Increased resolution of African Swine Fever Virus genome patterns based on profile HMM protein domains

Charles Masembe[1], My V.T. Phan[2,3], David L. Robertson[4], Matthew Cotten[2,4,5*]

1. College of Natural Sciences, Makerere University, Kampala, Uganda
2. Viral Genomics, Wellcome Trust Sanger Institute, Hinxton, UK
3. Department of Viroscience, Erasmus Medical Centre, Rotterdam, The Netherlands.
4. MRC-University of Glasgow Centre for Virus Research, Glasgow, UK
5. MRC/UVRI & LSHTM Uganda Research Unit, Entebbe, Uganda

* To whom correspondence should be addressed. Tel: +256 701 509 685; **Email:** matthew.cotten@lshtm.ac.uk

ABSTRACT

African Swine Fever Virus (ASFV) was originally described in Africa almost 100 years ago and is now spreading uncontrolled across Europe and Asia and threatening to destroy the domestic pork industry. Neither effective antiviral drugs nor a protective vaccine are currently available. Efforts to understand the basis for viral pathogenicity and the development of attenuated potential vaccine strains are complicated by the large and complex ASFV genome. We report here a novel method of documenting viral diversity based on profile Hidden Markov Model domains on a genome scale. The method can be used to infer genomic relationships independent of genome alignments and also reveal ASFV genome sequence differences that alter the presence of functional protein domains in the virus. We show that the method can quickly identify differences and shared patterns between virulent and attenuated ASFV strains and will be a useful tool for developing much-needed vaccines and antiviral agents to help control this virus. The tool is rapid to run and easy to implement, readily available as a simple Docker image.

INTRODUCTION

African Swine Fever Virus (ASFV), belonged to the *Asfarviridae* family, was first described in Kenya nearly 100 years ago (1). The virus is endemic in most sub-Saharan African countries where it naturally infects warthogs and bush pigs and is frequently transmitted via soft ticks. In sub-Saharan Africa, infections of warthogs and bush pigs have a typically mild disease outcome. In domestic swine or wild boars, ASFV infections can result in a more serious disease with much greater mortality between 90 – 100%. Of great concern for animal welfare and the food industry, ASFV infections are responsible for increasing swine mortality in several parts of the world (2). Outside of Africa, the virus has been previously reported in Portugal, and in Haiti in sporadic outbreaks, probably as an import from West Africa (3)(4). Since the virus's first European appearance in Georgia in 2007, the virus has spread in wild boar populations in Europe (reviewed in (5)), with currently 3,608 cases reported in wild boar and 1,413 cases in swine as of 1 June, 2019. Disturbingly high prevalence of ASFV has been found in Chinese dried pig blood used as porcine feed additives with all 21 tested samples testing positive by PCR as well as the generation of a full ASFV genome sequence (6). Furthermore, ASFV sequences have been identified in Chinese pork imported into Korea (7). These recent European and Asian incursions and outbreaks involve p72-Genotype II ASFV and appear not to involve the soft tick stage as originally observed in some parts in Africa. At the time of writing, neither antiviral drugs/agents nor an effective vaccine are available to stop the epidemic.

The ASFV virion is enveloped, spherical or pleomorphic in shape with a diameter of 175-215 nm. The virus has a linear, dsDNA genome of 170-195 kb with complementary terminal sequences. The ASFV genome encodes >150 open reading frames (ORFs) (8). In addition to known viral structural and replication proteins, there are a large number of ORFS with undefined functions. These include the multi-gene families (MGFs) that show frequent duplication, deletion or inversion across the virus family (8). Multiple examples of attenuated ASFV strains encoding changes in MGF content, indicate that MGFs have a role in ASFV virulence (9) (10) (11) (12) (13) (14) (15) (16) (17). However, the complexity of the MGF families and the nature of their sequence changes in ASFV evolution make it difficult to accurately ascribe specific changes in the ASFV genome to changes in phenotype. A simplified tool for monitoring these potentially functional changes would benefit the field and may aid in making a safe attenuated vaccine strain as well as to guide efforts to develop antiviral therapies.

The p72 gene (1,942 bp) is frequently used for PCR diagnosis of ASFV (18). Additional genes used for the diagnosis include the central variable region (CVR) of pB602L gene and p54 protein (encoded by E183L gene, an antigenic structural protein involved in the viral entry). Currently, there are 24 ASFV genotypes described based on p72 sequences (19), with the two most recent genotypes found in Ethiopia (20) and Mozambique (21). There have been efforts to classify ASFV strains, including using 3 ORFs (22) (23) (24) (25), the p72 gene (26), and the pB602L gene (27). In general, these methods have been limited to small portions of the ASFV genome (i.e. < 1% of the genome size), which are not likely to capture the full evolutionary history of the virus. Important drivers for this activity are efforts to understand the pathology of the virus infection, the components of a protective immune response and most important for vaccine development, the generation of attenuated but still immunogenic virus strains that may be used for vaccine applications. Altogether, this would help prevent and control the transmission of this virus across continents.

We have been developing the use of encoded protein domains as a classification tool for viral genomic sequence data, for example, applied to *Coronaviridae* genome sequences (28). A domain is a functional unit of a protein; different combinations of domains will give rise to different functional proteins. Instead of using differences in nucleotide or protein sequences to identify possible changes across sets of evolutionary related viral genomes, employing the domain classification would inform not only the genome changes but also potentially functional alterations of the virus genomes. All protein domains are well described in the Pfam collection, available at https://pfam.xfam.org. Novel instances of a domain and its relative distance to a reference domain can be rapidly identified in query sequences using the software HMMER-3 (29). HMMER (available at http://hmmer.org/) was developed by Eddy *et al* (29) to rapidly search a profile database for sequence homologs employing profile hidden Markov models (profile HMMs) probabilistic models. This strategy can be used to describe all domains encoded by a viral genome. A matrix of these domain scores can then be used to compare and cluster sets of ASFV genomes similar to a sequence-based phylogenetic analysis. We have developed these ideas further in this work to explore ASFV genome diversity and evolutionary relationships, to provide some functional clues for differences in viral genomes and to help identify viral elements associated with attenuation, virulence or transmissibility.

MATERIALS AND METHODS

ASFV Genome collection. All ASFV full genomes were retrieved from GenBank (5 April 2019) using the query: txid137992[Organism] AND 170000[SLEN]:200000[SLEN] yielding 48 complete genomes. Two genomes were identical MK333180 and a genome derived from dried blood products MK333181, only MK333180 was retained for a final set of 47 genomes. The GenBank entries and original literature were searched for country, date and original host (tick, warthog, wildboar or domestic pig) as well as any indication of virulence derived from the original literature. A summary of the 47 genomes used for the analysis is in Supplementary Table 1.

Pfam-A domain content. The Pfam domains encoded by ASFV genomes were identified using hmmsearch function of HMMER-3.2.1 (29), searching against the most recent Pfam database

(Pfam 32.0, September 2018, 17929 entries) (30) (31). For each genome in the collection, all ORFs ≥ 75 amino acids (aa) were collected from both reading strands and then examined for the presence of Pfam content. A domain hit was retained if the domain_i-Evalue was ≤ 0.0001. Details of each domain instance were gathered including the position in the query genome, the length, the domain_i-Evalue, and the bit-score.

Custom profile HMMs for the MGFs. All ASFV encoded MGF protein coding sequences were retrieved from GenBank as follows. An initial query to the NCBI Nucleotide database was made to retrieve complete or nearly complete ASFV genomes (txid137992[Organism] AND 170000[SLEN]:200000[SLEN] NOT patent). From the "Send to" menu, the option "all coding sequences" was selected and these entries were retrieved to a fasta file. MGF entries were selected from the complete ASFV coding sequence file by sorting for the presence of the term "MGF" in the coding sequence ID with a simple python script. This yielded a set of 660 MGF entries.

When screened for Pfam content, 127 of the 660 protein coding sequences failed to return a domain hit (at a lenient domain_i-Evalue cutoff of 0.01). These were classified in GenBank as MGF_100 (38 entries); MGF_110 (9 entries); MGF_300 (39 entries); and MGF_360 (41 entries). To increase resolution for ASFV genome comparisons, profile HMMs were prepared for these proteins as follows. The 660 MGF ORFs were clustered using Usearch (32) at an aa fraction sequence identity of 0.75. Initially clustering pilots were performed at identities of 0.95, 0.90, 0.85, 0.80. 0.75, 0.70 and 0.65 (the lowest ID cut-off recommended for Usearch clustering). The 0.75 clustering gave the best separation of the coding regions into groups that corresponded to the GenBank annotation. In general, clustering followed the annotation, however several MGFs were further divided into subfamilies at this identity cut-off resulting in a set of 45 MGF subfamilies. Each MGF subfamily was aligned with Mafft (33), and a profile HMM was built using hmmbuild (29).These custom profile HMMs were used in combination with the identified Pfam profile HMMs (see Results).

The computational tools for performing this analysis are openly available as a platform independent Docker image of the tool and instructions for installing and using the tool have been made available (see Availability section and Readme document in the Supplementary Data). The Docker image contains the Unix, python, biopython SciKit and HMMER-3 modules need to run the classification, and the set of 511 HMMs (469 from Pfam plus 45 custom profileHMMs from

MGF families) that were used to classify ASFV genomes. Outputs from the classification tool are a clustermap showing the relationship between the genomes, and a CSV table listing all domains identified in each genome, their position, length and coding strand in the genome and a flag indication high or low variance. This CSV table is useful for investigators wishing to explore the identified domains further or to investigate differences between genomes.

UK domain analysis. The UK protein coding sequence was retrieved from the GenBank entry NC_001659 for the BA71V strain and used in an online blast search (megablast default settings) to identify closely related sequences. Using the download menu, all hits (39 entries, 1 October 2019) were retrieved to a fasta file, the UK domain coding sequence from the NC_001659 genome was added, and the set was translated into protein sequences using Geneious, aligned in Mafft (33) (mafft --auto --preservecase ASFV_UKorf_set_aa.fas > ASFV_UKorf_set_aa_aln.fas) and Geneious was used to calculate pairwise aa differences and to visualize protein changes across the alignment. The Pfam domain content of the UK protein coding sequence set was determined as described above, identifying only the UK domain at a domain_i-Evalue cutoff of $\leq$ 0.0001. The domain bit-scores were collected for the set and compared to the pairwise aa differences (see Supplementary Figure 1).

The 47 ASFV full genome sequences available in GenBank were aligned using Mafft (33) and manually checked in AliView (34). Maximum-likelihood (ML) phylogenetic tree of the p72 gene was constructed in RAxML (35) under the GTRGAMMA model of substitutions and bootstrapped for 100 pseudo-replicates. The tree was mid-point rooted for clarity and branches were drawn to the scale of nucleotide substitutions per site, and bootstrap values $\geq$75% are indicated.

RESULTS

Documenting Pfam content of ASFV. Initially, we identified all profile HMM domains from the Pfam collection that were encoded in a set of 47 ASFV genomes. Using a domain_i-Evalue cutoff of 0.0001 (a measure of the probability of finding the domain by chance), 82 domains were identified at least once in the set of 47 genomes, and 17 domains were found twice or more in the set indicating repeat occurrences in some genomes (see Supplementary Table 2). The domain content and their scores (from Pfam plus custom MGF domains) were then used to examine patterns of the 47 ASFV genomes in GenBank in the following manner. Briefly, for each genome a total score for each domain was generated by summing the individual domain scores (taking into account multiple instances of the same domain). For each domain column in the matrix, the scores were normalized by dividing each value by the maximum value; domains that showed > 0.03 variance in their score across the set of 47 genomes were retained and used for hierarchical clustering. A schematic presentation of the process is shown in Figure 1.

Domain variability measured by this method. As an illustration of the domain-classification approach, we examined the UK gene's ORF encoding a 96 aa protein expressed early in ASFV infection (36). Although the protein is nonessential for growth in porcine macrophage cell cultures, deletion of the UK coding region reduces the virulence of ASFV in domestic pigs (36). A set of ASFV "UK" coding regions was retrieved from GenBank, an alignment of the proteins set is shown in Supplementary Figure 1A, revealing 22 aa differences between the most divergent forms of the protein. Following the HMMER-2 search of the UK ORFs, the Pfam domain score (bit-score) for the UK domain varies across the set with a bit-score value of 227.7 for perfect match. In support of the use of this metric, there is a highly significant negative correlation between Pfam domain score with the pairwise aa distance (Supplementary Figure 1B). Of note, the Pfam UK domain entry was constructed using the ASFV reference strain NC_001659 UK protein as a model and the HMMER-3 score is correlated with the differences of query domains from this early ASFV sequence. Thus, a HMMER-3 search can be used both to find members of a domain family in a query genome as well as to provide a quantitative score (bit-score) of the distance of the query domain from the model domain.

Documenting Pfam content of ASFV. We identified all profile HMM domains from the Pfam collection that were encoded in a set of 47 ASFV genomes. Using a domain_i-Evalue cutoff of 0.0001 (a measure of the probability of finding the domain by chance), 82 domains were identified at least once in the set of 47 genomes and 17 domains were found twice or more in the set indicating repeat occurrences in some genomes (see Supplementary Table 2). As described above, the domain content and their scores (from Pfam plus custom MGF domains) were then used to examine patterns of the 47 ASFV genomes in GenBank.

The 47 full ASFV genomes were ordered by hierarchical clustering based on the Pfam + MGF domain scores and compared to a p72 ML tree with the major genotypes in each analysis indicated by colored boxes (Figure 2). In validation of our approach, the domain-clustering (Figure 2, panel B) groups genomes in nearly the same pattern as p72 ML tree topology (Figure 2, panel

A), which is a current standard practice to genotype ASFV strains. Differences include the phylogenetic position of older genomes and those genomes obtained from tick samples. Of note, the genotype II (GII) viruses, that are spreading globally, clustered into a monophyletic group on the p72 ML tree (green shaded, Figure 2A). Interestingly, the domain clustering showed that the Estonian genome (GenBank LS478113, identified from a wildboar in 2014 (37)) possesses a large 14kb deletion, lacking functional domains MGF_110 1L-12L compared to other genotype II ASFV viruses (Figure 2B). Additionally, within the GII ASFV viruses, strains FR682468 and MH766894 show changes in the DUF4509 domain (associated with MGF_360 genes). In addition to diversity in the MGF domains, there is diversity (with variance ≥ 0.03) in the 11 domains (AAA_22, Ank_2, Ank_5, ATPase_2, mRNA_cap_enzyme, Nodulin_late, P12,RIO1, SHS2_Rpb7-N, TFIIS_M, UK) observed across different genotypes. None of these domain absence/presence are revealed from a p72 ML tree (Figure 2A) that is typically used to genotype these viruses.

Domains associated with Multigene Families (MGFs). Five MGFs have been defined (MGF 100, 110, 300, 360 and 505/530) with the naming based on the mean number of amino acids in the gene product. All annotated ORFs from 47 complete genome entries in GenBank were collected (660 total entries, MGF_100: 38; MGF_110: 148; MGF_300: 46; MGF_360: 267; MGF_505: 160 entries) and examined for Pfam domains. Three MGFs consistently encoded at least one domain (i.e. all members of that MGF family were found to encode a particular domain). These were MGF_110: domain v110, MGF_360: domain ASFV_360, MGF_505: domain DUF249. To capture the diversity in these MGFs, we prepared individual profile HMMs from a comprehensive set of MGF ORFs. Briefly, we grouped each MGF protein by aa sequence identity and identified 45 MGF subfamilies and then constructed custom profile HMMs for each of these (see Methods). We then analyzed the clustering pattern of all MGF ORFs based on their custom profile HMMs (Figure 3). Most MGFs clustered within their annotated family, evidenced by the rectangle of shared score similarities surrounding the large clusters of MGF_100 and MGF-110, MGF_360, MGF_505 (Figure 3). However, a subset of 10 MGFs appeared different from the main MGF group bearing their name (Figure 3, red boxes, IDs with asterisks). For example, several ORFs annotated as MGF_505-11L have less than 0.85 aa sequence identity (fractional identity (32)) with other MGF_505 family member and their domain scores cluster them to a unique sector of the graph (Figure 3 red box). There is a similar pattern for MGF_360-15R, MGF_300-1L and 2R, MGF_360-18R, MGF_300-4L and MGF110-12L revealing greater domain/functional variety in these genes than previously appreciated.

Changes in domain copy number. It has been previously noted that MGF counts vary with ASFV genotype and also between attenuated and virulent strains. This is illustrated in Figure 4 where we have plotted specific domain counts by sample date and virus genotype. As clearly shown in Figure 4, viruses of genotypes GII and GIX possess higher levels of MGF_110 and MGF_360 specific domains. A few domains were observed to be absent from GII and GIX genomes, for example an Ankyrin 4 domain found in some genotypes is not present in GII or GIX (Figure 4).

Of potential importance to disease status, it has been observed in several analyses that changes in MGF numbers might result in altered viral properties. A deletion of a large 5' region including multiple MGF_110 elements was associated with attenuation of an Estonian ASFV strain (37). Two GI viruses Lisboa60 (L60, KM262844, a virulent strain) and NH/P68 (NHV, KM262845, a non-virulent strain) studied for their differences in virulence revealed differences in 4 MGF families (MGF_100, MGF_110, MGF_360, MGF_505 (38). The attenuated strain NHV showed increases in MGF_100 and MGF-110 scores and decreases in MGF_360, MGF_505 scores. MGF_110-12La, an unconventional MGF_110 family member, has higher domain counts in GII strains (Figure 4, Panel C), while MGF_110-12Lb, an unconventional MGF_110 family member, has the highest domain counts in GIX Uganda viruses (Figure 4, Panel D). The Ank-4 domain is not detected in GII, GIX viruses. Ankyrin motifs are typically found in scaffolding and signaling molecules.

Analyses of paired viruses. Finally, we applied the genome-scale domain comparison method to examine pairs of ASFV strains with reported differences in virulence. Such analyses are crucial in efforts to understand the molecular basis for attenuation or virulence and to guide efforts for vaccine design.

For example, a naturally-occurring ASFV variant was recently described from Estonia that displayed attenuation in animal tests (37). The original report noted that the Estonian variant was missing 26 genes including 13 members of the MGF_110 family, 3 members of the MGF_360 family, deletions of MGF_100_1R, L83L, L60L and KP177R as well as a duplication and rearrangements (37). We applied the domain classification tool to compare the variant Estonian strain to contemporary viruses from Georgia, changes in protein domains are shown in Figure 5A with domains showing variation across the set of four related genomes indicated by changes in the cluster map. The MGF_110 and MGF_360 changes previously noted are clearly visible with reduced signals for these two families of genes (Figure 5A). Additional domain changes were observed including variations in the DUF4509, UK, PP1c_bdg and ASFV_L11L domains. The DUF4509 domain is found on a subset of MGF_360 domains and is consistent with the reported MGF_360 changes. The PP1c_bdg domain is found on a Phosphatase-1 catalytic subunit binding region that may influence apoptosis (39) and may be relevant for ASFV virulence. The ASFV_L11L domain also shows changes, this domain is found on the L11L gene which although reported to be non-essential for virus growth (40) was previously noted to be missing from attenuated viruses (37).

Other examples include the Lisboa60 (L60) virulent strain and the NH/P68 (NHV) non-virulent strain, which have been described and compared for virulence differences (38). Domain differences between the two strains confirms the previously reported changes in MGFs (100, 110, 360 and 505, Figure 5B). Also, BA71 and BA71V are a pair of virulent/attenuated ASFV strains. The BA71V strain was adapted to cell culture and showed attenuation accompanied by the loss of MGF_360 and 505 genes (41) (42). The domain differences between the two strains are consistent with previously reported differences in the MGF_360 and MGF_505 genes. In addition, the ASFV_L11L domain and a Nodulin_late domain show a change in signal in the attenuated strain (Figure 5C). The observed changes in ASFV_L11L in two quite different pairs of virulent/avirulent ASFV strains is notable and the role of the ASFV_L11L membrane protein should be re-examined in more detail.

DISCUSSION

We have demonstrated the utility of a novel method of characterizing ASFV-encoded protein diversity on a genome-scale based on profile Hidden Markov Model descriptions of conserved protein domains. The method exploits the Pfam collection of profile HMMs (43) as well as the rapid and sensitive HMMER3 software (29). The standard methods of accurately comparing large virus genomes requires the careful preparation of a full-length genome alignment of the ~190 kb ASFV genome combined with a maximum likelihood phylogenetic tree inference coupled with bootstrapping to check cluster reliability. The combined phylogenetic analysis might take several days to complete and is further complicated by the large size and frequent gene deletions and duplications in the ASFV genome making an accurate and reproducible alignment quite difficult to generate. In comparison, the domain method described here requires no alignment and can be performed from an unaligned fasta file of the genome sequences through to hierarchical clustering in minutes. The clustermap analyses reported for 47 ASFV full genomes was performed in approximately 3 minutes run-time on a standard laptop (in this case a 2018

MacBook Pro with 2.7 GHz Intel Core i7, and 16 GB of memory). The method will be useful for quality control of newly assembled genomes and for exploring novel ASFV genomes as they are sequenced and annotated, as well as for comparing genomes with varied clinical, epidemiological and phenotypic outcomes. The combination of our approaches with the viral outcomes are important in efforts to develop an effective and safe ASFV vaccine.

We have identified greater diversity in the 5 MGF families than previously noted. We further reveal the presence of a set of unconventional MGFs (Figure 3) that appear distinct to ASFV. Their presence and evolution will need to be monitored in future studies. Indeed, the process of MGF evolution may be an important part of ASFV evolution and the current work provides novel tools for monitoring changes in these possibly high consequence genes. Grouping MGF genes in only 5 categories may result in a loss of information, obscuring important details necessary for understanding ASFV transmission, virulence and attenuation.

The domain method described here also allows a rapid assessment in both the qualitive features of encoded domains, and a reported a bit-score for each identified domain, which is a protein distance from the model domain. Furthermore, the method also reports copy number changes in domains. For example, examining changes in domain instances showed that the GII ASFV strains, responsible for large global outbreak of ASF, encoded a substantial increase in several MGF gene families (Figure 4). These changes may be an important part of the replication success of the virus and warrant further investigation.

The added benefit of domain-based classification is its alignment-free feature. The resolution of any phylogenetic constructions relies heavily on accurate alignment of homologous regions of sequences. In the case of ASFV, there are differences in MGFs across different ASFV strains, either duplications or deletions, which are very difficult and time-consuming to reliably align. Furthermore, if certain genes are missing from some of the genomes for some of the alignment, this region of the alignment may be masked in the entire alignment and will not contribute to the phylogenetic signal. However, such deletions, duplications or inversions of domains are captured by the domain scoring system used and may be an important component of the increased resolution of the domain method.

In conclusion, hierarchical clustering based on profile HMM domain scores has provided a rapid method of comparing similar genomes to identify differences in the encoded proteins. We applied the method to three sets of ASFV genomes from contemporary outbreaks with known phenotype differences in their ability to replicate in and kill pigs (Figure 5). The novel method identified previously noted differences (primarily in the encoded MGF genes) but revealed an additional set of changes that should be further explored as potential virulence factors. These functions may be important to remove or alter in efforts to generate attenuated yet immunogenic viruses.

Finally, we note that the computational tools for performing this analysis are openly available as a platform independent Docker image of the tool and instructions for installing and using the tool have been made available. We hope that by providing these computational methods as easy to implement tools they may help contribute to efforts to control this virus.

AVAILABILITY

The computational tools for performing this analysis can be downloaded as a platform independent Docker image using this command (docker pull matthewcotten/asfv_class_tool). Instructions for installing and using the tool are available in the Supplementary Data Readme file.

FUNDING

This work was supported by a Marie Sklodowska-Curie Individual Fellowship, funded by European Union's Horizon 2020 research and innovation programme (MVTP, grant agreement No 799417), by a Wellcome Trust Intermediate Fellowship (CM, grant number 105684/Z/14/Z), by the ASF-RESIST African Union Commission (CM, DR, Grant AURG-II-1-196-2016) and MRC (MC, DR, MC UU 1201412).

CONFLICT OF INTEREST

All authors declare no conflict of interest.

REFERENCES

1. Eustace Montgomery,R. (1921) On A Form of Swine Fever Occurring in British East Africa (Kenya Colony). *J. Comp. Pathol. Ther.*, **34**, 159–191.

2. Pikalo,J., Zani,L., Hühr,J., Beer,M. and Blome,S. (2019) Pathogenesis of African swine fever in domestic pigs and European wild boar - lessons learned from recent animal trials. *Virus Res.*, 10.1016/j.virusres.2019.04.001.

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

4. Phologane,S.B., Bastos,A.D.S. and Penrith,M.-L. (2005) Intra- and Inter-Genotypic Size Variation in the Central Variable Region of the 9RL Open Reading Frame of Diverse African Swine Fever Viruses. *Virus Genes*, **31**, 357–360.

5. Cwynar,P., Stojkov,J. and Wlazlak,K. (2019) African Swine Fever Status in Europe. *Viruses*, **11**, 310.

6. Wen,X., He,X., Zhang,X., Zhang,X., Liu,L., Guan,Y., Zhang,Y. and Bu,Z. (2019) Genome sequences derived from pig and dried blood pig feed samples provide important insights into the transmission of African swine fever virus in China in 2018. *Emerg. Microbes Infect.*, **8**, 303–306.

7. Kim,H.-J., Lee,M.-J., Lee,S.-K., Kim,D., Seo,S.-J., Kang,H.-E. and Nam,H.-M. (2019) African Swine Fever Virus in Pork Brought into South Korea by Travelers from China, August 2018. *Emerg. Infect. Dis.*, **25**, 1231–1233.

8. Dixon,L.K., Chapman,D.A.G., Netherton,C.L. and Upton,C. (2013) African swine fever virus replication and genomics. *Virus Res.*, **173**, 3–14.

9. Burrage,T.G., Lu,Z., Neilan,J.G., Rock,D.L. and 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.

10. Afonso,C.L., Piccone,M.E., Zaffuto,K.M., Neilan,J., Kutish,G.F., Lu,Z., Balinsky,C.A., Gibb,T.R., Bean,T.J., Zsak,L., *et al.* (2004) African Swine Fever Virus Multigene Family 360 and 530 Genes Affect Host Interferon Response. *J. Virol.*, **78**, 1858–1864.

11. González,A., Calvo,V., Almazán,F., Almendral,J.M., Ramírez,J.C., de la Vega,I., Blasco,R. and Viñuela,E. (1990) Multigene families in African swine fever virus: family 360. *J. Virol.*, **64**, 2073–2081.

12. Rodriguez,J.M., Yanez,R.J., Pan,R., Salas,M.L. and Vinuela,E. (1994) Multigene Families in African Swine Fever Virus: Family 505. **68**, 6.

13. Almendral,J.M., Almazán,F., Blasco,R. and Viñuela,E. (1990) Multigene families in African swine fever virus: family 110. *J. Virol.*, **64**, 2064–2072.

14. Netherton,C., Rouiller,I. and Wileman,T. (2004) The Subcellular Distribution of Multigene Family 110 Proteins of African Swine Fever Virus Is Determined by Differences in C-Terminal KDEL Endoplasmic Reticulum Retention Motifs. *J. Virol.*, **78**, 3710–3721.

15. Agüero,M., Blasco,R., Wilkinson,P. and Viñuela,E. (1990) Analysis of naturally occurring deletion variants of african swine fever virus: Multigene family 110 is not essential for infectivity or virulence in pigs. *Virology*, **176**, 195–204.

16. Golding,J.P., Goatley,L., Goodbourn,S., Dixon,L.K., Taylor,G. and Netherton,C.L. (2016) Sensitivity of African swine fever virus to type I interferon is linked to genes within multigene families 360 and 505. *Virology*, **493**, 154–161.

17. Zsak,L., Lu,Z., Burrage,T.G., Neilan,J.G., Kutish,G.F., Moore,D.M. and 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.

18. Atuhaire,D.K., Afayoa,M., Ochwo,S., Mwesigwa,S., Okuni,J.B., Olaho-Mukani,W. and Ojok,L. (2013) Molecular characterization and phylogenetic study of African swine fever virus isolates from recent outbreaks in Uganda (2010–2013). *Virol. J.*, **10**, 247.

19. Mulumba-Mfumu,L.K., Saegerman,C., Dixon,L.K., Madimba,K.C., Kazadi,E., Mukalakata,N.T., Oura,C.A.L., Chenais,E., Masembe,C., Ståhl,K., *et al.* (2019) African swine fever: Update on Eastern, Central and Southern Africa. *Transbound. Emerg. Dis.*, 10.1111/tbed.13187.

20. Achenbach,J.E., Gallardo,C., Nieto-Pelegrín,E., Rivera-Arroyo,B., Degefa-Negi,T., Arias,M., Jenberie,S., Mulisa,D.D., Gizaw,D., Gelaye,E., *et al.* (2017) Identification of a New Genotype of African Swine Fever Virus in Domestic Pigs from Ethiopia. *Transbound. Emerg. Dis.*, **64**, 1393–1404.

21. Quembo,C.J., Jori,F., Vosloo,W. and Heath,L. (2018) Genetic characterization of African swine fever virus isolates from soft ticks at the wildlife/domestic interface in Mozambique and identification of a novel genotype. *Transbound. Emerg. Dis.*, **65**, 420–431.

22. Michaud,V., Randriamparany,T. and Albina,E. (2013) Comprehensive Phylogenetic Reconstructions of African Swine Fever Virus: Proposal for a New Classification and Molecular Dating of the Virus. *PLoS ONE*, **8**, e69662.

23. Gallardo,C., Mwaengo,D.M., Macharia,J.M., Arias,M., Taracha,E.A., Soler,A., Okoth,E., Martín,E., Kasiti,J. and 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.

24. Alkhamis,M.A., Gallardo,C., Jurado,C., Soler,A., Arias,M. and Sánchez-Vizcaíno,J.M. (2018) Phylodynamics and evolutionary epidemiology of African swine fever p72-CVR genes in Eurasia and Africa. *PLOS ONE*, **13**, e0192565.

25. Rock,D.L. (2017) Challenges for African swine fever vaccine development—“… perhaps the end of the beginning.” *Vet. Microbiol.*, **206**, 52–58.

26. Onzere,C.K., Bastos,A.D., Okoth,E.A., Lichoti,J.K., Bochere,E.N., Owido,M.G., Ndambuki,G., Bronsvoort,M. and Bishop,R.P. (2018) Multi-locus sequence typing of African swine fever viruses from endemic regions of Kenya and Eastern Uganda (2011–2013) reveals rapid B602L central variable region evolution. *Virus Genes*, **54**, 111–123.

27. Sanna,G., Dei Giudici,S., Bacciu,D., Angioi,P.P., Giammarioli,M., De Mia,G.M. and Oggiano,A. (2017) Improved Strategy for Molecular Characterization of African Swine Fever Viruses from Sardinia, Based on Analysis of p30, CD2V and *I73R / I329L* Variable Regions. *Transbound. Emerg. Dis.*, **64**, 1280–1286.

28. Phan,M.V.T., Ngo Tri,T., Hong Anh,P., Baker,S., Kellam,P. and Cotten,M. (2018) Identification and characterization of Coronaviridae genomes from Vietnamese bats and rats based on conserved protein domains. *Virus Evol.*, **4**.

29. Eddy,S.R. (2011) Accelerated Profile HMM Searches. *PLoS Comput. Biol.*, **7**, e1002195.

30. Finn,R.D., Coggill,P., Eberhardt,R.Y., Eddy,S.R., Mistry,J., Mitchell,A.L., Potter,S.C., Punta,M., Qureshi,M., Sangrador-Vegas,A., *et al.* (2016) The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.*, **44**, D279–D285.

31. El-Gebali,S., Mistry,J., Bateman,A., Eddy,S.R., Luciani,A., Potter,S.C., Qureshi,M., Richardson,L.J., Salazar,G.A., Smart,A., *et al.* (2019) The Pfam protein families database in 2019. *Nucleic Acids Res.*, **47**, D427–D432.

32. Edgar,R.C. (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, **26**, 2460–2461.

33. Katoh,K. and Standley,D.M. (2013) MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol. Biol. Evol.*, **30**, 772–780.

34. Larsson,A. (2014) AliView: a fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics*, **30**, 3276–3278.

35. Stamatakis,A. (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, **30**, 1312–1313.

36. Zsak,L., Caler,E., Lu,Z., Kutish,G.F., Neilan,J.G. and Rock,D.L. (1998) A Nonessential African Swine Fever Virus Gene UK Is a Significant Virulence Determinant in Domestic Swine. *J VIROL*, **72**, 8.

37. Zani,L., Forth,J.H., Forth,L., Nurmoja,I., Leidenberger,S., Henke,J., Carlson,J., Breidenstein,C., Viltrop,A., Höper,D., *et al.* (2018) Deletion at the 5'-end of Estonian ASFV strains associated with an attenuated phenotype. *Sci. Rep.*, **8**, 6510.

38. Portugal,R., Coelho,J., Hoper,D., Little,N.S., Smithson,C., Upton,C., Martins,C., Leitao,A. and Keil,G.M. (2015) Related strains of African swine fever virus with different virulence: genome comparison and analysis. *J. Gen. Virol.*, **96**, 408–419.

39. Jousse,C., Oyadomari,S., Novoa,I., Lu,P., Zhang,Y., Harding,H.P. and Ron,D. (2003) Inhibition of a constitutive translation initiation factor 2α phosphatase, *CReP* , promotes survival of stressed cells. *J. Cell Biol.*, **163**, 767–775.

40. Kleiboeker,S.B., Kutish,G.F., Neilan,J.G., Lu,Z., Zsak,L. and Rock,D.L. (1998) A conserved African swine fever virus right variable region gene, I11L, is non-essential for growth in vitro and virulence in domestic swine. *J. Gen. Virol.*, **79 (Pt 5)**, 1189–1195.

41. Lacasta,A., Monteagudo,P.L., Jiménez-Marín,Á., Accensi,F., Ballester,M., Argilaguet,J., Galindo-Cardiel,I., Segalés,J., Salas,M.L., Domínguez,J., *et al.* (2015) Live attenuated African swine fever viruses as ideal tools to dissect the mechanisms involved in viral pathogenesis and immune protection. *Vet. Res.*, **46**, 135.

42. Rodríguez,J.M., Moreno,L.T., Alejo,A., Lacasta,A., Rodríguez,F. and Salas,M.L. (2015) Genome Sequence of African Swine Fever Virus BA71, the Virulent Parental Strain of the Nonpathogenic and Tissue-Culture Adapted BA71V. *PLOS ONE*, **10**, e0142889.

43. Finn,R.D., Bateman,A., Clements,J., Coggill,P., Eberhardt,R.Y., Eddy,S.R., Heger,A., Hetherington,K., Holm,L., Mistry,J., *et al.* (2014) Pfam: the protein families database. *Nucleic Acids Res.*, **42**, D222–D230.

TABLE AND FIGURES LEGENDS

Figure 1. The process of genome clustering with HMMs. Each full ASFV genome was scanned for Pfam and MGF domain content (step 1), the domain scores were collected, built into a matrix and normalized to fraction of highest score in the set (step 2). Domains with low variance across the entire set were removed and hierarchical clustering of the genomes was performed using the high variance domains (step 3).

Figure 2. Panel A. The p72 maximum-likelihood phylogenetic tree. The coding sequences of p72 gene from the 47 ASFV genomes available in GenBank were aligned in Aliview. ML tree was inferred using RAxML under GTRGAMMA model of substitutions with 100 bootstraps (see Methods for further details). The tree was mid-point rooted for clarity and branches were drawn to the scale of nucleotide substitutions per site (indicated in nucleotide substitutions/site), and bootstrap values ≥75% are indicated. Genotypes are indicated by colored boxes, with the Genotype II in green. **Panel B. The domain cluster-map classification of 47 ASFV genomes.** The 47 ASFV genomes were examined by their Pfam content (see Methods). The bit-scores for all domains identified with domain_i-Evalue ≤ 0.0001 were collected for each domain, a matrix was prepared and subjected to hierarchical clustering (see Materials and Methods) based on domain whose normalized values showed ≥ 0.03 variance. In both panels, the genotypes are indicated with colored boxes. Genome IDs shown on node labels (panel A) and Y axis (panel B) include GenBank accession number, strain name, country, date, host, virulence and length in nucleotides. For both panels, genomes with incongruent placement between the two methods are highlighted with a red asterisk.

Figure 3. Hierarchical clustering of all available ASFV MGF protein sequences. All available ASFV MGF proteins (N=660) were retrieved from GenBank, clustered at an amino acid fractional identity 0.85 and a profile HMM was prepared from each of the 45 alignments (ASFV_HMM45) using HMMER3 (29). The same set of 659 proteins were then examined for ASFV_HMM45 content at an domain_i-Evalue threshold of 0.0001, bit-scores were collected and used to prepare a matrix describing the set of proteins. The matrix was then subjected to hierarchical clustering and a clustermap prepared. Each column represents one of the 45 profile HMMs, each row represents an MGF protein. Major clusters are indicated to the right, unconventional domains that do not cluster with other members bearing the same GenBank MGF family annotation are marked in the red box.

Figure 4. Changes in domain copy numbers. The total number of domains detected per genome was plotted per genome, organized by sample date and coloured by ASFV genotype (see legend insert for color code). Domains examined are panel A: Pfam v110 domain (found on MGS_110 family members), panel B: Pfam ASFV_360 domain (found on MGS_360 family members), panel C: the custom domain MGF_110-12La, panel D: The custom domain MGF_110-12Lb and panel E: the Pfam doman Ank_4. Genome ids (X axis) include Genbank accession number, strain_name, country, date, host, virulence and length in nucleotides.

Figure 5. Differences in domains between paired ASFV strains. For each panel, the indicated genomes were examined for Pfam and MGF domain content, the bit-scores for all domains identified with domain_i-Evalue ≤ 0.0001 were collected for each domain, a matrix was prepared and subjected to hierarchical clustering (see Materials and Methods) based on domain whose normalized values showed ≥ 0.03 variance. Genome ids (Y axis) include GenBank accession number, strain_name, country, date, host and virulence (lovir = low reported virulence, hivir = high reported virulence).

Supplementary Figure 1. Panel A. All available ASFV "UK" ORF sequences from Genbank full genomes were translated into protein sequences, aligned in Mafft (33) and differences in the sequences relative to the consensus were visualized using Geneious (see Methods for details). **Panel B.** HMMR3 was used to screen the protein set for Pfam profile HMMs, the UK domain was detected and the bit-score for the domain from each sequence was plotted as a function of the pairwise protein sequence distance from the consensus. The Pearson correlation coefficient for the two sets of measurements was -0.995.

Figure 1

Genomes scanned for Pfam domains

Genome1

Genome2

Genome3

etc.

Hmmer-3

+ Library of all ASFV Pfam domains

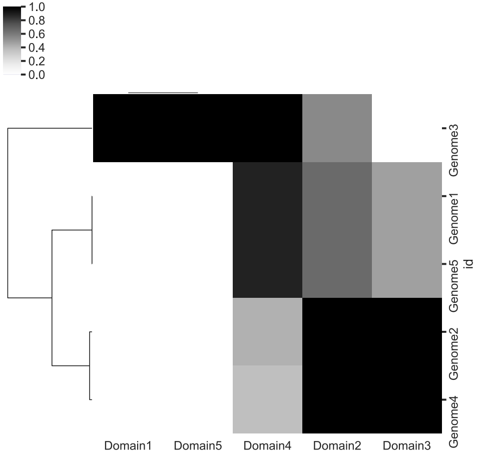
Generate array of Pfam scores

Genome	Domain1	Domain2	Domain3	Domain4	Domain5
Genome1	0	43.8	19.6	1739.7	0
Genome2	0	66.7	41.9	810.2	0
Genome3	16.8	36.4	0	1962.6	36.3
Genome4	0	66.7	41.9	715.9	0
Genome5	0	43.8	19.6	1739.5	0

Hierarchical Clustering

Remove domains with low variance

Cluster on remaining domains



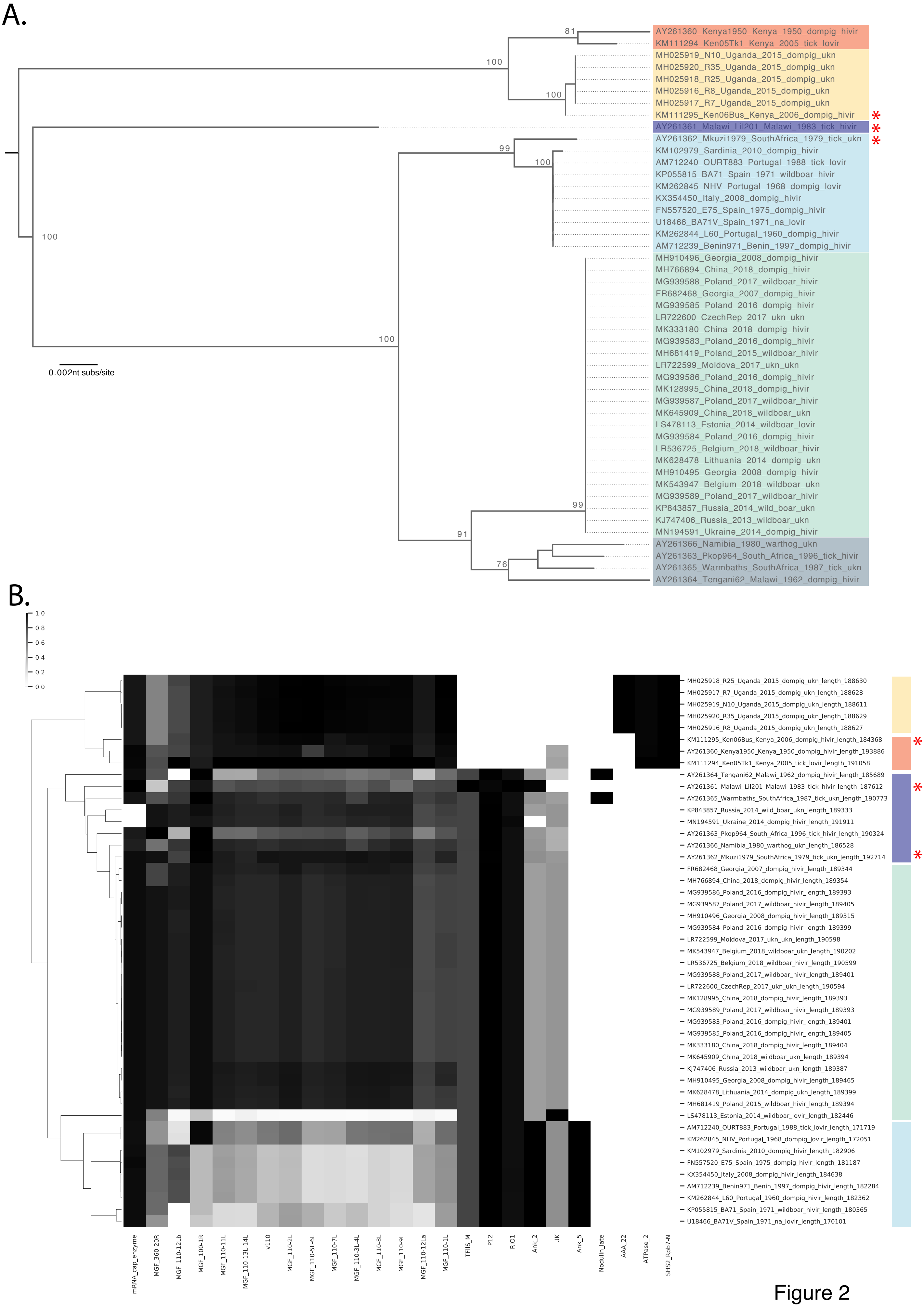

A.
81
100
100
99
100
100
100
99
91
76
0.002nt subs/site
AY261360_Kenya1950_Kenya_1950_dompig_hivir
KM111294_Ken05Tk1_Kenya_2005_tick_lovir
MH025919_N10_Uganda_2015_dompig_ukn
MH025920_R35_Uganda_2015_dompig_ukn
MH025918_R25_Uganda_2015_dompig_ukn
MH025916_R8_Uganda_2015_dompig_ukn
MH025917_R7_Uganda_2015_dompig_ukn
KM111295_Ken06Bus_Kenya_2006_dompig_hivir *
AY261361_Malawi_Lil201_Malawi_1983_tick_hivir *
AY261362_Mkuzi1979_SouthAfrica_1979_tick_ukn *
KM102979_Sardinia_2010_dompig_hivir
AM712240_OURT883_Portugal_1988_tick_lovir
KP055815_BA71_Spain_1971_wildboar_hivir
KM262845_NHV_Portugal_1968_dompig_lovir
KX354450_Italy_2008_dompig_hivir
FN557520_E75_Spain_1975_dompig_hivir
U18466_BA71V_Spain_1971_na_lovir
KM262844_L60_Portugal_1960_dompig_hivir
AM712239_Benin971_Benin_1997_dompig_hivir
MH910496_Georgia_2008_dompig_hivir
MH766894_China_2018_dompig_hivir
MG939588_Poland_2017_wildboar_hivir
FR682468_Georgia_2007_dompig_hivir
MG939585_Poland_2016_dompig_hivir
LR722600_CzechRep_2017_ukn_ukn
MK333180_China_2018_dompig_hivir
MG939583_Poland_2016_dompig_hivir
MH681419_Poland_2015_wildboar_hivir
LR722599_Moldova_2017_ukn_ukn
MG939586_Poland_2016_dompig_hivir
MK128995_China_2018_dompig_hivir
MG939587_Poland_2017_wildboar_hivir
MK645909_China_2018_wildboar_ukn
LS478113_Estonia_2014_wildboar_lovir
MG939584_Poland_2016_dompig_hivir
LR536725_Belgium_2018_wildboar_hivir
MK628478_Lithuania_2014_dompig_ukn
MH910495_Georgia_2008_dompig_hivir
MK543947_Belgium_2018_wildboar_ukn
MG939589_Poland_2017_wildboar_hivir
KP843857_Russia_2014_wild_boar_ukn
KJ747406_Russia_2013_wildboar_ukn
MN194591_Ukraine_2014_dompig_hivir
AY261366_Namibia_1980_warthog_ukn
AY261363_Pkop964_South_Africa_1996_tick_hivir
AY261365_Warmbaths_SouthAfrica_1987_tick_ukn
AY261364_Tengani62_Malawi_1962_dompig_hivir
B.
1.0
0.8
0.6
0.4
0.2
0.0
MH025918_R25_Uganda_2015_dompig_ukn_length_188630
MH025917_R7_Uganda_2015_dompig_ukn_length_188628
MH025919_N10_Uganda_2015_dompig_ukn_length_188611
MH025920_R35_Uganda_2015_dompig_ukn_length_188629
MH025916_R8_Uganda_2015_dompig_ukn_length_188627
KM111295_Ken06Bus_Kenya_2006_dompig_hivir_length_184368 *
AY261360_Kenya1950_Kenya_1950_dompig_hivir_length_193886
KM111294_Ken05Tk1_Kenya_2005_tick_lovir_length_191058
AY261364_Tengani62_Malawi_1962_dompig_hivir_length_185689
AY261361_Malawi_Lil201_Malawi_1983_tick_hivir_length_187612 *
AY261365_Warmbaths_SouthAfrica_1987_tick_ukn_length_190773
KP843857_Russia_2014_wild_boar_ukn_length_189333
MN194591_Ukraine_2014_dompig_hivir_length_191911
AY261363_Pkop964_South_Africa_1996_tick_hivir_length_190324
AY261366_Namibia_1980_warthog_ukn_length_186528
AY261362_Mkuzi1979_SouthAfrica_1979_tick_ukn_length_192714 *
FR682468_Georgia_2007_dompig_hivir_length_189344
MH766894_China_2018_dompig_hivir_length_189354
MG939586_Poland_2016_dompig_hivir_length_189393
MG939587_Poland_2017_wildboar_hivir_length_189405
MH910496_Georgia_2008_dompig_hivir_length_189315
MG939584_Poland_2016_dompig_hivir_length_189399
LR722599_Moldova_2017_ukn_ukn_length_190598
MK543947_Belgium_2018_wildboar_ukn_length_190202
LR536725_Belgium_2018_wildboar_hivir_length_190599
MG939588_Poland_2017_wildboar_hivir_length_189401
LR722600_CzechRep_2017_ukn_ukn_length_190594
MK128995_China_2018_dompig_hivir_length_189393
MG939589_Poland_2017_wildboar_hivir_length_189393
MG939583_Poland_2016_dompig_hivir_length_189401
MG939585_Poland_2016_dompig_hivir_length_189405
MK333180_China_2018_dompig_hivir_length_189404
MK645909_China_2018_wildboar_ukn_length_189394
KJ747406_Russia_2013_wildboar_ukn_length_189387
MH910495_Georgia_2008_dompig_hivir_length_189465
MK628478_Lithuania_2014_dompig_ukn_length_189399
MH681419_Poland_2015_wildboar_hivir_length_189394
LS478113_Estonia_2014_wildboar_lovir_length_182446
AM712240_OURT883_Portugal_1988_tick_lovir_length_171719
KM262845_NHV_Portugal_1968_dompig_lovir_length_172051
KM102979_Sardinia_2010_dompig_hivir_length_182906
FN557520_E75_Spain_1975_dompig_hivir_length_181187
KX354450_Italy_2008_dompig_hivir_length_184638
AM712239_Benin971_Benin_1997_dompig_hivir_length_182284
KM262844_L60_Portugal_1960_dompig_hivir_length_182362
KP055815_BA71_Spain_1971_wildboar_hivir_length_180365
U18466_BA71V_Spain_1971_na_lovir_length_170101
mRNA_cap_enzyme
MGF_360-20R
MGF_110-12Lb
MGF_100-1R
MGF_110-11L
MGF_110-13L-14L
v110
MGF_110-2L
MGF_110-5L-6L
MGF_110-7L
MGF_110-3L-4L
MGF_110-8L
MGF_110-9L
MGF_110-12La
MGF_110-1L
TFIIS_M
P12
RIO1
Ank_2
UK
Ank_5
Nodulin_late
AAA_22
ATPase_2
SHS2_Rpb7-N

Figure 2

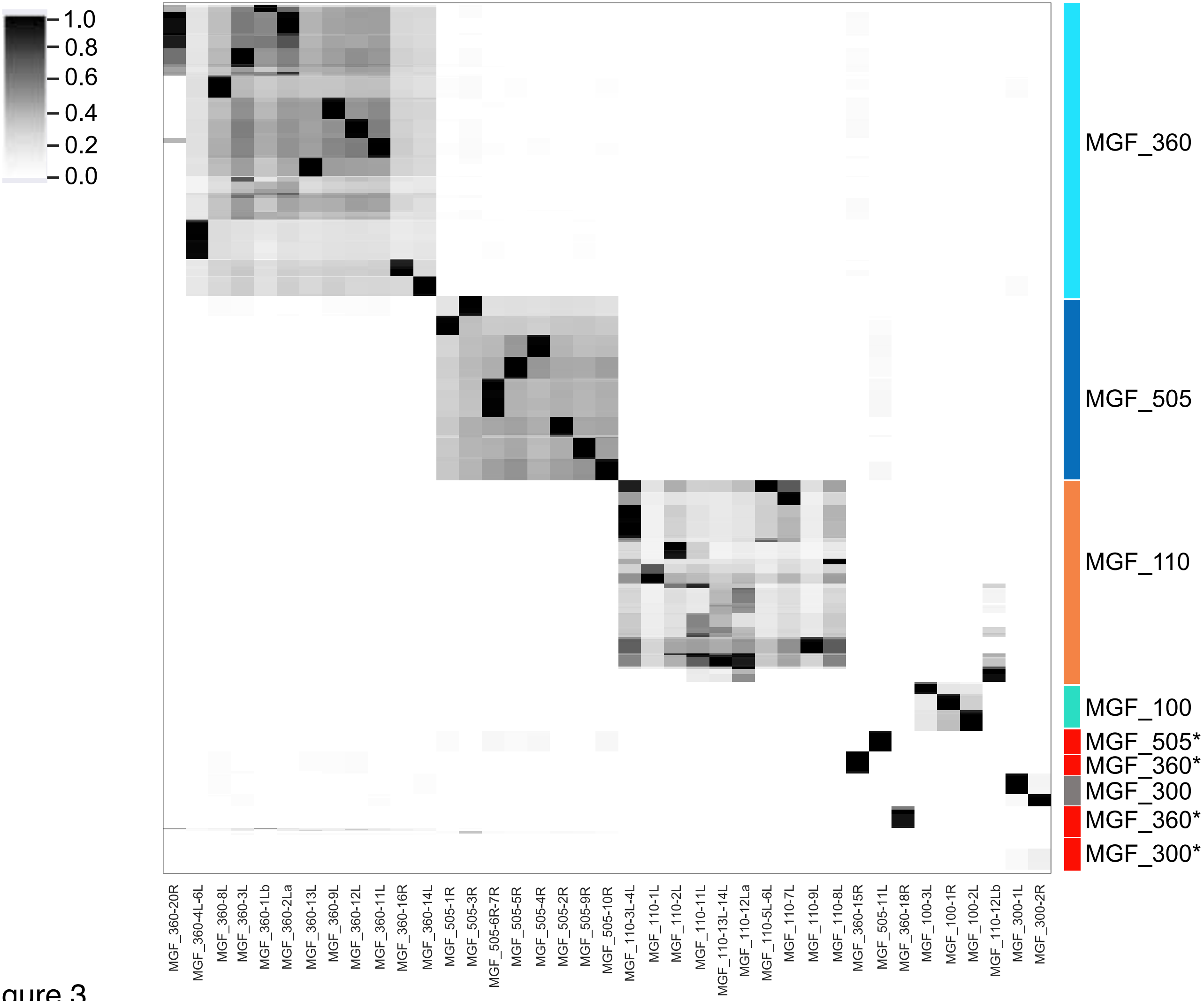

1.0
0.8
0.6
0.4
0.2
0.0
MGF_360
MGF_505
MGF_110
MGF_100
MGF_505*
MGF_360*
MGF_300
MGF_360*
MGF_300*
MGF_360-20R
MGF_360-4L-6L
MGF_360-8L
MGF_360-3L
MGF_360-1Lb
MGF_360-2La
MGF_360-13L
MGF_360-9L
MGF_360-12L
MGF_360-11L
MGF_360-16R
MGF_360-14L
MGF_505-1R
MGF_505-3R
MGF_505-6R-7R
MGF_505-5R
MGF_505-4R
MGF_505-2R
MGF_505-9R
MGF_505-10R
MGF_110-3L-4L
MGF_110-1L
MGF_110-2L
MGF_110-11L
MGF_110-13L-14L
MGF_110-12La
MGF_110-5L-6L
MGF_110-7L
MGF_110-9L
MGF_110-8L
MGF_360-15R
MGF_505-11L
MGF_360-18R
MGF_100-3L
MGF_100-1R
MGF_100-2L
MGF_110-12Lb
MGF_300-1L
MGF_300-2R

Figure 3

Figure 4

A. V110

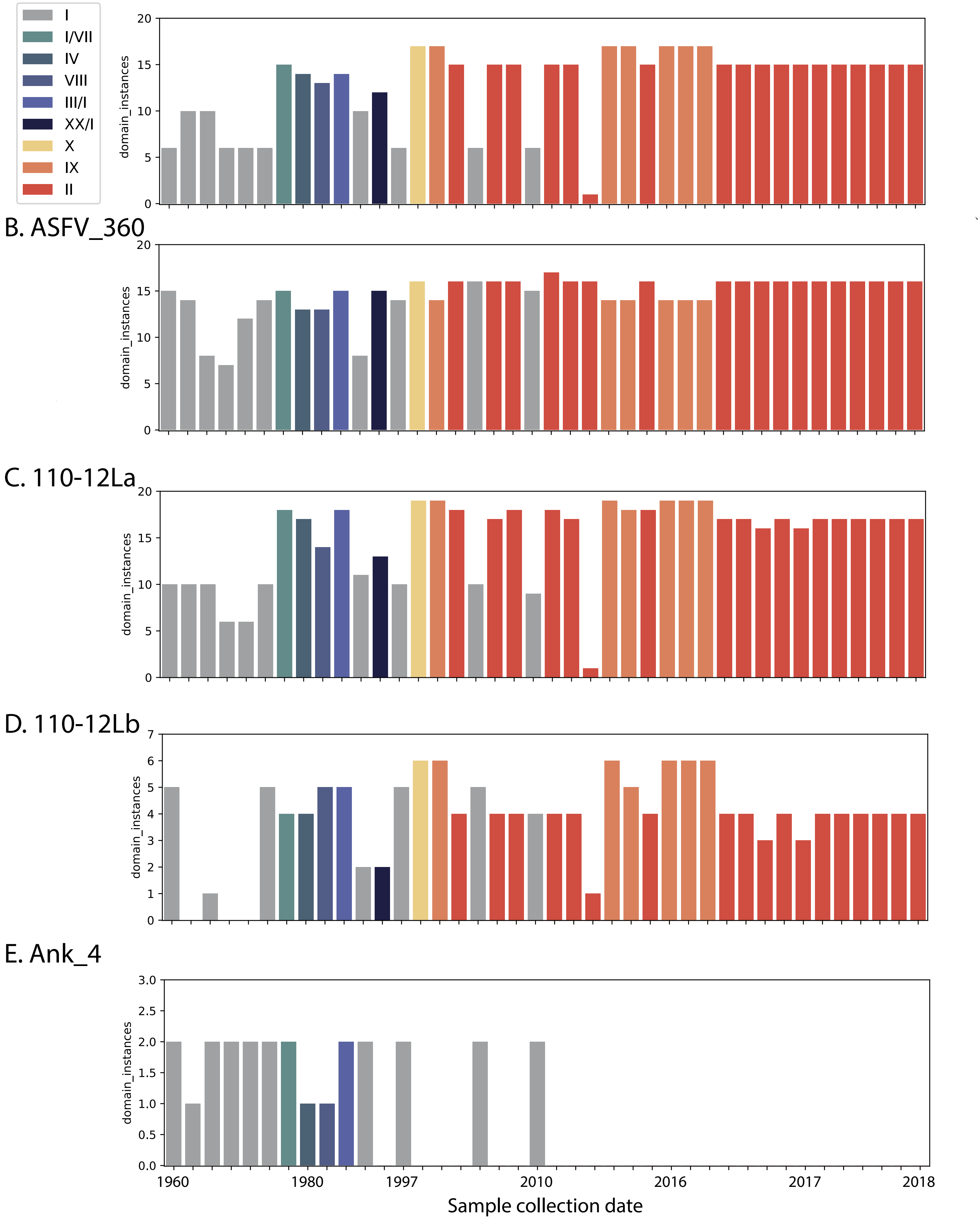
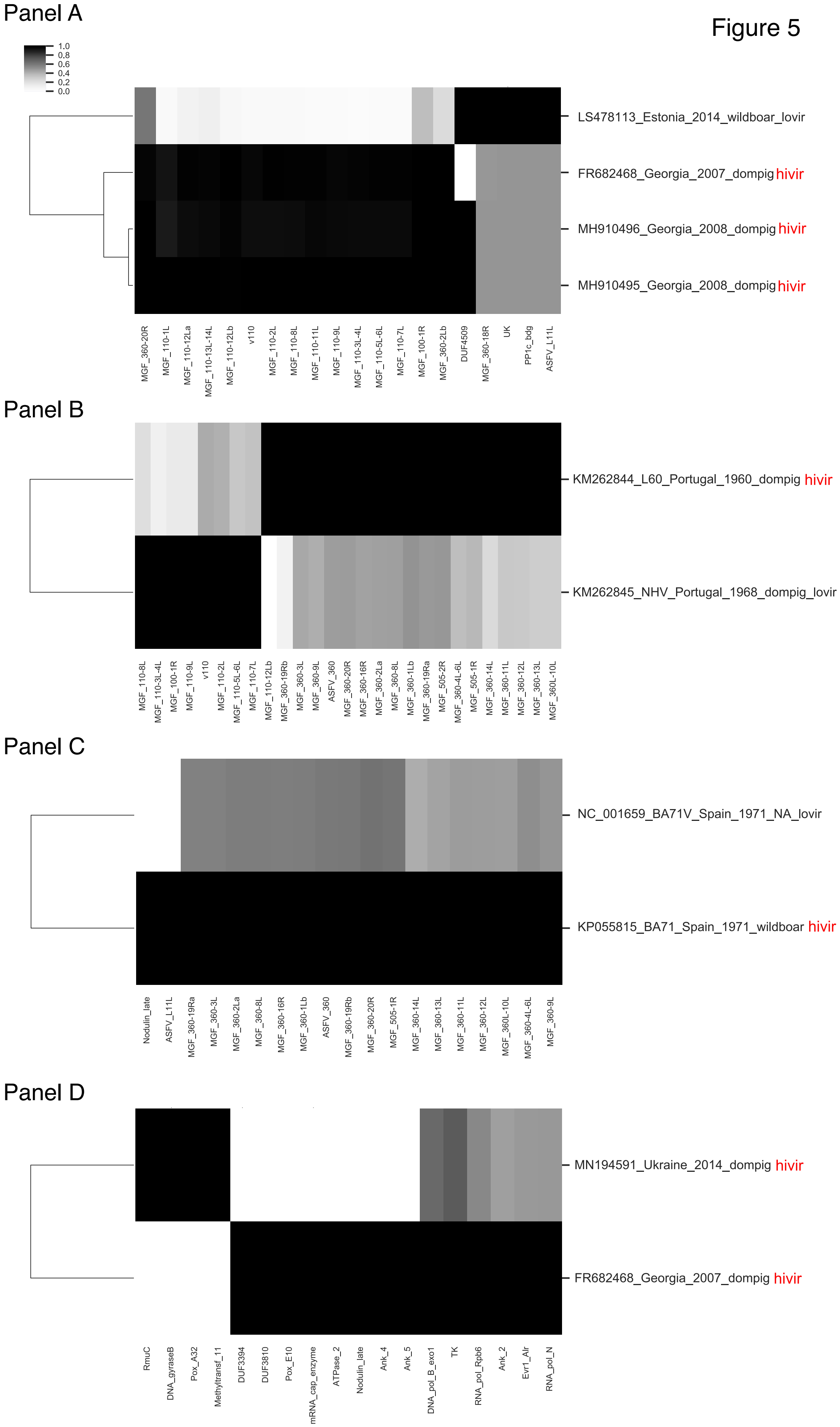
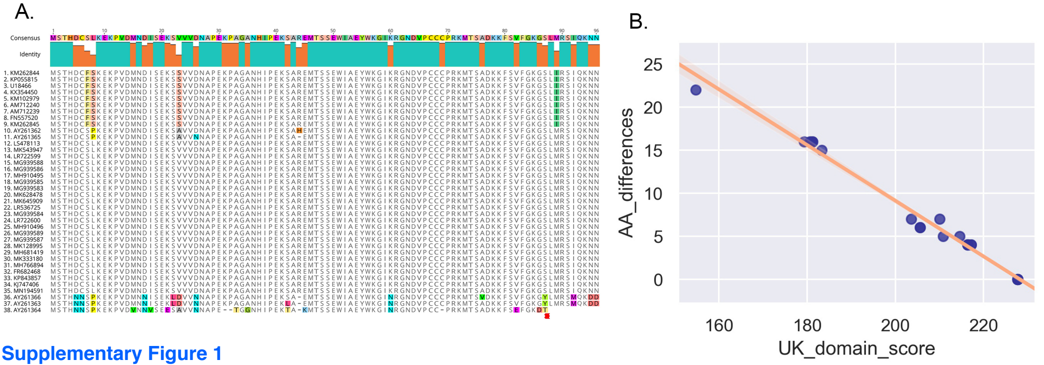


Figure 5


Panel A
1.0
0.8
0.6
0.4
0.2
0.0
LS478113_Estonia_2014_wildboar_lovir
FR682468_Georgia_2007_dompig hivir
MH910496_Georgia_2008_dompig hivir
MH910495_Georgia_2008_dompig hivir
MGF_360-20R
MGF_110-1L
MGF_110-12La
MGF_110-13L-14L
MGF_110-12Lb
v110
MGF_110-2L
MGF_110-8L
MGF_110-11L
MGF_110-9L
MGF_110-3L-4L
MGF_110-5L-6L
MGF_110-7L
MGF_100-1R
MGF_360-2Lb
DUF4509
MGF_360-18R
UK
PP1c_bdg
ASFV_L11L
Panel B
KM262844_L60_Portugal_1960_dompig hivir
KM262845_NHV_Portugal_1968_dompig_lovir
MGF_110-8L
MGF_110-3L-4L
MGF_100-1R
MGF_110-9L
v110
MGF_110-2L
MGF_110-5L-6L
MGF_110-7L
MGF_110-12Lb
MGF_360-19Rb
MGF_360-3L
MGF_360-9L
ASFV_360
MGF_360-20R
MGF_360-16R
MGF_360-2La
MGF_360-8L
MGF_360-1Lb
MGF_360-19Ra
MGF_505-2R
MGF_360-4L-6L
MGF_505-1R
MGF_360-14L
MGF_360-11L
MGF_360-12L
MGF_360-13L
MGF_360L-10L
Panel C
NC_001659_BA71V_Spain_1971_NA_lovir
KP055815_BA71_Spain_1971_wildboar hivir
Nodulin_late
ASFV_L11L
MGF_360-19Ra
MGF_360-3L
MGF_360-2La
MGF_360-8L
MGF_360-16R
MGF_360-1Lb
ASFV_360
MGF_360-19Rb
MGF_360-20R
MGF_505-1R
MGF_360-14L
MGF_360-13L
MGF_360-11L
MGF_360-12L
MGF_360L-10L
MGF_360-4L-6L
MGF_360-9L
Panel D
MN194591_Ukraine_2014_dompig hivir
FR682468_Georgia_2007_dompig hivir
RmuC
DNA_gyraseB
Pox_A32
Methyltransf_11
DUF3394
DUF3810
Pox_E10
mRNA_cap_enzyme
ATPase_2
Nodulin_late
Ank_4
Ank_5
DNA_pol_B_exo1
TK
RNA_pol_Rpb6
Ank_2
Evr1_Alr
RNA_pol_N



Supplementary Figure 1

Supplementary Table 1

	GenBank_Acc	ID	Collection_date	Host	Virulence	Genotype
1	AY261360	AY261360_Kenya_1950_Kenya_1950_dompig_hivir	1950	dompig	hivir	X
2	KM262844	KM262844_L60_Portugal_1960_dompig_hivir	1960	dompig	hivir	I
3	AY261364	AY261364_Tengani62_Malawi_1962_dompig_hivir	1962	dompig	hivir	I
4	KM262845	KM262845_NHV_Portugal_1968_dompig_lovir	1968	dompig	lovir	I
5	KP055815	KP055815_BA71_Spain_1971_wildboar_hivir	1971	wildboar	hivir	I
6	U18466	U18466_BA71V_Spain_1971_na_lovir	1971	ukn	lovir	I
7	FN557520	FN557520_E75_Spain_1975_dompig_hivir	1975	dompig	hivir	I
8	AY261362	AY261362_Mkuzi1979_SouthAfrica_1979_tick_ukn	1979	ukn	ukn	I/VII
9	AY261366	AY261366_Warthog_Namibia_1980_warthog_ukn	1980	warthog	ukn	IV
10	AY261361	AY261361_Malawi_Lil201_Malawi_1983_tick_hivir	1983	tick	hivir	VIII
11	AY261365	AY261365_Warmbaths_SouthAfrica_1987_tick_ukn	1987	tick	ukn	III/I
12	AM712240	AM712240_OURT883_avirulent_Portugal_1988_tick_lovir	1988	tick	lovir	I
13	AY261363	AY261363_Pretorisuskop964_South_Africa_1996_tick_hivir	1996	tick	hivir	XX/I
14	AM712239	AM712239_Benin971_Benin_1997_dompig_hivir	1997	dompig	hivir	I
15	KM111294	KM111294_Ken05Tk1_Kenya_2005_tick_lovir	2005	tick	lovir	X
16	KM111295	KM111295_Ken06Bus_Kenya_2006_dompig_hivir	2006	dompig	hivir	IX
17	FR682468	FR682468_Georgia_Georgia_2007_dompig_hivir	2007	dompig	hivir	II
18	KX354450	KX354450_47Ss2008_Italy_2008_dompig_hivir	2008	dompig	hivir	I
19	MH910495	MH910495_Georgia_2008_dompig_hivir	2008	dompig	hivir	II
20	MH910496	MH910496_Georgia_2008_dompig_hivir	2008	dompig	hivir	II
21	KM102979	KM102979_26544OG10_Sardinia_2010_dompig_hivir	2010	dompig	hivir	I
22	KJ747406	KJ747406_Kashino0413_Kashino_Russia_2013_wildboar_ukn	2013	wildboar	ukn	II
23	KP843857	KP843857_Odintsovo_Russia_2014_wildboar_ukn	2014	wildboar	ukn	II
24	LS478113	LS478113_Estonia_Estonia_2014_wildboar_lovir	2014	wildboar	lovir	II
25	MK628478	MK628478_Lithuania_2014_dompig_ukn	2014	dompig	ukn	II
26	MN194591	MN194591_Ukraine_2014_dompig_hivir	2014	dompig	hivir	II
27	MH025916	MH025916_R8_Uganda_2015_dompig_ukn	2015	dompig	ukn	IX
28	MH025917	MH025917_R7_Uganda_2015_dompig_ukn	2015	dompig	ukn	IX
29	MH025918	MH025918_R25_Uganda_2015_dompig_ukn	2015	dompig	ukn	IX
30	MH025919	MH025919_N10_Uganda_2015_dompig_ukn	2015	dompig	ukn	IX
31	MH025920	MH025920_R35_Uganda_2015_dompig_ukn	2015	dompig	ukn	IX
32	MH681419	MH681419_POL2015Podlaskie_Poland_2015_wildboar_hivir	2015	wildboar	hivir	II
33	MG939583	MG939583_20186_Poland_2016_dompig_hivir	2016	dompig	hivir	II
34	MG939584	MG939584_20538_Poland_2016_dompig_hivir	2016	dompig	hivir	II
35	MG939585	MG939585_20540_Poland_2016_dompig_hivir	2016	dompig	hivir	II
36	MG939586	MG939586_29413_Poland_2016_dompig_hivir	2016	dompig	hivir	II
37	MG939587	MG939587_03029_Poland_2017_wildboar_hivir	2017	wildboar	hivir	II
38	MG939588	MG939588_04461_Poland_2017_wildboar_hivir	2017	wildboar	hivir	II
39	MG939589	MG939589_05838 _Poland_2017_wildboar_hivir	2017	wildboar	hivir	II
40	LR722599	LR722599_Moldova_2017_ukn_ukn	2017	ukn	ukn	II
41	LR722600	LR722600_CzechRep_2017_ukn_ukn	2017	ukn	ukn	II
42	MH766894	MH766894_SY18_China_2018_dompig_hivir	2018	dompig	hivir	II
43	MK128995	MK128995_AnhuiXCGQ_China_2018_dompig_hivir	2018	dompig	hivir	II
44	LR536725	LR536725_Belgium_2018_wildboar_hivir	2018	wildboar	hivir	II
45	MK333180	MK333180_HLJ_China_2018_dompig_hivir	2018	dompig	hivir	II
46	MK543947	MK543947_Belgium_2018_wildboar_ukn	2018	wildboar	ukn	II
47	MK645909	MK645909_China_2018_wildboar_ukn	2018	wildboar	ukn	II

Supplementary_Table_2_ASFV_PFAM_domains, page 1

Domain	Hits_eval0.01	Hits_eval0.001	Hits_eval0.0001	Hits_eval0.00001
ASFV_360	622	615	615	615
v110	547	546	546	546
DUF249	406	406	406	406
Helicase_C	169	169	168	85
ResIII	168	168	125	84
DEAD	120	120	120	84
SNF2_N	89	84	84	84
Herpes_ori_bp	84	84	84	84
AP_endonuc_2	43	43	43	43
Pox_MCEL	43	43	43	43
DNA_pol_B_exo1	43	43	43	43
RNA_pol_Rpb1_5	43	43	43	43
ASFV_J13L	43	43	43	43
TK	43	43	43	43
PP1c_bdg	43	43	43	43
polyprenyl_synt	47	42	42	42
zf-FCS	42	42	42	42
D5_N	42	42	42	42
RNA_pol_Rpb2_2	42	42	42	42
RNA_pol_L_2	42	42	42	42
Ribonuc_red_lgC	42	42	42	42
YqaJ	42	42	42	42
Thymidylate_kin	42	42	42	42
BIR	42	42	42	42
ASFV_p27	42	42	42	42
UL45	42	42	42	42
DNA_pol_B	42	42	42	42
RNA_pol_Rpb1_3	42	42	42	42
DNA_topoisoIV	42	42	42	42
Pkinase_Tyr	42	42	42	42
Ribonuc_red_lgN	42	42	42	42
Capsid_NCLDV	42	42	42	42
TFIIS_C	42	42	42	42
RNA_pol_Rpb6	42	42	42	42
Pkinase	42	42	42	42
ASFV_L11L	42	42	42	42
L1R_F9L	42	42	42	42
Pox_VLTF3	42	42	42	42
A2L_zn_ribbon	42	42	42	42
TOPRIM_C	42	42	42	42
UQ_con	42	42	42	42
Bac_DNA_binding	42	42	42	42
NUDIX	42	42	42	42
PriCT_2	42	42	42	42
Bcl-2	42	42	42	42
Ribonuc_red_sm	42	42	42	42
RNA_pol_Rpb1_2	42	42	42	42
RNA_pol_Rpb1_1	42	42	42	42
RNA_pol_Rpb1_4	42	42	42	42
Aminotran_5	42	42	42	42
ERCC4	42	42	42	42
FtsJ	42	42	42	42

Supplementary_Table_2_ASFV_PFAM_domains, page 2

RNA_pol_N	42	42	42	42
RNA_pol_Rpb5_C	42	42	42	42
Evr1_Alr	42	42	42	42
DNA_ligase_A_M	42	42	42	42
dUTPase	42	42	42	42
RNA_pol_Rpb2_1	42	42	42	42
Herpes_UL52	42	42	42	42
RNA_pol_Rpb2_4	42	42	42	42
RNA_pol_Rpb2_7	42	42	42	42
RNA_pol_Rpb2_6	42	42	42	42
RNA_pol_Rpb2_3	42	42	42	42
Peptidase_C48	42	42	41	41
DNA_pol_B_thumb	41	41	41	41
UK	38	38	38	38
RIO1	34	34	34	34
RNA_pol_Rpb1_7	42	42	42	3
Pox_E10	42	42	42	2
TFIIS_M	42	42	34	1
mRNA_cap_enzyme	41	41	41	1
Pentapeptide_4	2	2	2	1
Ank_2	97	75	45	0
Pox_A32	83	1	0	0
DUF4064	54	0	0	0
ATPase_2	52	41	9	0
DX	49	0	0	0
AAA_22	48	6	5	0
Ank_4	46	44	0	0
RNA_pol_A_bac	42	8	0	0
Kinase-like	42	42	0	0
PCNA_N	42	42	0	0
P12	42	42	34	0
JIP_LZII	42	10	0	0
T5orf172	42	36	0	0
DUF3169	42	0	0	0
Spc7	42	34	0	0
DUF3810	42	0	0	0
HATPase_c	42	42	42	0
Baculo_LEF5_C	42	8	0	0
O-antigen_lig	42	0	0	0
DNA_gyraseB	42	0	0	0
RWD	42	2	0	0
SHS2_Rpb7-N	42	42	8	0
Amnionless	42	22	0	0
RNA_pol_Rpb1_6	42	10	0	0
RIFIN	42	1	0	0
RNA_pol_Rpb2_5	42	10	0	0
YibE_F	41	0	0	0
Bcl-2_3	41	0	0	0
Colicin_V	41	0	0	0
Methyltransf_11	40	0	0	0
DNA_pol3_a_NII	40	0	0	0
Nodulin_late	38	31	2	0
DUF4509	38	30	1	0

Supplementary_Table_2_ASFV_PFAM_domains, page 3

NTP_transf_2	37	10	0	0
Flavodoxin_3	36	0	0	0
APH	34	0	0	0
RmuC	34	0	0	0
EzrA	33	32	0	0
ODV-E18	33	8	0	0
Ank_5	31	29	9	0
DUF1510	31	0	0	0
Cation_ATPase_C	30	0	0	0
AAA_18	29	0	0	0
Lectin_C	26	25	0	0
dNK	25	1	0	0
SRTM1	23	0	0	0
AAA_17	20	20	0	0
MGDG_synth	19	0	0	0
EB	18	0	0	0
Ank	17	0	0	0
DUF3394	17	1	0	0
Actin_micro	14	0	0	0
HTH_26	14	0	0	0
Prok-E2_B	14	0	0	0
Rep_2	13	1	0	0
Ank_3	11	0	0	0
PAN_4	11	2	0	0
Stk19	10	0	0	0
Methyltransf_23	9	2	0	0
LapA_dom	9	0	0	0
AAA_14	9	0	0	0
CTD	8	0	0	0
Renin_r	7	0	0	0
AAA_30	7	0	0	0
DUF5381	7	0	0	0
DUF302	6	6	0	0
NB-ARC	6	0	0	0
DUF1189	6	0	0	0
SirB	6	0	0	0
SieB	6	0	0	0
DUF2105	6	0	0	0
KTI12	5	0	0	0
DUF4153	5	0	0	0
Synaptobrevin	4	0	0	0
TMEM154	4	0	0	0
Pentapeptide_3	3	1	0	0
Neurensin	2	0	0	0
Cpta_toxin	2	1	0	0
3-HAO	2	0	0	0
DUF4512	1	0	0	0
TIL	1	0	0	0
DUF587	1	0	0	0
LRRNT_2	1	0	0	0
DUF3772	1	0	0	0
ABC2_membrane_5	1	0	0	0
DUF3719	1	0	0	0

CCDC158	1	0	0	0
DUF5354	1	0	0	0
DUF4904	1	0	0	0
FUSC_2	1	0	0	0
DNA_pol_B_palm	1	0	0	0
Sigma_reg_N	1	0	0	0
TRP	1	0	0	0
Imm39	1	0	0	0
SelK_SelG	1	0	0	0
OAD_gamma	1	0	0	0
Clostridium_P47	1	0	0	0
IncA	1	0	0	0
DUF4131	1	0	0	0
Gemini_AL1_M	1	0	0	0
GP41	1	0	0	0
Methyltransf_25	1	0	0	0
Taxilin	1	0	0	0
DUF2937	1	0	0	0
7TM_GPCR_Sra	1	0	0	0
SpoIIIAH	1	0	0	0
DUF3749	1	0	0	0
RNA_helicase	1	0	0	0
DUF4952	1	0	0	0
Gly-zipper_OmpA	1	0	0	0
RE_SinI	1	0	0	0
Rft-1	1	0	0	0
Arc_trans_TRASH	1	0	0	0
Orf78	1	0	0	0

Readme document

African Swine Fever Virus (ASFV) clustering tool

# Below are instructions how to use the Docker image of the ASFV_classification tool to characterize ASFV genomes.

# This tool is written and developed by Matthew Cotten (matthew.cotten@lshtm.ac.uk) and My Phan (v.t.m.phan@erasmusmc.nl) and described in the manuscript Masembe et al. (2019) "Increased resolution of African Swine Fever Virus genome patterns based on profile HMM protein domains"

Please do not re-distribute this tool.

The tool uses HMMR-3 from Sean R. Eddy and the HMMER development team (http://hmmer.org/) as well as a subset of the Pfam 31.0 library from (https://pfam.xfam.org/). We are extremely grateful to the developers of these tools for their open sharing of software and information.

Briefly, a fasta file of ASFV genome sequences is analyzed by the tool as follows: First, query sequences are screened for Pfam and custom MGF domains identified in the query *Asfarviridae* sequences at an e-value $\leq$0.0001. The e-value, bit-score and position in the query sequence are gathered and the bit-scores are assembled into a matrix. Domains whose bit-scores vary across the set at variance >= to the variance argument are then used to cluster related ASFV sequences.

To use the classification tool:
1. Download a Docker from the website https://docs.docker.com/
Depending on your machine, you may download the Docker version for Windows, Linux, or Mac. If your current MacOS system is older than MacOS Yosemite 10.10.3, you may want to either upgrade your MacOS system. Alternatively, you can download the Docker Toolbox version at: https://docs.docker.com/docker-for-mac/docker-toolbox/

2. Follow the instructions on Docker webpage to install and test Docker.

3. Make a test_directory include your query fasta file in the directory and move to it:

```
mkdir test_directory
cd test_directory
```

4. Within that directory, retrieve and load the docker image with the following command:

```
docker pull matthewcotten/asfv_class_tool
```

5. To run the classification tool on sequences contained in the fasta file <asfv_sequences.fas>, use the following command:
Note: Replace <path_to_test_directory> with the actual path to the directory,
Replace <asfv_sequences.fas> with the actual name of the ASFV sequence file to be examined.
/workdir is an actual directory within the docker image and this should not be confused with the test_directory.
The <variance> argument sets the variance threshold for reporting domains in the final matrix used for clustering. We typically use 0.03 but investigators can experiment with alternate cutoffs.

```
docker run -ti --rm -w /workdir -v <path_to_test_directory>:/workdir matthewcotten/asfv_class_tool Drop_hunt_ASFV_for_Docker4.py <asfv_sequences.fas> <variance>
```

Output files:
A. A pdf of the clustermap.

B. A CSV table of domains identified with variance >= to the threshold that was set. For each domain identified the table lists the target genome and the location of the domain in the target genome (position, length and strand) and a variance flag (>= threshold variance in the set = high_variance, < threshold variance in the set = low_variance).