ORGANIZATION OF THE MULTIGENE FAMILIES OF AFRICAN SWINE FEVER VIRUS

JACOB IMBERY & CHRIS UPTON *
BIOCHEMISTRY AND MICROBIOLOGY, UNIVERSITY OF VICTORIA, VICTORIA, BC V8W 2Y2, CANADA

MANUSCRIPT RECEIVED 18 APRIL 2017;
ACCEPTED 13 JUNE 2017

CORRESPONDING AUTHOR

Chris Upton *
Biochemistry and Microbiology, University of Victoria, Victoria, BC V8W 2Y2, Canada

KEYWORDS

- African swine fever virus
- Multigene family
- Annotation
- Bioinformatics
- Genomics

ABSTRACT

African swine fever virus is a complex DNA virus that infects swine and is spread by ticks. Mortality rates in domestic pigs are very high and the virus is a significant threat to pork farming. The genomes of 16 viruses have been sequenced completely, but these represent only a few of the 23 genotypes. The viral genome is unusual in that it contains 5 multigene families, each of which contain 3–19 duplicated copies (paralogs). There is significant sequence divergence between the paralogs in a single virus and between the orthologs in the different viral genomes. This, together with the fact that in most of the multigene families there are numerous gene indels that create truncations and fusions, makes annotation of these regions very difficult; it has led to inconsistent annotation of the 16 viral genomes. In this project, we have created multiple sequence alignments for each of the multigene families and have produced gene maps to help researchers more easily understand the organization of the multigene families among the different viruses. These gene maps will help researchers ascertain which members of the multigene families are present in each of the viruses. This is critical because some of the multigene families are known to be associated with virus virulence.

INTRODUCTION

African Swine Fever Virus (ASFV) is a large dsDNA virus in the family *Asfarviridae*; 23 genotypes have been characterized by sequencing of the p72 gene (1, 12, 13). The virus is endemic in many regions of Africa where it infects primarily warthogs and is spread *via* the bites of soft ticks (9). Although ASFV causes mild symptoms in warthogs and produces no symptoms while replicating in the ticks, it causes very serious haemorrhagic disease in domestic pigs and wild boar. In these animals, the mortality rate approaches 100% for some ASFV strains (18). The relatively recent (2014) but extensive spread of ASFV through Africa to parts of Central Europe takes a significant toll on both small and large-scale pig farming operations in these regions, putting a large strain on the global pig trade (5).

To date, most successful viral prevention methods rely on routine degenerate PCR screening of wild pig and tick populations together with a rapid and competent diagnosis program when an outbreak is suspected. In addition, strict sanitary control procedures must be implemented to reduce the possibility of infected wild hosts interacting with domestic pigs (7). When outbreaks occur, currently, the only effective response is culling of an infected herd and the imposition of a ban on the movement of adjacent herds (2). This produces serious economic problems for the farmers and may incentivize noncompliance. Clearly, an unhindered pork trade would be very beneficial and benefit a large proportion of the population. Between 2014 and 2015, close to $55 million was spent on ASFV prevention in the Baltic States alone, which was considered to have prevented US$4.5 billion in potential losses (7).

Sixteen full ASFV genomes have been sequenced to date, and more than 100 will be sequenced in the next 2–3 years (E. Okoth, personal communication). The availability of these genome sequences is important because comparative genomics analyses will allow researchers to better correlate gene content and amino acid sequence variation with virulence and antigenic variation. However, all ASFV isolates have at least 5 multigene families (MGF) that are made up of sets of paralogs, which are frequently but not always arranged in tandem. Not only do the different viruses have different numbers of these paralogs, but they frequently have indels that remove multiple genes and partial genes resulting in some gene fusions (4,6,14). Consequently, when these viral genomes are aligned by software tools these regions are not aligned correctly. The problem is made more complicated by the fact that the individual MGFs have sometimes been mis-annotated due to failure to identify the correct ortholog groups within the sets of paralogs. Correct identification of the members of these MGFs is especially important because these genes have been linked with virus virulence (6).

The challenge of developing an effective vaccine stems, in part, from the high antigenic diversity distributed among the different strains of ASFV and therefore from the genomic variation. Although it is possible to induce immunity in pigs that protects from challenge with a homologous genotype, the

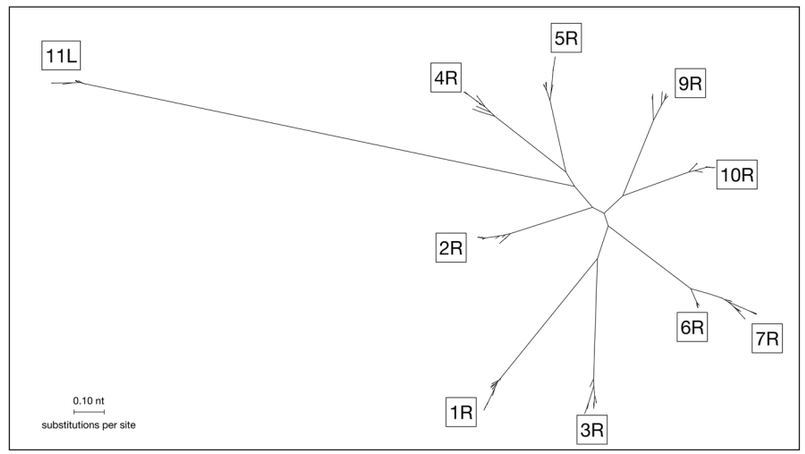


Figure 1. Maximum likelihood phylogenetic tree of MGF 505 DNA sequences was created with RAxML using the GTRGAMMA base substitution model. Sequences were aligned with MAFFT and trees visualized using MEGA7.

MGF	Paralogs	Size (bp)	Conservation	Mis-annotations
MGF 100	3	375-440	High	1
MGF 110	13	315-875	Low	27
MGF 300	3	315-800	High	5
MGF 360	19	960-1100	Moderate	17
MGF 505	10	1500-1630	Moderate	12

Table 1. Summary of the 5 MGFs for the 16 ASFV strains. The relative conservation between the paralogs inversely correlates with the number of paralogs. The "Mis-annotations" column indicates the number of

generation of protection against a heterologous genotype has proved unreliable (15). In fact, vaccination does not always adequately protect against viruses of the same genotype (19).

Here we describe the reannotation of the ASFV MGFs using a common nomenclature that will facilitate future ASFV genome comparisons and provide clarity for the discussion of the differences between viral gene sets.

METHODS

DATA SET

Genomes of the following ASFV isolates were used, GenBank accession numbers are given in parentheses: ASFV-Benin_97_1 (AM712239); ASFV-L60 (KM262844); ASFV-E75 (FN557520); ASFV-OURT_88_3 (AM712240); ASFV-NHV (KM262845); ASFV-Mkuzi_1979 (AY261362); ASFV-BA71V (U18466); ASFV-Georgia_2007/1 (FR682468); ASFV-Pretorisuskop_96_4 (AY261363); ASFV-Warmbaths (AY261365); ASFV-Warthog (AY261366); ASFV-Tengani62 (AY261364); ASFV-MWI_Lil_20_1_1983 (AY261361); ASFV-Ken05_Tk1 (KM111294); ASFV-Ken06 (KM111295); ASFV-KEN_1950 (AY261360).

PHYLOGENETIC TREE AND DOTPLOT CONSTRUCTION

A multiple sequence alignment (MSA) of the 16 complete ASFV genomes was generated using MAFFT (10). Base-By-Base (BBB; (8)) was used to visualize the MSA and highlight the differences between the genomes. Maximum-likelihood phylogenetic trees were constructed using RAxML (16) under the GTRGAMMA base substitution model using 1000 bootstrap replicates. MEGA7 (11) was used to visualize and manipulate the phylogenetic tree output.

Since alignment tools such as MAFFT treat genomes as linear syntenic sequences, they are unable to accurately display any sequence transpositions. Similarly, it can be difficult to assess small differences in the quality of the various possible alignments for the ASFV MGFs from a MSA. Therefore dotplots, which provide a 2-dimensional visualization of all nucleotides-against-all nucleotides were used to supplement genome alignments (JDotter; (3)). A dotplot was created for each individual MGF gene compared against a full length ASFV reference genome. The series of matrix alignments across the dotplots created a unique "barcode" describing the relationship of the gene to all the paralogs in the MGFs. The dotplots were especially useful for determining the breakpoints between fused paralogs.

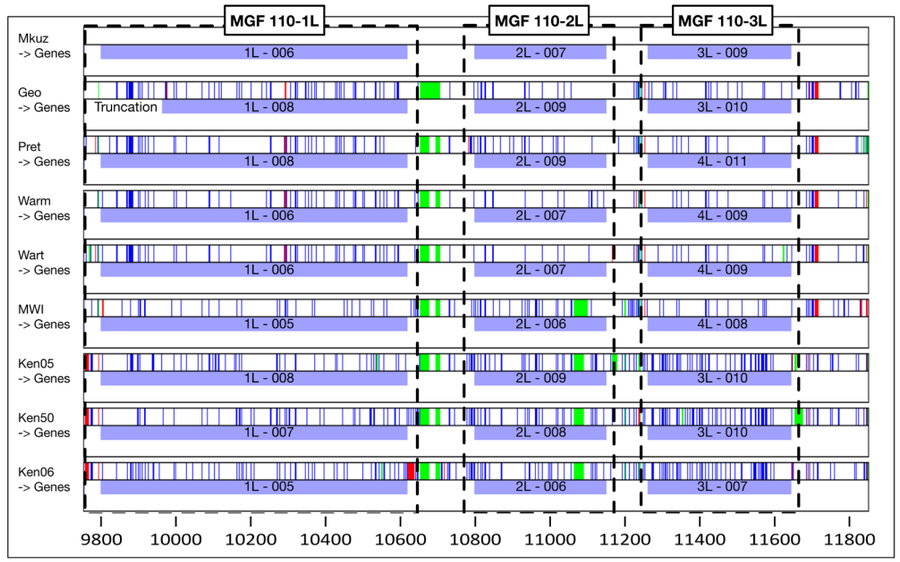


Figure 2. Visual summary from BBB of MGF 110-1L/-2L/-3L paralogs from 9 ASFV genomes. Thin vertical dark blue bars, red and green blocks represent SNPs and insertions and deletions with respect to the topmost sequence. Light blue blocks represent ORFs. Gene labels, positioned on the ORFs, indicate previous annotations (note inconsistencies) as well as gene number. Boxed gene labels indicate new nomenclature.

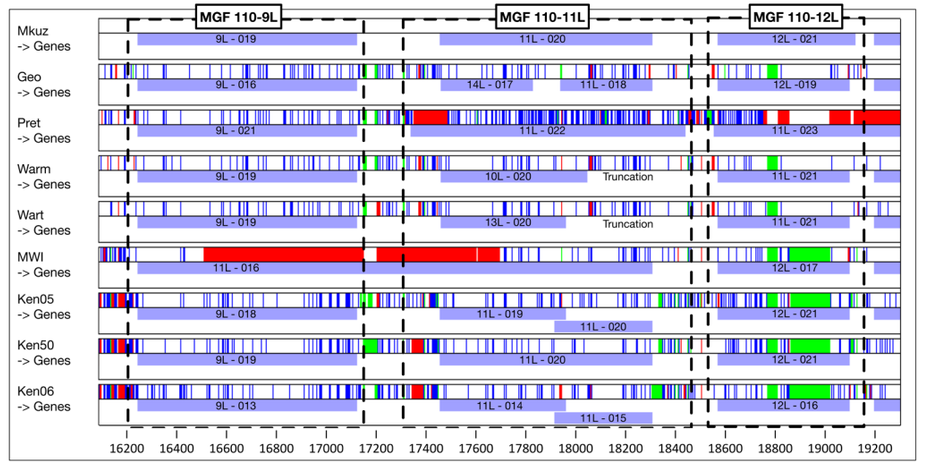


Figure 3. Visual summary displaying fragmentation of MGF 110-11L orthologs. Gene labels indicate the previously annotated orthologs. Thin vertical dark blue bars, red and green blocks represent SNPs and insertions and deletions with respect to the topmost sequence. The MGF 110-10L is not represented in this diagram due to annotation only in Warmbaths strain that is more likely a truncated MGF 110-11L ortholog. Nucleotide positions are mapped at the bottom of the figure.

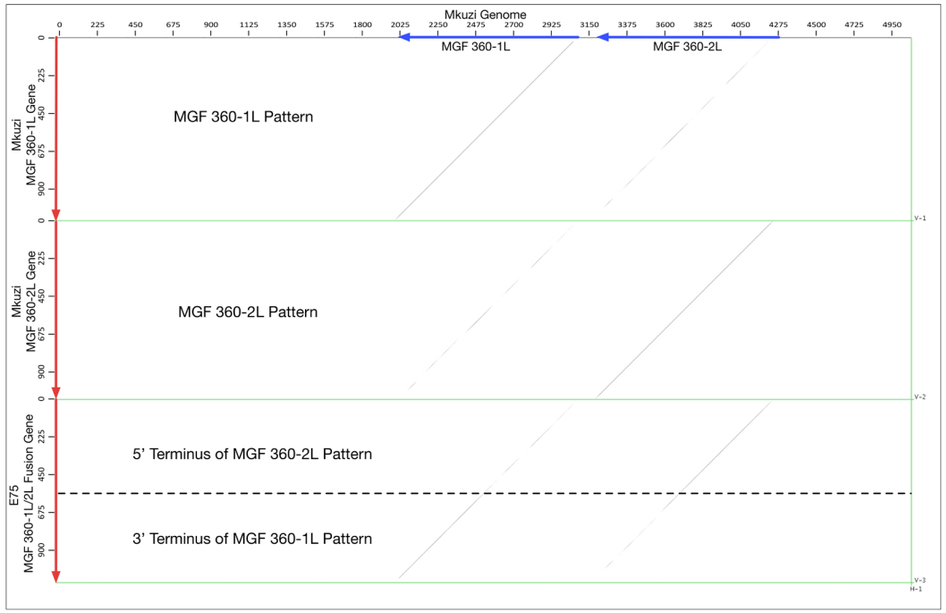


Figure 4. Dotplot of MGF-360 genes. The horizontal sequence contains ASFV-Mkuzi MGF-360-1L and MGF-360-2L genes (blue arrows). The vertical axis represents 3 sequences; the top 2 panels are the ASFV-Mkuzi MGF-360-1L and MGF-360-2L genes as controls and the bottom panel represents the ASFV-E75 MGF-360-1L/2L fusion (red arrow). The dashed line in the bottom panel separates the 5' half of 360-2L and the 3' half of 360-1L. Self-plots of genes generate a solid diagonal black line whereas plots of paralogs produce fainter intermittent lines.

CONSTRUCTION OF MGF MAPS

The MGF maps were created as vector graphics with Omnigraffle (Omni Group, Seattle) on iMac computers. These diagrams can be fully edited to incorporate new genomes and new MGF orthologs as they are discovered.

RESULTS

The goal of this analysis was to create an accurate reference map of the distribution of MGF members throughout the 16 sequenced ASFV genomes. Since the MGF members are not simply present or absent, an annotation scheme was also required to describe the various gene fragmentation/truncation/fusion patterns that exist in the different virus strains. Since we do not yet know the functional consequences of these multiple rearrangements on the biology of the viruses, the purpose of the map is primarily to flag the various differences that exist between the ASFV MGFs. In addition, due to the extremely complex nature of the indels in the MGF regions, which compound when MSAs are generated, we opted to illustrate general variations in the open reading frame (ORF) patterns rather than try to capture every single difference. Our results are sufficient to flag differences between the paralogs so that a detailed DNA sequence alignment of the region can be performed if more information is required for a particular study.

There are currently 5 known MGF series observed in ASFV. Paralogs within an MGF series are numbered chronologically as they appear in the ASFV genome and are classified

Figure 5. Example of summaries showing organization of MGF-505. The ASFV strain is given in the red boxes at the left. Paralog names are shown in the boxes at the top of the diagram along with the size of the ortholog from the reference genome, Mkuzi. The annotation of each gene (arrow head indicates direction of transcription) is in two parts: the paralog group followed by the gene number of the virus strain. If the paralog annotation is incorrect, it is labelled in red. The gene box size represents the relative gene size of orthologs, but is not to scale, nor does it reflect the size relationship between paralogs. Genes connected by grey boxes indicate the fusion of two orthologs.

as "R" or "L" indicating that this gene is either transcribed on the forward or reverse strand respectively.

Our first step in reviewing the relationships between the paralogs/orthologs of each MGF was to create a phylogenetic tree. The MSA was generated using MAFFT with the DNA sequences and the phylogenetic trees were constructed with RAxML. Figure 1 illustrates the value of the trees by showing a visual representation of the relationship between the paralogs of the MGF-505 series (MGF average size 505 amino acids). For example, the tree shows that paralogs MGF-505-6R and MGF-505-7R result from a relatively recent duplication. However, it must be appreciated that phylogenetic trees also hide the raw data, which may have valuable information about the sequences. For example, recombination events and deletions that fuse two paralogs are likely to be lost if only the tree is viewed. Therefore, to ensure that the tree generation step was not flawed by faulty input data, we checked the MSAs with BBB, a MSA editor that also provides highlighting of the differences between pairs of sequences in the MSAs. This helps the researcher recognize MSA regions that have inconsistent similarity levels and may be the result of gene fusion events.

In addition to displaying sequence alignments, BBB is capable of generating a "summary view" of sequence alignments, that captures the positions of SNPs and indel information to allow large MSAs to be shown on a single page. Figures 2 and 3 illustrate the use of BBB to view parts of MGF-110. A truncation of ASFV-Geo MGF-110-1L is shown together with previous alternate naming of ASFV MGF-110-3L orthologs (Figure 2). In MGF-110, we propose 13 paralogs whereas 14 had been previously annotated among the 16 genomes because fragments of a single gene had been annotated separately. Figure 3 shows several of the difficulties faced

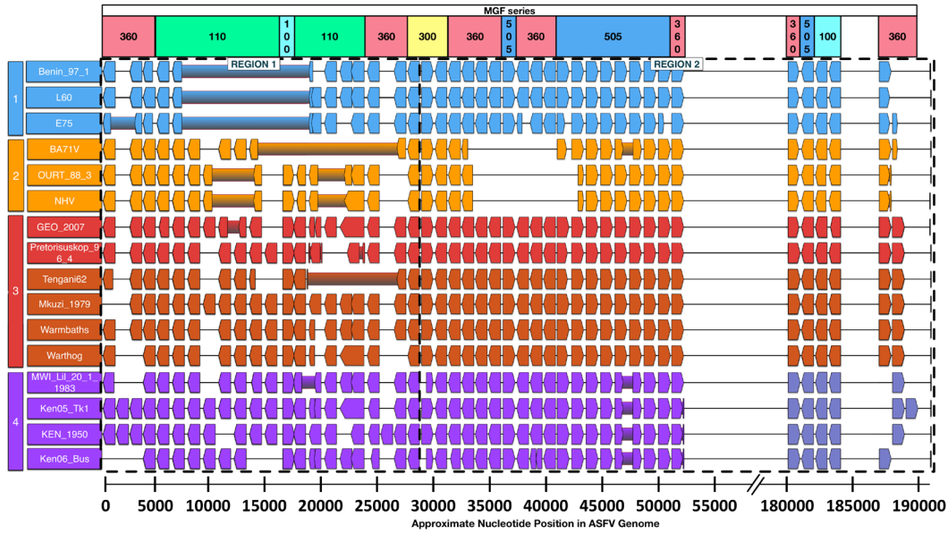


Figure 6. Display of all MGF paralogs as they appear, left to right, in the 16 ASFV strains. Columns of boxes represent individual MGF orthologs as in Figure 5. Location of the different MGF series is annotated above the diagram. The arrangement of MGF orthologs separates the genomes (coloured rows) into 4 groups that correlate well with the phylogenetic tree. It is notable that the MGF compilation can be split into 2 regions, between MGF 300-1L and MGF 300-2R, reflecting significantly more variation in region1 than 2. Nucleotide positions are given at the bottom of the figure.

in trying to annotate the MGFs consistently: 1) MGF-110-11L is fragmented in several viruses, 2) A large deletion in ASFV-MWI creates a fusion of MGF-110-9L/11L, and 3) Large indels (red and green blocks, which illustrate insertion or deletion with respect to the reference) create orthologs of significantly different sizes (MGF-110-12L). Since the ORFs are displayed by BBB across gapped alignments, they are not accurate representations of their true size.

Although MSAs do show raw data (the actual aligned DNA sequences), because they display a one-dimensional representation of the alignment they are of less use when regions of sequences may have been rearranged. In such situations, the two-dimensional presentation of global sequence comparisons from a dotplot can better show rearrangements. For example, Figure 4 shows the comparison of the ASFV-E75 MGF-360-1L/2L fusion with the 2 parental orthologs. It also shows the results of paralog comparison (1L and 2L for ASFV Mkuzi) and ortholog comparison (1L for ASFV-E75 and ASFV-Mkuzi).

After reviewing data from phylogenetic trees, MSAs and dotplots, we constructed a summary diagram for each MGF. These are presented as Supplementary Figures 1-6, which are provided at a large scale to present much greater detail . For each summary diagram, there is also an "information sheet" that explains the representations of the MGFs (Supplementary Figures 7-12). These figures are also available from the Viral Bioinformatics Resource Center, in the ASFV section (https://virology.uvic.ca/organisms/dsdna-viruses/asfarviridae/). Figure 5 is an example of one of the summary diagrams and Table 1 shows the varying complexity of the individual MGFs.

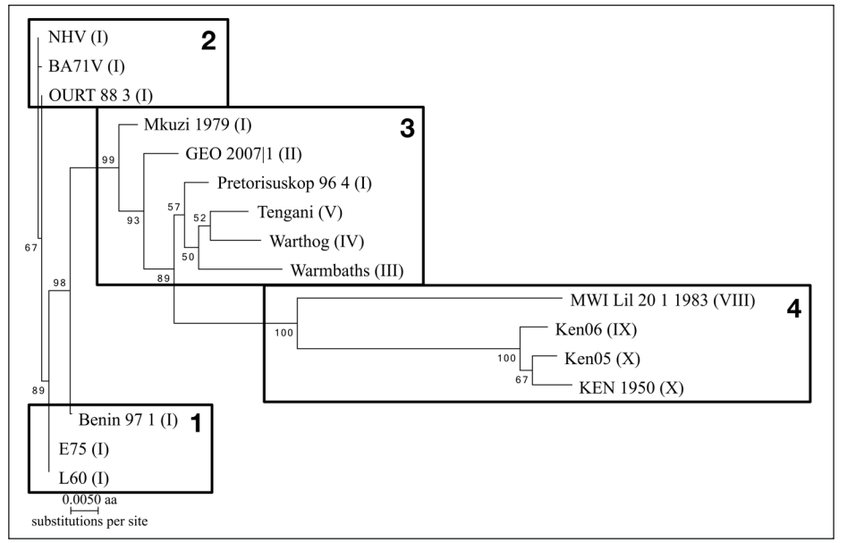


Figure 7. Phylogenetic tree for the 16 ASFV strains. Genotypes are shown in parentheses. Boxes indicate groupings defined in Figure 6. MAFFT was used to align the concatenated amino acid sequences of phosphoprotein p32, structural proteins p72 and p54, and the chaperone protein B602L. The tree was created with RAxML under a GTRGAMMA base substitution model using 1000 bootstrap replicates and visualized using MEGA7. This phylogenetic tree places ASFV strains into clades that replicate the groupings generated by overall MGF gene characteristics.

Although many of the MGFs are similar among the different ASFV strains, there are some differences that are specific to particular clades of the ASFVs. Examples of these are shown in Figure 6 with a full genome phylogenetic tree provided in Figure 7. From these figures, it is clear that the MGFs are relatively fluid, with differences appearing even between ASFV-E75 and –L60, which are very similar. However, some of this variation is expected given the overall variation between the ASFV strains. Interestingly, although these viruses are all denoted as ASFV strains, there is significant divergence between them. A comparison of the ASFV B602L, p32, p54, and p72 genes (as concatenated amino acid sequences) of the 16 ASFVs revealed that there are several viruses that have diverged to be 94 – 95 % identical (aa). In contrast, we found that poxviruses, which belong to a different family of large DNA viruses, that are classified as separate species within the Capripox or Orthopox genera may be 97 – 99 % identitical (aa) in pairwise alignments. Thus, ASFVs that are currently classified as different genotypes within a single species may well be classified into different species if taxonomic standards that are used with poxviruses were applied to ASFV.

DISCUSSION

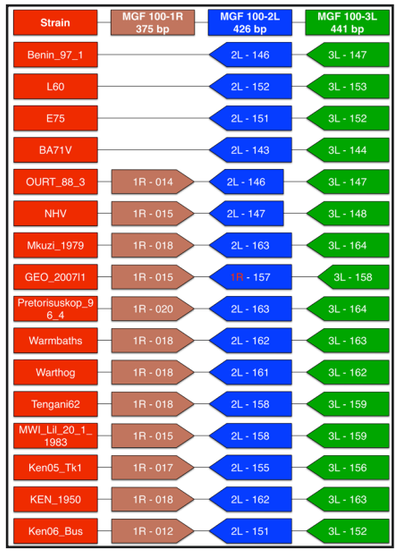
The genomic regions that encode the 5 MGFs of ASFV presented here are highly variable and are hotspots for indels making both sequence alignment and accurate annotation difficult. This has resulted in inconsistent annotation among the 16 ASFV genomes for the identification of paralogs and especially the naming of gene fusions. Since

a large number of ASFV genomes will be sequenced in the near future, we decided that standardizing the annotation and presentation of the ASFV MGFs would greatly simplify genome annotation in the future. With a better reference system for the ASFV MGFs available, we envision a 3-part process in the annotation process for ASFV genomes sequenced in the future. First, there would be a basic sequence similarity search with a set of 10 conserved genes to identify the most similar reference genome to be used with the Genome Annotation Transfer Utility tool (GATU; (17)). Second, a dotplot would be used to confirm co-linearity between the proposed reference genome and the newly sequenced target genome. Third, GATU would be used to transfer as many annotations as possible (> 95 %) from the reference genome to the target, with the use of a full genome alignment of the reference and target in BBB to confirm the positions and numbering of the members of the MGFs. As the number of sequenced and annotated genomes increases, fewer differences will be found between the new target genomes and their references. Thus, GATU will become more efficient and less annotator intervention will be required to annotate those few ORFs that GATU leaves unfinished.

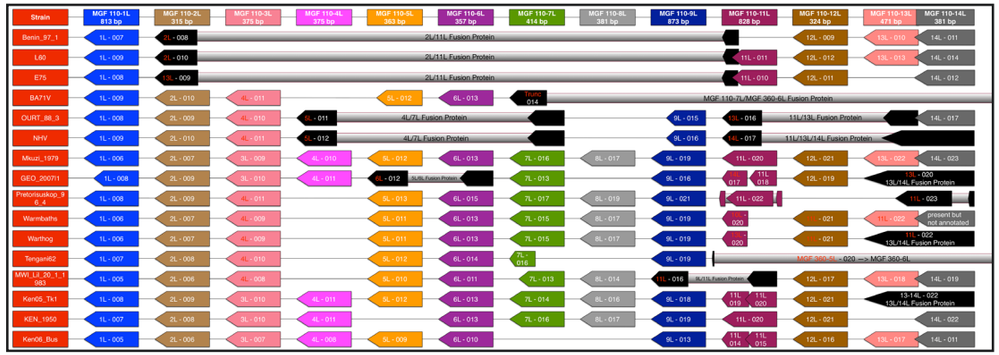
In conclusion, we believe that our figures are an intuitive visualization of the arrangement of genes within the MGFs, especially when there is a need to compare the MGFs of different viruses. It is envisioned that the maps of the ASFV MGFs will be living documents, updated, by a volunteer curator from the research community, with any new paralogs that may be discovered in newly sequenced genomes. This is likely to be required since genomes have yet to be sequenced from a large proportion of the ASFV genotypes and the MGFs are the most variable parts of the genomes. To this end, the diagrams can be easily updated when new MGF paralogs are discovered by the addition of new columns. Although new ASFV genomes can be added to the diagrams by simply copying the most similar existing row and editing the sizes of the gene boxes and labels, as the number of genomes grows, space could be saved by showing a single representative if multiple viruses have identical MGFs.

These MGF maps will speed up the annotation process and simplify the comparison of new ASFV genomes. With more accurate genome alignments, researchers will also be better able to correlate genomic features with virulence levels of the various ASFV isolates.

SUPPLEMENTAL FIGURES



Supp. Figure 1. Diagram of MGF 100 organization. See Supp. Figure 7 for additional information describing this MGF.

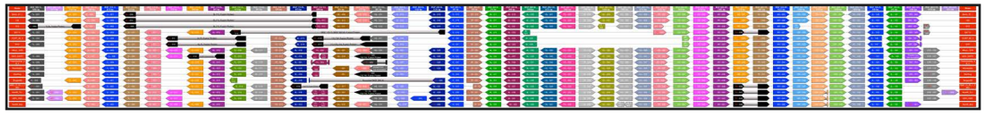


Supp. Figure 2. Diagram of MGF 110 organization. See Supp. Figure 8 for additional information describing this MGF.

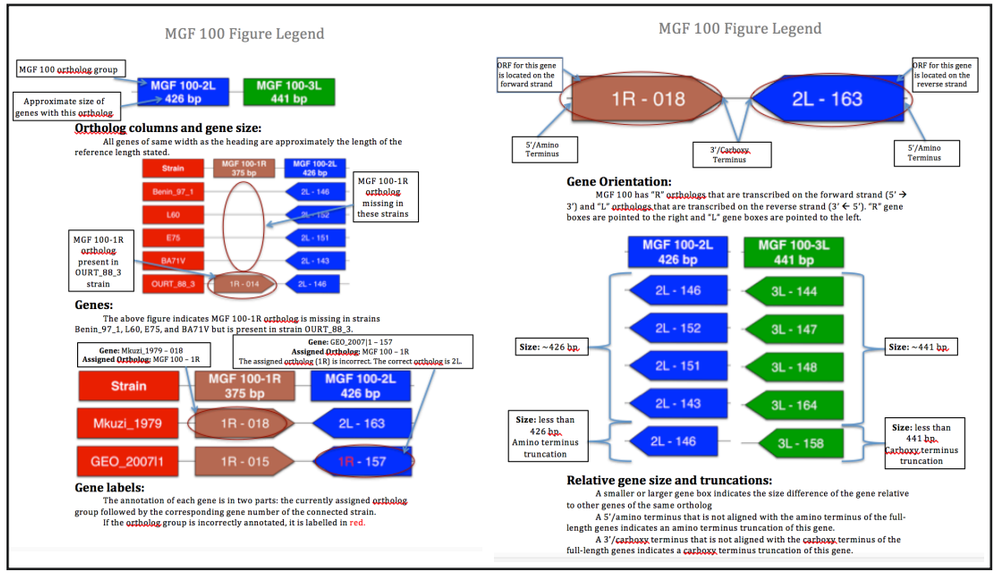
Supp. Figure 3. Diagram of MGF 300 organization. See Supp. Figure 9 for additional information describing this MGF.

Supp. Figure 4. Diagram of MGF 360 organization. See Supp. Figure 10 for additional information describing this MGF.

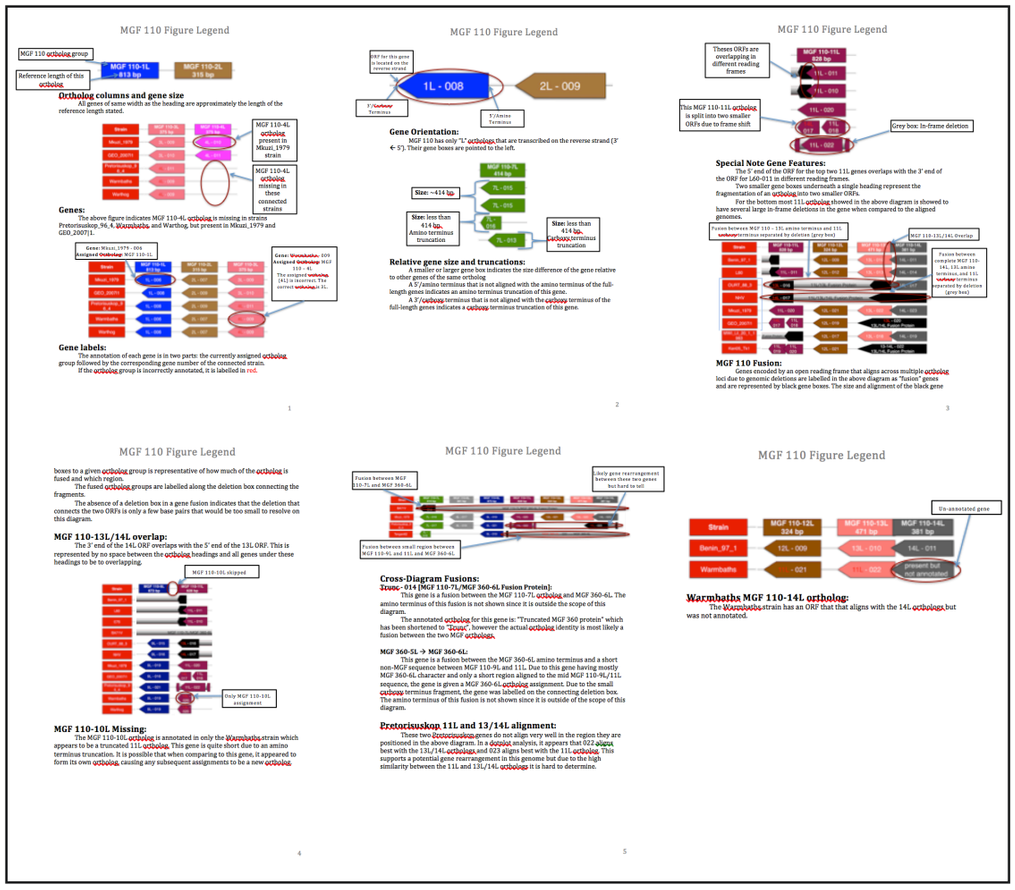
Supp. Figure 5. Diagram of MGF 505 organization. See Supp. Figure 11 for additional information describing this MGF.



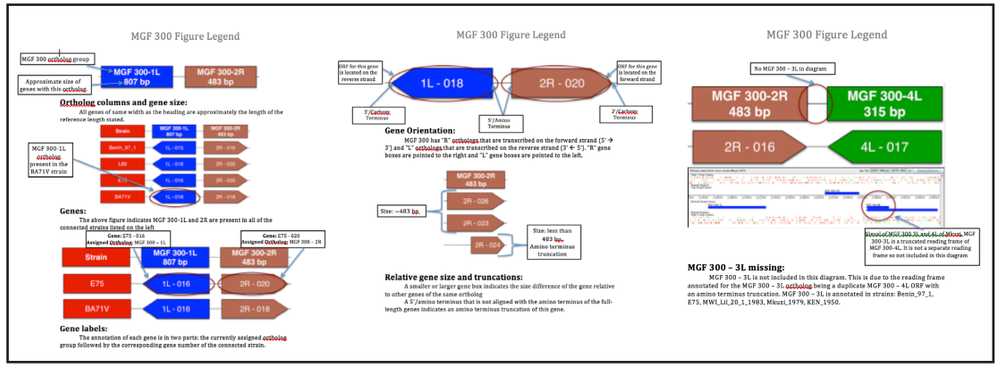
Supp. Figure 6. Diagram of MGF compilation. See Supp. Figure 12 for additional information describing this MGF compilation.



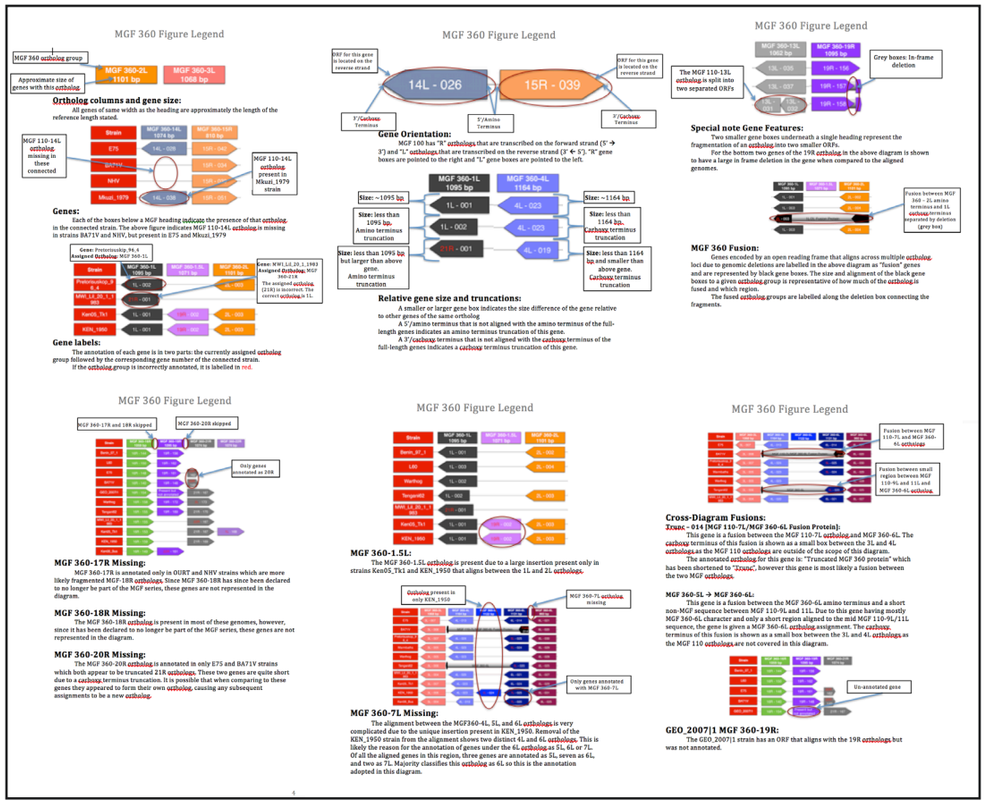
Supp. Figure 7. Information sheet describing MGF 100.



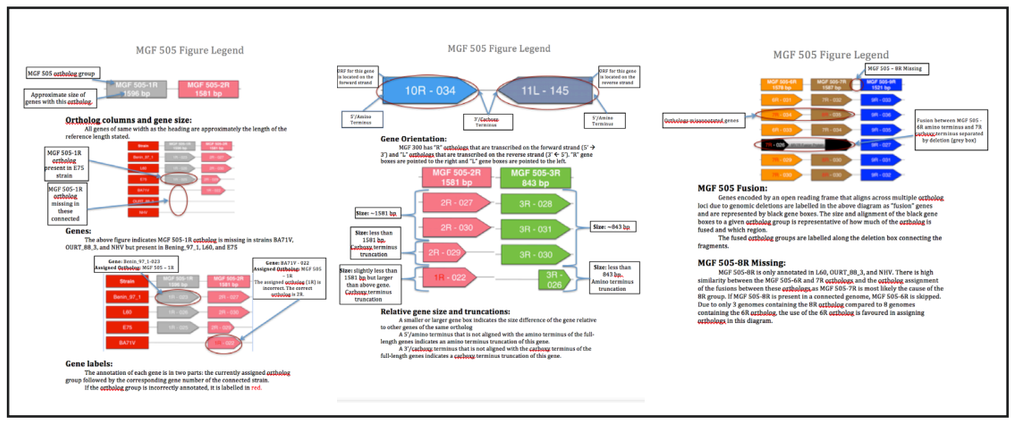
Supp. Figure 8. Information sheet describing MGF 110



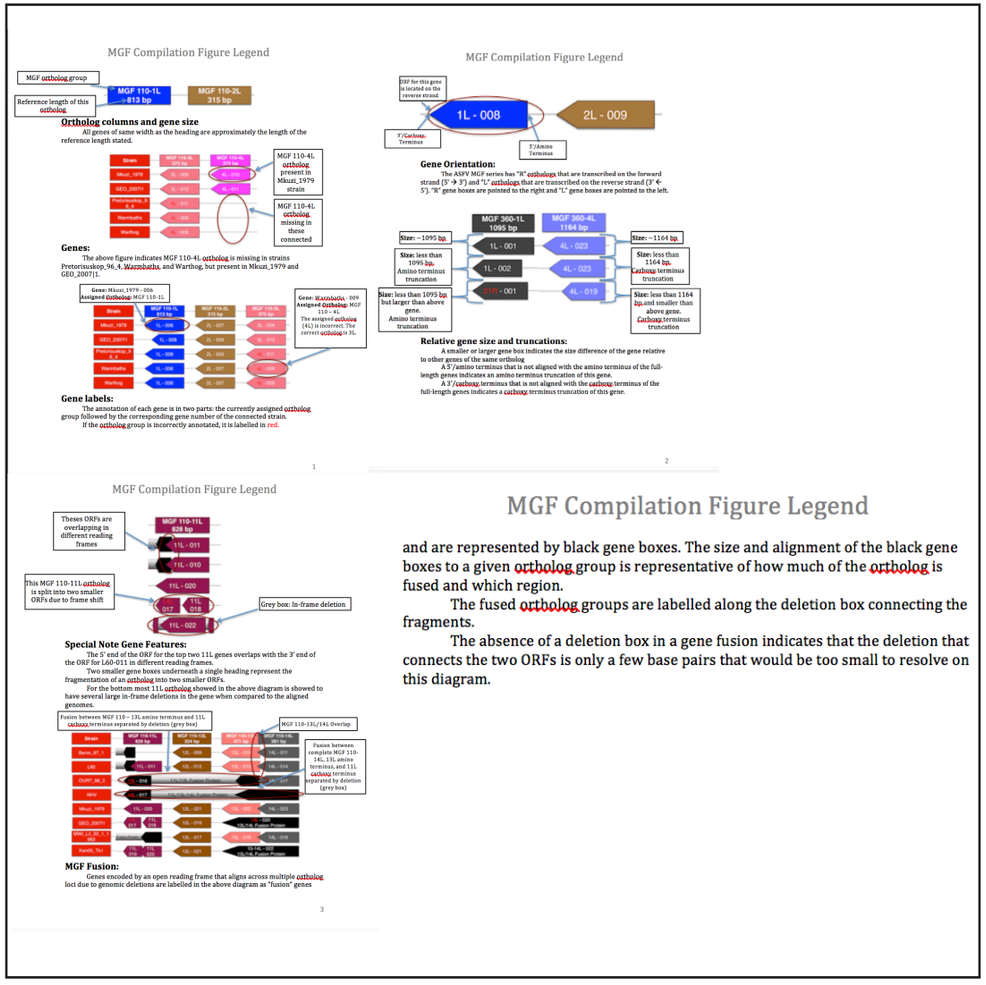
Supp. Figure 9. Information sheet describing MGF 300.



Supp. Figure 10. Information sheet describing MGF 360.



Supp. Figure 11. Information sheet describing MGF 505.



Supp. Figure 12. Information sheet describing MGF compilation.

ACKNOWLEDGMENTS

We would like to thank Chad Smithson for his help and the many Co-op students from the University of Victoria who have built the Viral Bioinformatics Resource Center (virology.uvic.ca).

REFERENCES

1. Achenbach JE, Gallardo C, Nieto-Pelegrín E, et al. 2016. Identification of a New Genotype of African Swine Fever Virus in Domestic Pigs from Ethiopia. *Transbound Emerg Dis* doi:10.1111/tbed.12511

2. Bellini S, Rutili D, Guberti V. 2016. Preventive measures aimed at minimizing the risk of African swine fever virus spread in pig farming systems. *Acta Vet Scand* 58:191.

3. Brodie R, Roper RL, Upton C. 2004. JDotter: a Java interface to multiple dotplots generated by dotter. *Bioinformatics* 20:279–81.

4. Chapman DAG, Tcherepanov V, Upton C, Dixon LK. 2008. Comparison of the genome sequences of non-pathogenic and pathogenic African swine fever virus isolates. *J. Gen. Virol* 89:397–408.

5. Cisek AA, Dąbrowska I, Gregorczyk KP, Wyżewski Z. 2016. African Swine Fever Virus: a new old enemy of Europe. *Ann Parasitol* 62:161–67.

6. Golding JP, Goatley L, Goodbourn S, Dixon LK, Taylor G, Netherton CL. 2016. Sensitivity of African swine fever virus to type I interferon is linked to genes within multigene families 360 and 505. *Virology* 493:154–61.

7. Guinat C, Gogin A, Blome S, Keil G, Pollin R, et al. 2016. Transmission routes of African swine fever virus to domestic pigs: current knowledge and future research directions. *Veterinary Record* 178:262–67.

8. Hillary W, Lin S-H, Upton C. 2011. Base-By-Base version 2: single nucleotide-level analysis of whole viral genome alignments. *Microb Inform Exp* 1:2.

9. Jori F, Vial L, Penrith ML, Pérez-Sánchez R, Etter E, et al. 2013. Review of the sylvatic cycle of African swine fever in sub-Saharan Africa and the Indian ocean. – PubMed – NCBI. *Virus Res* 173:212–27.

10. Kazutaka K, Standley DM. 2013. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol. Biol. Evol* 30:772–80.

11. Kumar S, Stecher G, Tamura K. 2016. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Mol. Biol. Evol* 33:1870–74.

12. Malogolovkin A, Burmakina G, Titov I, Sereda A, Gogin A, et al. 2015. Comparative analysis of African swine fever virus genotypes and serogroups. *Emerg Infect Dis* 21:312–15.

13. Muangkram Y, Sukmak M, Wajjwalku W. 2015. Phylogeographic analysis of African swine fever virus based on the p72 gene sequence. – PubMed – NCBI. *Genet. Mol Res* 14:4566–74.

14. O'Donnell V, Holinka LG, Gladue DP, Sanford B, Krug PW, et al. 2015. African Swine Fever Virus Georgia Isolate Harboring Deletions of MGF360 and MGF505 Genes Is Attenuated in Swine and Confers Protection against Challenge with Virulent Parental Virus. *J Virol* 89:6048–56.

15. Rock DL. 2016. Challenges for African swine fever vaccine development—"... perhaps the end of the beginning.." *Vet. Microbiol*

16. Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30:1312–13.

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

18. Tulman ER, Delhon GA, Ku BK, Rock DL. 2009. African swine fever virus. *Curr Top Microbiol Immunol* 328:43–87.

19. Zakaryan H, Revilla Y. 2016. African swine fever virus: current state and future perspectives in vaccine and antiviral research. *Vet. Microbiol* 185:15–19.