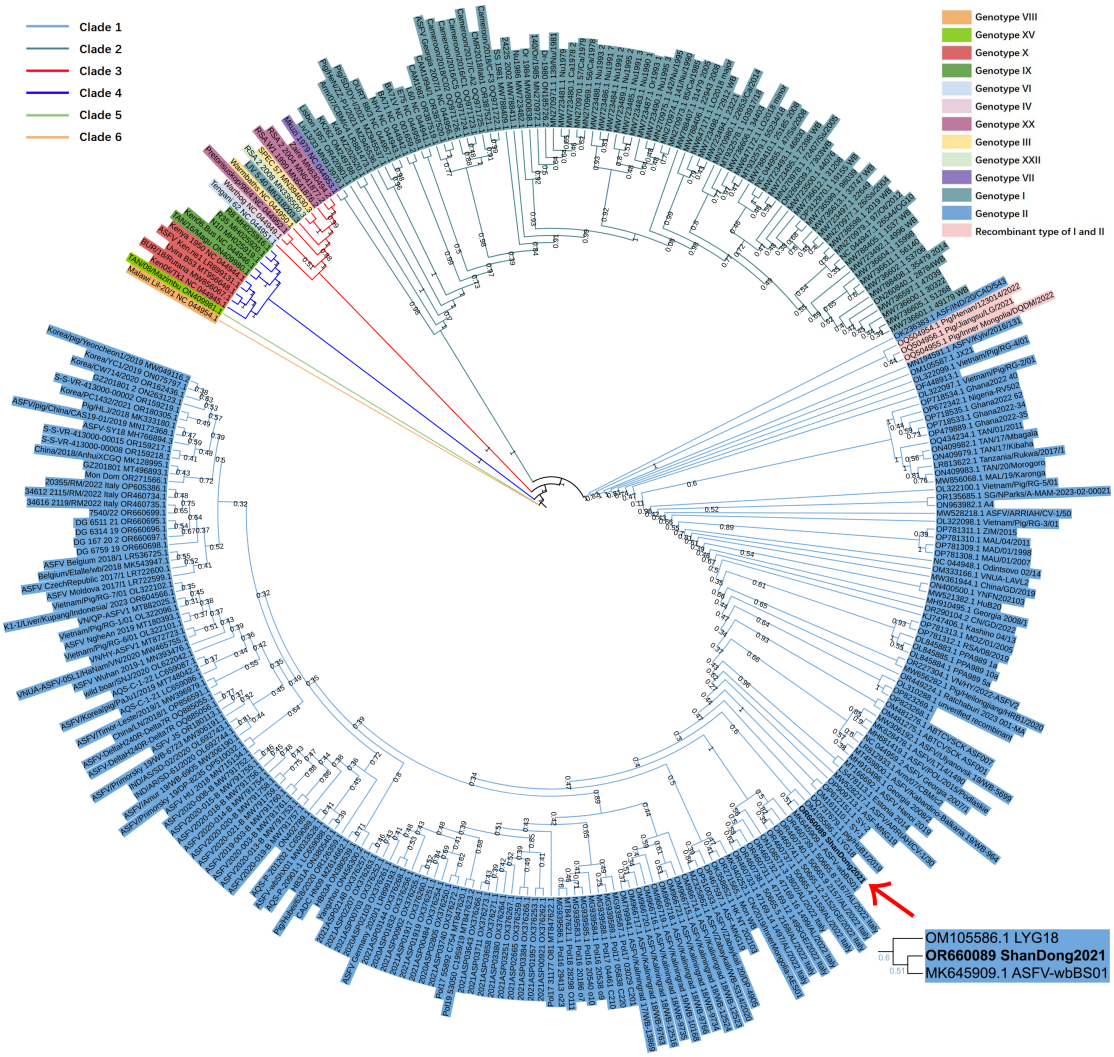
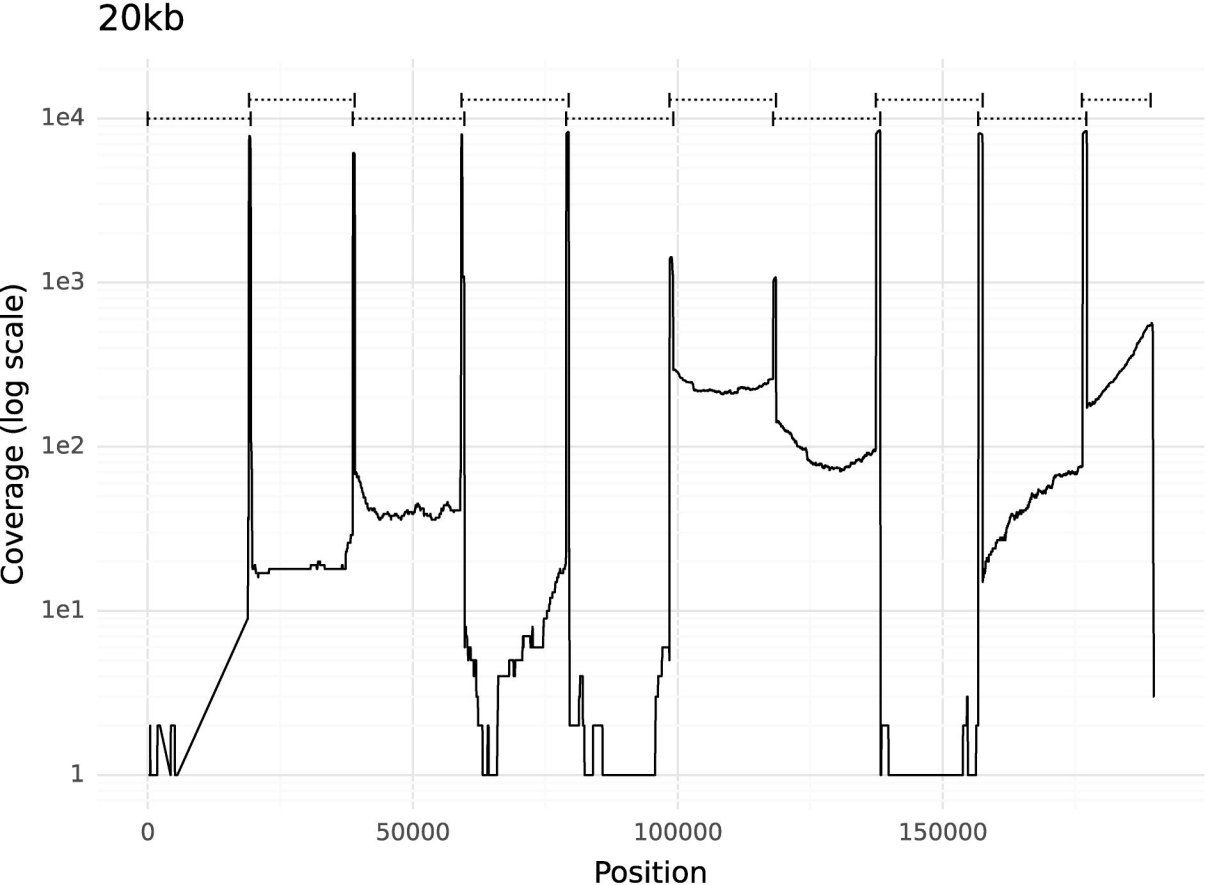
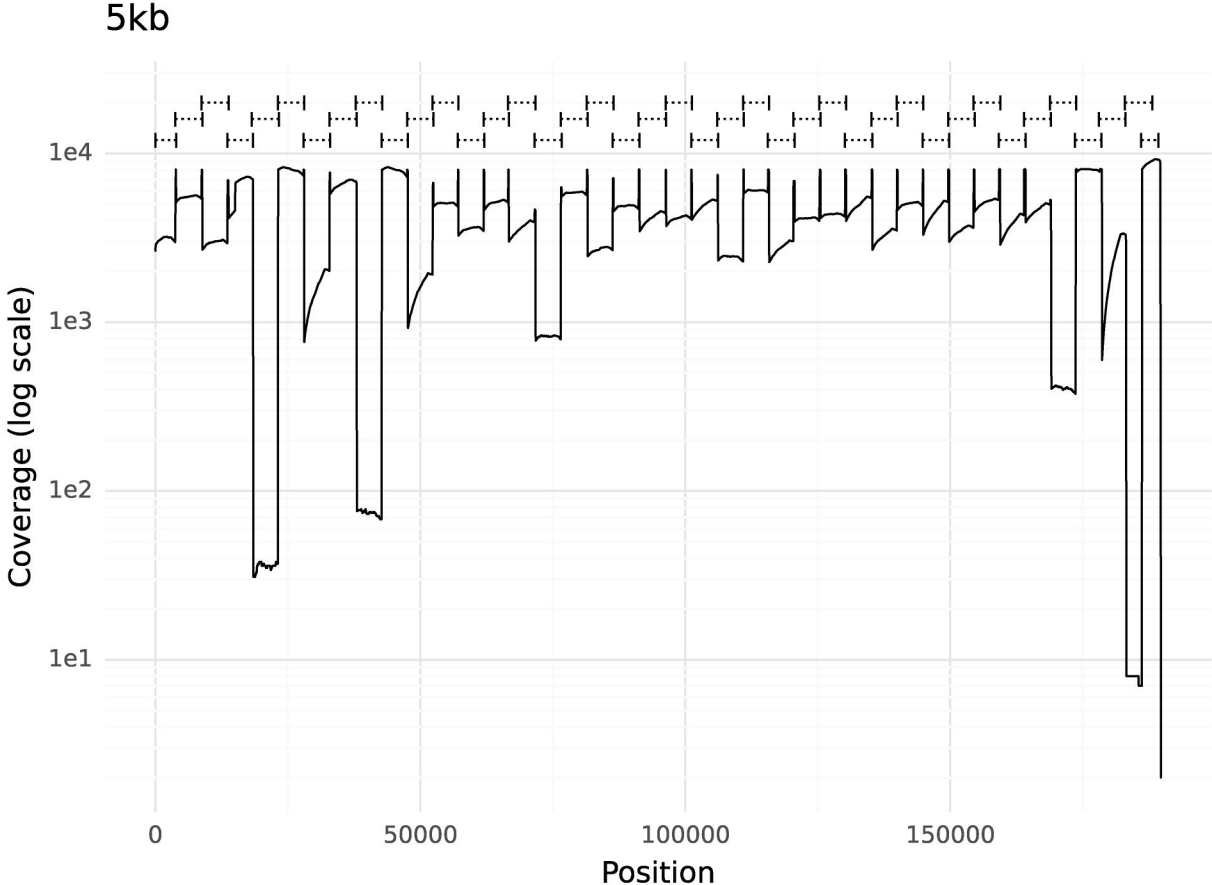
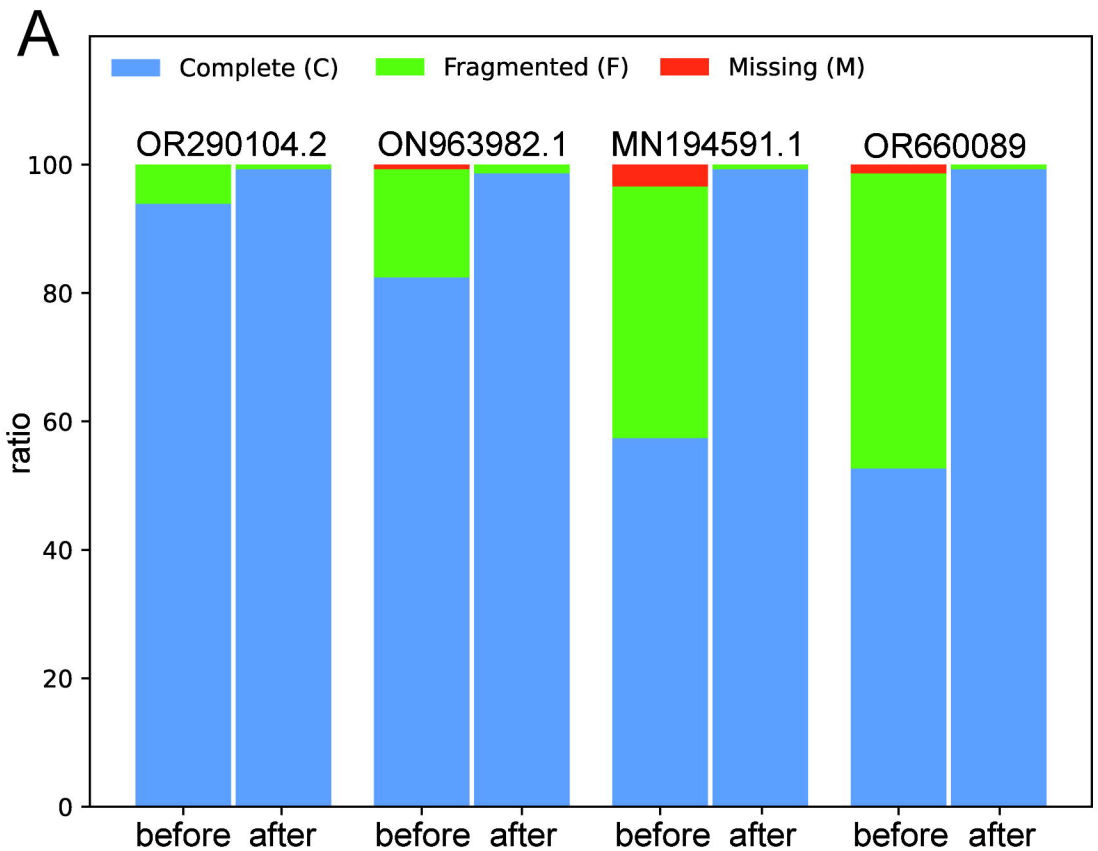
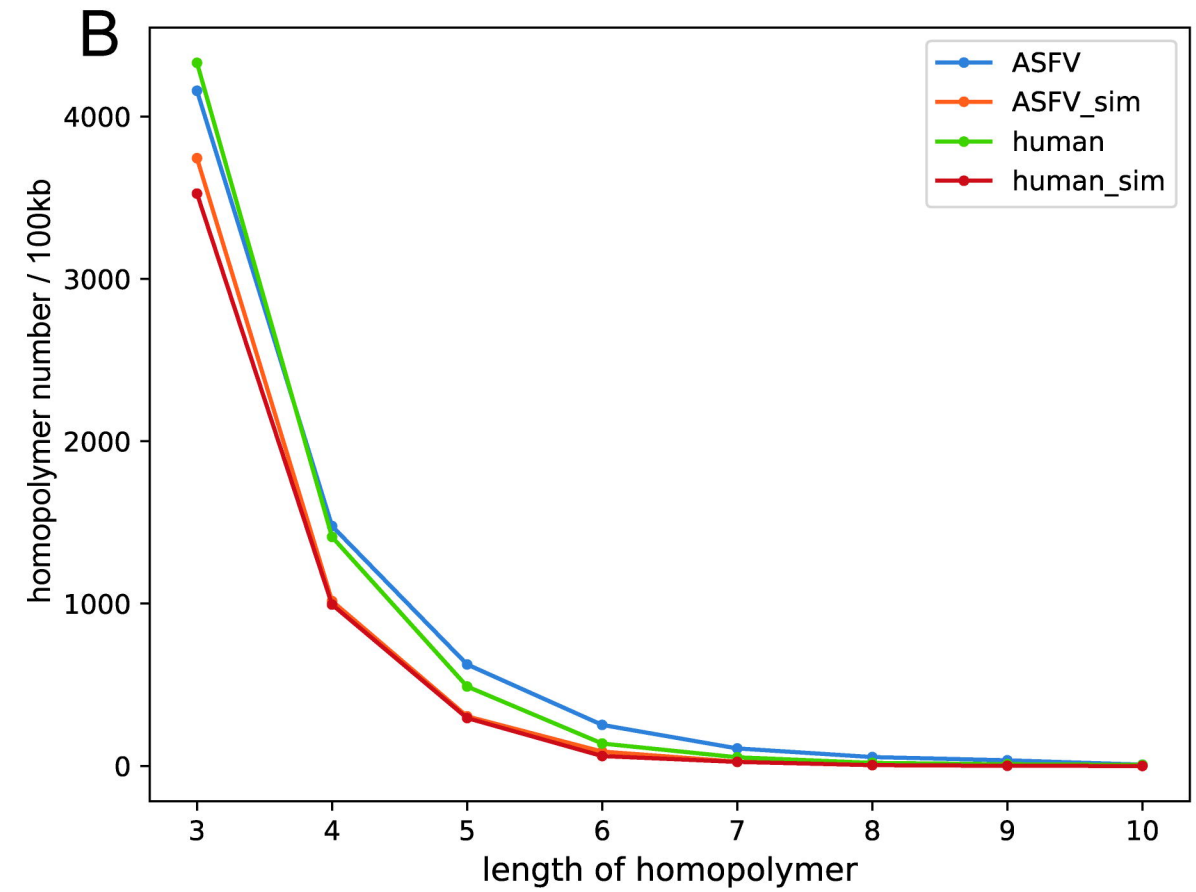


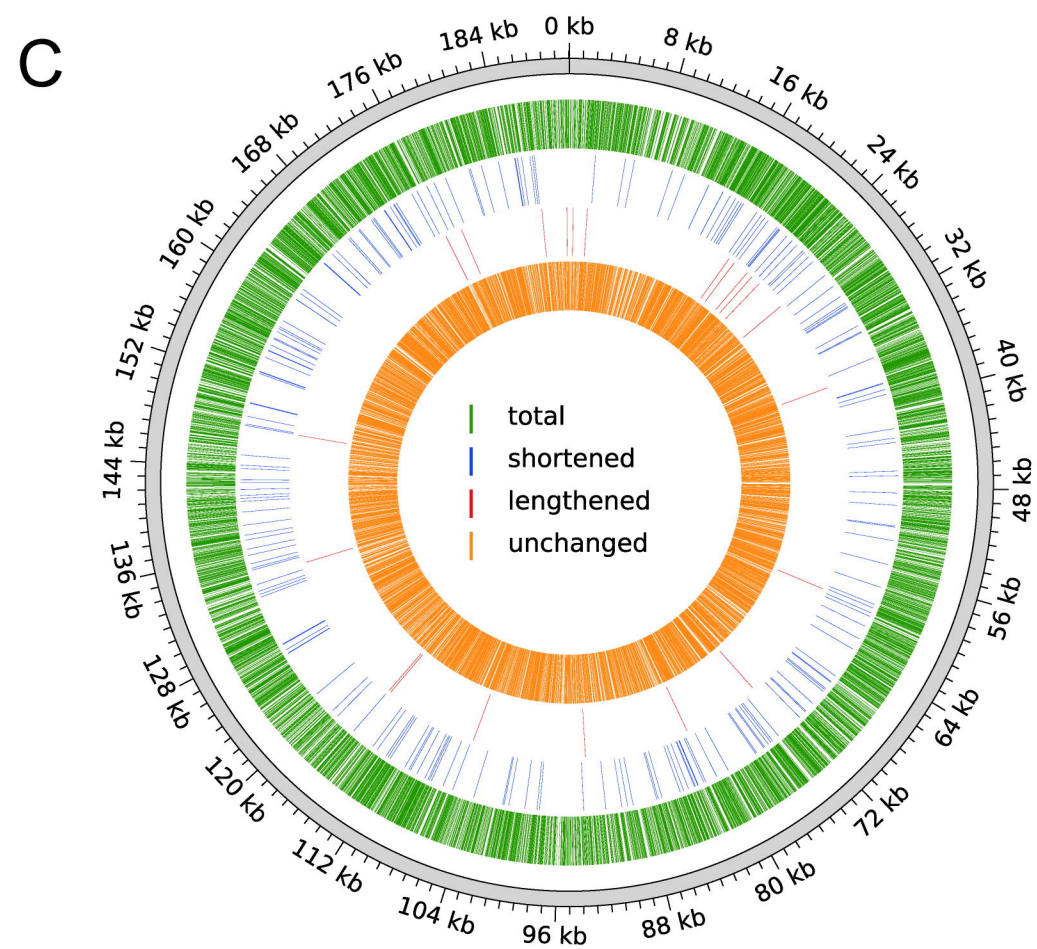
Figure 5. Phylogenetic analysis with all available ASFV whole genomes. The ShanDong2021 isolate is highlighted using bold letters and a red arrow. A magnified image of the branch of ShanDong2021 is shown in the lower right corner. Isolates are represented using their NCBI nucleotide accession numbers and their isolate names. The colors of the branches indicate the six clades identified by our study. The background color of the isolate name indicates different genotypes.

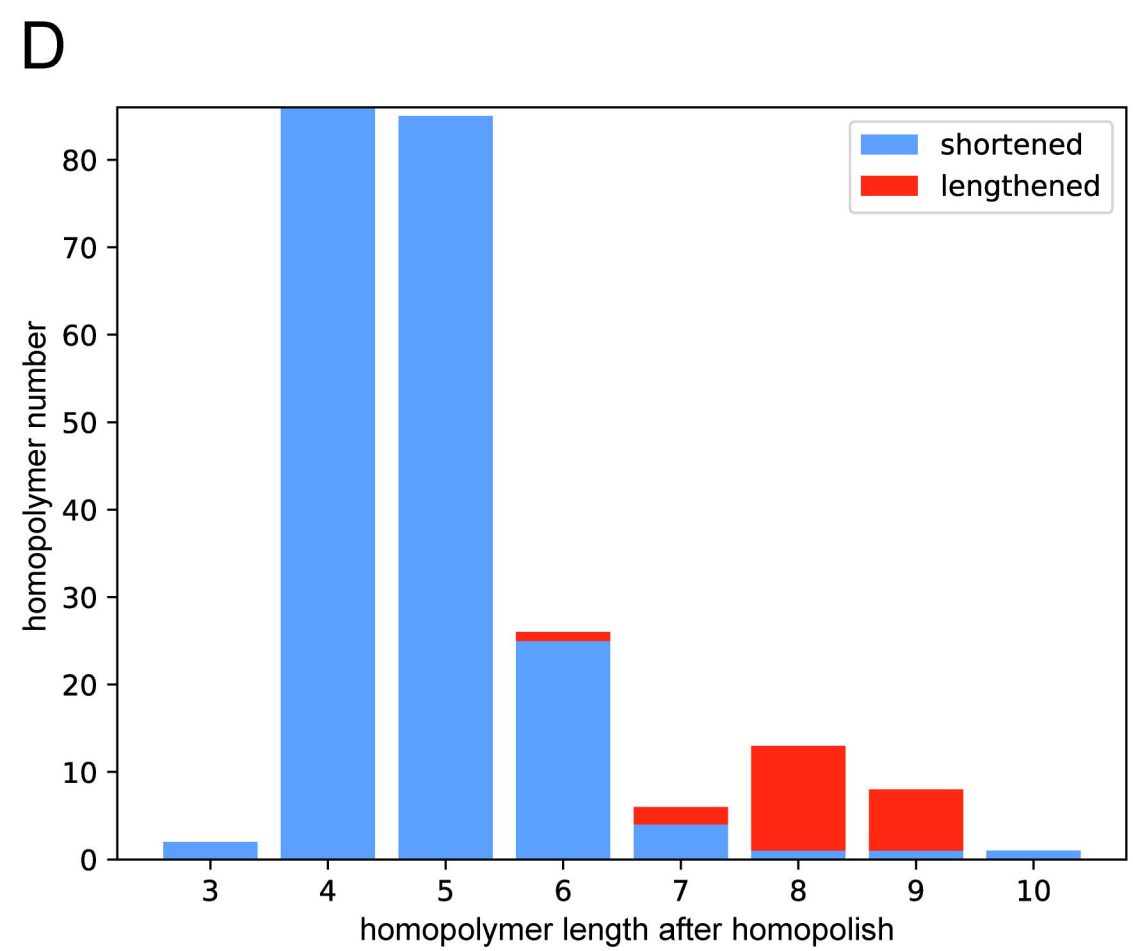

20kb
1e4
1e3
1e2
1e1
1
Coverage (log scale)
0
50000
100000
150000
Position

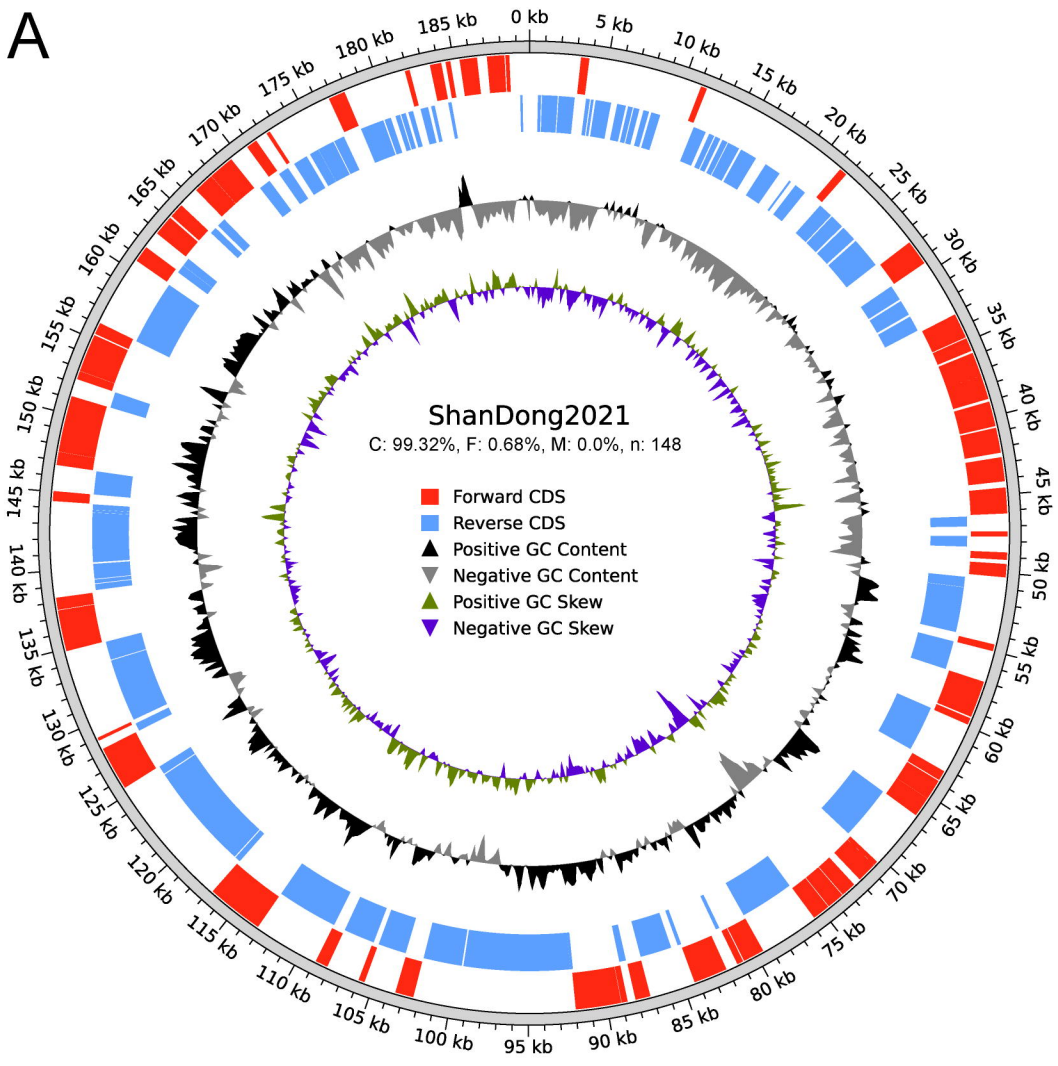

5kb
1e4
1e3
1e2
1e1
Coverage (log scale)
0
50000
100000
150000
Position

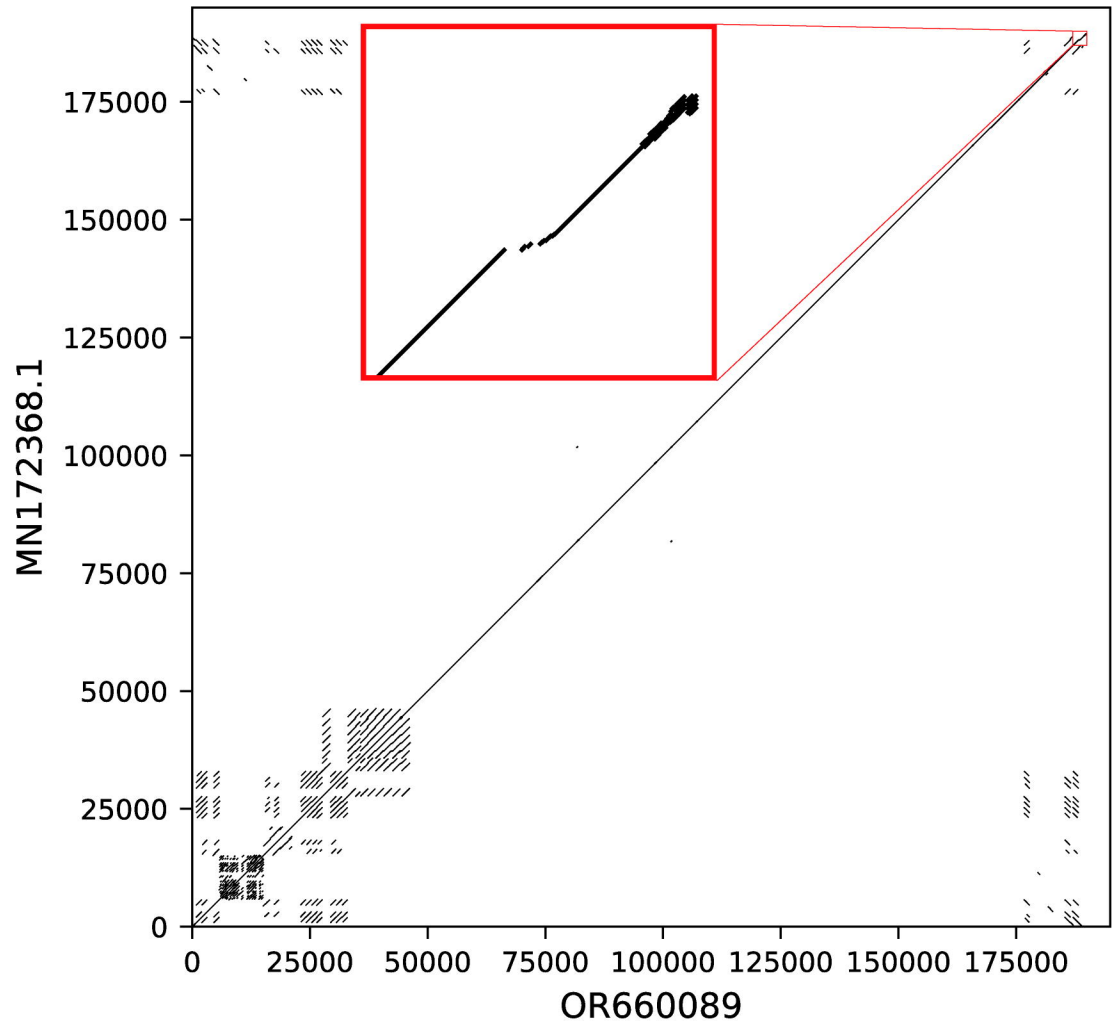

A
Complete (C)
Fragmented (F)
Missing (M)
OR290104.2
ON963982.1
MN194591.1
OR660089
ratio
0
20
40
60
80
100
before
after
before
after
before
after
before
after

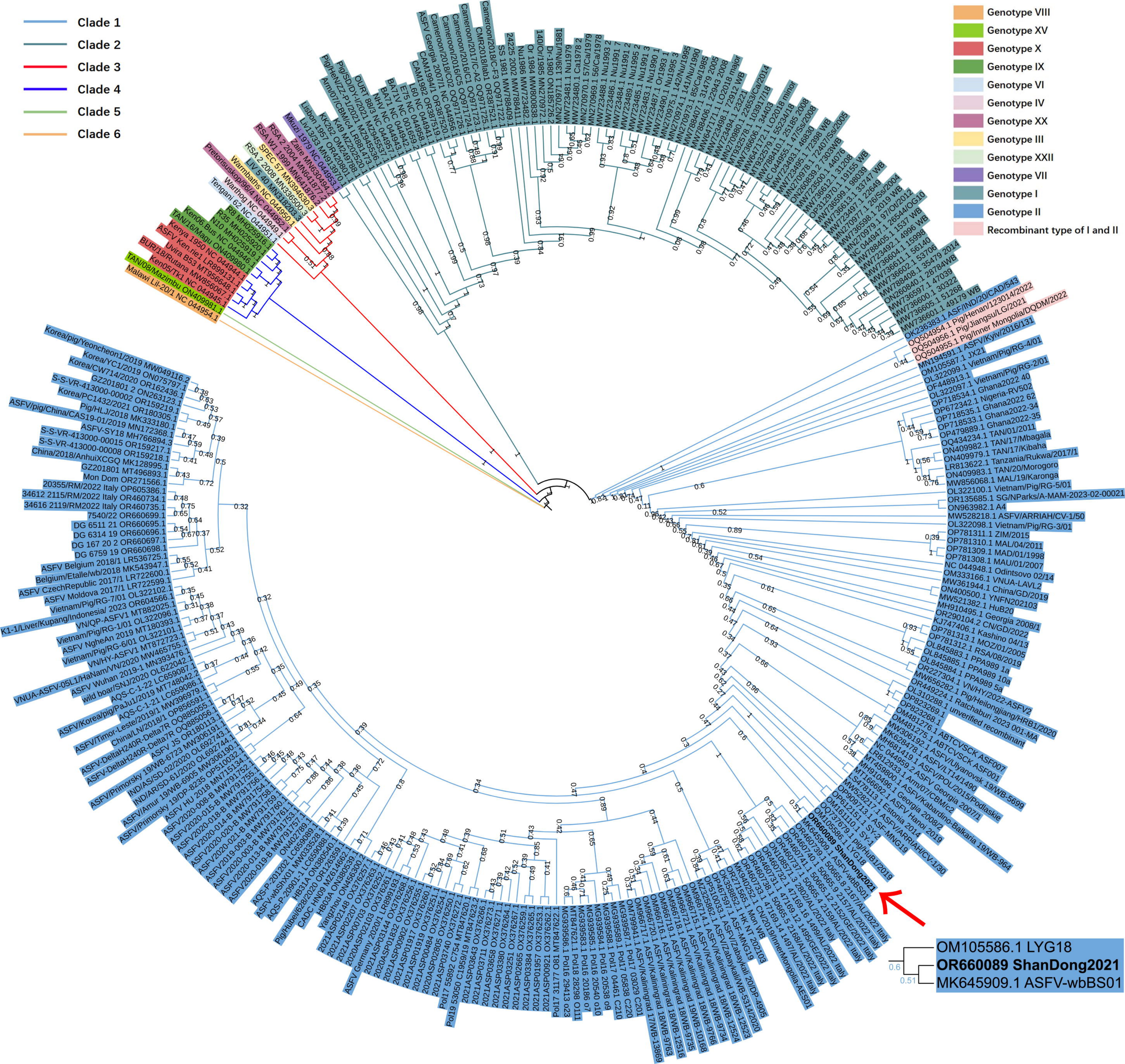

B
ASFV
ASFV_sim
human
human_sim
homopolymer number / 100kb
0
1000
2000
3000
4000
3
4
5
6
7
8
9
10
length of homopolymer


C
0 kb
8 kb
16 kb
24 kb
32 kb
40 kb
48 kb
56 kb
64 kb
72 kb
80 kb
88 kb
96 kb
104 kb
112 kb
120 kb
128 kb
136 kb
144 kb
152 kb
160 kb
168 kb
176 kb
184 kb
total
shortened
lengthened
unchanged


D
shortened
lengthened
homopolymer number
0
10
20
30
40
50
60
70
80
3
4
5
6
7
8
9
10
homopolymer length after homopolish


A
ShanDong2021
C: 99.32%, F: 0.68%, M: 0.0%, n: 148
Forward CDS
Reverse CDS
Positive GC Content
Negative GC Content
Positive GC Skew
Negative GC Skew
0 kb
5 kb
10 kb
15 kb
20 kb
25 kb
30 kb
35 kb
40 kb
45 kb
50 kb
55 kb
60 kb
65 kb
70 kb
75 kb
80 kb
85 kb
90 kb
95 kb
100 kb
105 kb
110 kb
115 kb
120 kb
125 kb
130 kb
135 kb
140 kb
145 kb
150 kb
155 kb
160 kb
165 kb
170 kb
175 kb
180 kb
185 kb


B
MN172368.1
OR660089
0
25000
50000
75000
100000
125000
150000
175000


Clade 1
Clade 2
Clade 3
Clade 4
Clade 5
Clade 6
Genotype VIII
Genotype XV
Genotype X
Genotype IX
Genotype VI
Genotype IV
Genotype XX
Genotype III
Genotype XXII
Genotype VII
Genotype I
Genotype II
Recombinant type of I and II
OM105586.1 LYG18
OR660089 ShanDong2021
MK645909.1 ASFV-wbBS01