ASFVdb: An integrative resource for genomics and proteomics analyses of African swine fever

Zhenglin Zhu[1*], Geng Meng[2*]

1. The School of Life Sciences, Chongqing University, Chongqing, China

2. Laboratory of Biomedical research and College of Veterinary Medicine, China Agricultural University, Beijing, China

[*]Corresponding authors

Zhenglin Zhu, The School of Life Sciences, Chongqing University, Chongqing, 400044 China

TEL: (86)23-6512-2686, FAX: (86)23-6512-2689, zhuzl@cqu.edu.cn

Geng Meng, Laboratory of Biomedical research and College of Veterinary Medicine, China Agricultural University, Beijing, 100094 China

TEL: (86)10-6273-3466, FAX: (86)10-6273-3466, mg@cau.edu.cn

Keywords: African swine fever virus, population genomics, proteomics, database

Abstract

The recent outbreaks of African swine fever (ASF) in China and Europe have threatened the swine industry globally. To control the transmission of the African swine fever virus (ASFV), we developed ASFVdb, the African swine fever virus database, an online data visualization and analysis platform for comparative genomics and proteomics. On the basis of known ASFV genes, ASFVdb reannotates the genomes of every strain and annotates 4833 possible ORFs. Moreover, ASFVdb performs a thorough analysis of the population genetics of all the published genomes of ASFV strains and performs functional and structural predictions for all genes. For each ASFV gene, visitors can obtain not only basic information of the gene but also the distribution of the gene in strains, conserved or high mutation regions, possible subcellular location of the gene and topology of the gene. In the genome browser, ASFVdb provides sliding window population genetics analysis results, which facilitate genetics and evolutional analyses at the genomic level. The web interface is constructed based on SWAV 1.0. ASFVdb is freely accessible at http://asfvdb.popgenetics.net.

KEYWORDS: African swine fever virus; population genomics; database

Introduction

African swine fever (ASF), caused by the ASF virus (ASFV), is a viral disease in domestic or wild swine (Parker et al., 1969; Thomson et al., 1980; Anderson et al., 1998). The mortality rate of ASF is nearly 100% for infected domestic swine. The outbreaks in China and central Europe in the summer of 2018 caused important economic losses in the global swine industry. However, efforts to produce effective vaccines or treatments for ASFV are obstructed by the complexity of the virus. The ASF virus, which has an icosahedral capsid with a multilayered membrane structure, contains many polypeptides, the function and identities of which are largely unknown. It has been reported that some generated antibodies are not protective but may enhance the disease (Detray, 1963; Stone and Hess, 1967; Mebus, 1988; Blome et al., 2014; Sunwoo et al., 2019) due to the complexity of the virion and the intracellular and extracellular localization of infectious particles. Consequently, identification of functional ASFV polypeptides is important to develop effective vaccines.

Due to the above observations, we developed ASFVdb, the African swine fever database, to assist ASFV research. Different from previous relevant ASFV databases, such as ViralZone (Hulo et al., 2011) , ASFVdb is specific for ASFV. ASFVdb compares all the published ASFV genomes (Chapman et al., 2008; de Villiers et al., 2010; Chapman et al., 2011; Bishop et al., 2015; Portugal et al., 2015; Olesen et al., 2018; Zani et al., 2018; Bao et al., 2019), annotates each genome and validates the persistence of each gene in different strains. ASFVdb was developed to facilitate the identification of gene function and identity. Aside from a basic annotation, it also provides information on the subcellular location of genes and functions of genes, topology predictions, comparative genomic alignments and population genetics analysis results.

Result and discussions

Overview of the genomic data in ASFVdb

ASFVdb collects a broad range of published ASFV genome data (Table S1) and analyzes and identifies 253 ASFV gene clusters (Table S2) from the genomes of 38 strains (for details, see Material and Methods). Among the 253 gene clusters, 145 have conserved ORFs in all strains (Table S2). Reannotation of the genomes of the 38 strains predicted 4833 genes and 924 genetic remains (Figure 1). In each strain, there are 181-263 genes, on average. These results did not contradict previous reports of 151-181 genes per strain (Galindo and Alonso, 2017; Bao et al., 2019), considering that the pipeline included all possible alternative splices (Figure S1) and the discrepancies between annotations for different strains with different strategies at different times (Chapman et al., 2008; de Villiers et al., 2010; Chapman et al., 2011; Bishop et al., 2015; Portugal et al., 2015; Olesen et al., 2018; Zani et al., 2018; Bao et al., 2019). We performed a subcellular localization analysis of the ASFV genes to determine their roles in the infection process (Figure 2A). We found that more than half of the ASFV (60%) are single-pass or multi-pass membrane proteins (Figure 2B). 8% of the ASV genes (23 cases) are capsid proteins, which may provide potential targets for vaccine design. In gene ontology, the largest number of ASFV genes are enriched in the membrane and had catalytic activity (Figure 3). Moreover, based on the population genetics test analysis, we found that 334 genes have a significantly low Tajima's D (Tajima, 1989) and significantly high composite likelihood ratio (CLR) (Nielsen et al., 2005; Zhu and Bustamante, 2005) (Rank Test, P-value<0.05), indicating that these genes were recently possibly under positive selection (Table S3). 77 genes have a significantly high Tajima's D value (Table S4, Rank Test, P-value<0.05), and they may be involved in balancing selection. These results are informative for future ASFV research.

ASFVdb offers a diverse search method

The main search input of ASFVdb, located at the center of the home page, allows

users to search a gene by its general gene name, gene accession or description of the gene's function. We created a gene alias list to make the search results as complete as possible. In the right column of the home page, there is a list of other gene search methods in ASFVdb. First, users can BLAT against one ASFV genome, like the workflow in the UCSC genome browser (Karolchik et al., 2007), or BLAST against CDS (coding sequence) or protein sequences. Second, users can search for genes in a specific subcellular location. In the results view, we list the topology information and function annotation of each gene to allow users to conveniently localize the gene in the host-virus interaction. Third, users can find genes that have a specific function according to the gene's GO annotation. "Gene Clusters" lists the persistence of the ASFV genes in the 38 strains. Users can view a specific gene by clicking an item in the gene list table. Moreover, to more conveniently track a list of genes, ASFVdb provides gene links by inputting a list of the genomic positions or a list of the accession numbers of genes. These operations facilitate analysis of ASFVdb with a personal gene list.

Evolutionary episodes of the ASFV strains in ASFVdb

On the home page, below the main search box, a phylogenetic tree displays the evolutionary history of the available published 38 ASFV strains. In contrast to the previous method used to make ASFV phylogenetic trees (de Villiers et al., 2010; Bao et al., 2019), we built the tree by performing multiple alignments on the genomic scale to better understand the processes and patterns of evolution (Wilgenbusch et al., 2017). The tree we constructed (Figure 4) is consistent with previous publications (Galindo and Alonso, 2017; Bao et al., 2019; Sanchez et al., 2019). Our tree reflects transmission events during the spread of ASFV. After it was first identified in Kenya in the 1920s (Montgomery, 1921), ASFV was subsequently reported in eastern and southern Africa (Thomson G. R., 2004). In the mid-last century, ASFV spread from Africa to Portugal (Sanchez-Botija, 1963) as its first stop in Europe. Then, ASFV spread to Malta (1978), Italy (1967, 1980), France (1964, 1967, 1977), Belgium (1985) and the Netherlands in 1986 (Costard et al., 2009). In 2007, ASFV was introduced to

Caucasus (Rowlands et al., 2008) and began to spread across Russia and Eastern Europe. In 2018, ASFV was found in China, and outbreaks in China and central Europe enhance the seriousness of this disease (OIE, 2019).

Gene Visualization in ASFVdb

Through clicking the taxon name in the tree or selecting an item in the "Gbrwoser" list, users can go to the genome browser page (Figure 5), where gene segments are subsequently arranged along the genome according to the genomic positions of genes. Genes annotated from the NCBI GFF file are colored in blood red as "Annotation from NCBI". Genes annotated by mapping (for details, see the Material and Methods) are colored in dark red as "Newly annotated". "Genetic Remains" are colored in light gray. The gene name is provided above each gene segment, and ">" or "<" indicates a gene's transcriptional direction. Four tracks of population genetics analysis, Pi, Theta (Watterson, 1975), Tajima's D (Tajima, 1989), and CLR (Nielsen et al., 2005; Zhu and Bustamante, 2005) are listed following the gene track. To help trace selective signatures, we drew a top 5% line for CLR and two 5% lines in ascending and descending order for Pi, Theta and Tajima's D. Using the tool bar at the top, users can move or zoom the genome in the browser, set the focus bar to a selected region, export the drawing of one track with gene segments and export the data of one track.

Clicking one gene segment will lead to the gene's information page. The full list of annotations includes basic information (strain, gene name, description, location in the genome, Genbank Accession and full name), sequence (CDS and protein), summary (function, UniProt accession, related Pubmed ID, related EMBL ID, corresponding Proteomes ID, related Pfam ID and correlated Interpro ID), ontologies (GO and KEGG), subcellular location, topology (transmembrane region prediction), genomic alignment in the CDS region, multiple alignment of orthologues, gene tree of corresponding NCBI annotated or newly annotated proteins, and orthologous genes in strains. Texts that link to internal or external sites are hyperlinked to facilitate viewing and analysis.

Analysis of CD2v (EP402R) in ASFVdb: a case study

CD2v (EP402R) is an ASFV gene related to hemadsorption (Rodriguez et al., 1993). CD2v (EP402R) plays a key role in viral attenuation and may be involved in the generation of immune protection from the host against ASFV infection (Monteagudo et al., 2017). Through searching for "EP402R" in the strain China_AnhuiXCGQ_2018, we found that this gene encodes a single-Pass membrane protein in the transmembrane region from 207 AA to 229 AA (Figure S2A). In the genome view with the population genetics analysis tracks, we found that the transmembrane region is composed of a Tajima's D valley and a CLR peak (Figure S2B), inferring the selection signatures. The region from 1AA to 206AA is predicted to be located outside of the viron membrane, with peaks of Pi and Theta, indicating its high diversity among strains. From the information regarding "Genomic alignment in the CDS region" and "Orthologues in Strains", we found that this gene is only highly conserved among 27 of all 38 compared strains (Figure S2C). This result may partly explain why pigs vaccinated with baculovirus-expressed EP402R/CD2v did not have full protection from ASFV infection (Ruiz-Gonzalvo et al., 1996; Escribano et al., 2013). AFSVdb annotates that this protein has similarity (E-value=0.000801919) to the cell adhesion molecule CD2 (Jones et al., 1992), the crystal structure of which (Figure S2D, from PDB) was used to build a homology modelling structure of EP402R.

Conclusion

To assist veterinary scientists and researchers to combat the threat of ASFV outbreaks, we developed a database-based toolbox, ASFVdb. ASFVdb integrates data from NCBI, UniProt (Apweiler et al., 2004), ViralZone (Hulo et al., 2011) and a broad range of published articles (Hulo et al., 2011). ASFVdb specializes in ASFV. ASFVdb compares all published ASFV genomes (Chapman et al., 2008; de Villiers et al., 2010; Chapman et al., 2011; Bishop et al., 2015; Portugal et al., 2015; Olesen et al., 2018; Zani et al., 2018; Bao et al., 2019), summarizes data on the proteins of all known ASFV strains, and provides not only basic information of genes but also external links, subcellular localization, topology, comparative genomic alignments, a gene tree and evolutionary analysis results. This information is helpful and convenient for the further accession of proteomics and biochemical data in ASFV research. Hopefully, ASFVdb will be useful for further research on ASFV and vaccine design.

Materials and methods

Data collection and processing

We downloaded the genome sequences, CDS, proteins and relevant annotations of 38 sequenced ASFV strains (Table S1) from the NCBI genome database. Among the 3 different versions (NC_001659.2, U18466.2 and KP055815.1) of Spain_BA71_1971 genome data, we used the original version, KP055815.1. Of the 38 different ASFV genomic data, only 23 are fully annotated, and their gene IDs or names are not unified. To harmonize this inconsistency, we constructed a unified dataset and used the dataset to reannotate the 38 genomes. Specifically, we performed pairwise alignments for all of the annotated protein sequences. The sequences were clustered together with an identity > 50% and coverage > 80% according to previous homologous gene identification methods (She et al., 2009; Yue et al., 2018). Finally, a list containing 308 unified genes was generated. Then, we mapped these unified genes to all 38 genomes. If a unique protein mapped with the highest score of identity > 50% and coverage > 80%, we considered the protein to be a unique protein in the strain, and the mapped region in the strain genome was marked. To identify the corresponding mapped region of one unique gene that might express a protein, we translated the mapped region using NCBI ORFfinder and aligned all the possible translations of the unique protein. We took the alignment with the highest score and checked whether its identity and coverage were both higher than 0.5. If its identity and coverage were both higher than 0.5, we consider the region to express proteins; if not, we marked the region as “Genetic Remains”. The genes that were annotated in the NCBI genome database were marked as “Annotation from NCBI”, and unannotated gene regions with a high expression probability were marked as “Newly Annotated”. Finally, all of the sequenced genes were clustered into 253 gene clusters with CD-HIT (Li and Godzik, 2006) and required identification > 50% and coverage > 70%.

To retrieve the annotation of the ASFV gene product from UniProt (Apweiler et al.,

2004; Patient et al., 2008), we BLASTed the protein sequence of the ASFV gene against the UniProt protein database (Patient et al., 2008) and took the hits with the highest scores as the best matches. If the best match of one ASFV gene had an E-value <0.05, we extracted the accession number of the mapped sequence as the UniProt ID of the matched ASFV gene. With the UniProt ID, we obtained the external annotations of the ASFV gene, such as the corresponding PubMed ID, EMBL ID, Proteomes ID, Pfam, Interpro, Gene Ontology, and KEGG and function annotations. We also BLASTed all of the ASFV protein sequences against the PDB database (Berman et al., 2000; Burley et al., 2019) in the same way and recorded the alignment results. Disappointedly, we only obtained 45 unique proteins from the PDB that had an E-value <0.05, which means that most of the ASFV-expressed proteins did not have structural information.

Subcellular localization and topology prediction

The subcellular location and topology annotations from UniProt only cover several ASFV genes. Thus, we reperformed these predictions for all of the genes. We predicted the subcellular location of the ASFV proteins through MSLVP (Thakur et al., 2016) using the parameters of “one-versus-one”, “Second-tier Algorithm” and a similarity >90%. A total of 253 gene clusters were grouped into 9 subcellular locations. We predicted transmembrane helixes within the protein sequences using TMHMM 2.0 (Krogh et al., 2001). The output images were converted into PNG format to be displayed on the website by Magick (www.imagemagick.org).

Comparative genomics and population genetics analysis

To trace the evolutionary history of ASFV, we performed a whole genome alignment of all 38 ASFV genomes by Mugsy (Angiuoli and Salzberg, 2011) and built a phylogenetic tree by FastTree 2.1 (Price et al., 2010) , with the parameter ‘-boot 5000’ to test the likelihood of the generated tree. To obtain a more detailed comparative genomic map of the 38 strains, we used LASTZ (Harris, 2007) to perform

genome-genome alignments between any two ASFV strains and outputted the results in AXT format. For one ASFV gene, we retrieved the corresponding sequences of other strains from the genome-genome alignment results, realigned these sequences by MUSCLE (Edgar, 2004b; Edgar, 2004a), and obtained the genomic alignment of the CDS region. According to the persistence of each gene in different strains, we also retrieved all of the protein counterparts of one unique gene and performed a multiple alignment by MUSCLE. To more clearly display the evolutionary history of each protein, we also drew trees from the multiple alignment as we did for the genome.

We used a window of 200 bp and a step size of 50 bp to slide along the ASFV genome. From the genome-genome alignments described above, we retrieved sequences within the sliding window that we used to calculated Pi, Theta (Watterson, 1975) and Tajima's D (Tajima, 1989) using VariScan 2.0 (Vilella et al., 2005; Hutter et al., 2006) . We wrote Perl scripts to call the allele frequencies and used SweapFineder2 (DeGiorgio et al., 2016) to calculate the composite likelihood ratio (CLR) (Nielsen et al., 2005; Zhu and Bustamante, 2005) with a step size of 50. We took the median of the population genetics test statistics in the region of each gene as the corresponding value.

Web Interfaces

The web interface of ASFVdb was developed using Mysql + PHP + CodeIgniter (www.codeigniter.com) + JQuery (jquery.com) and was based on SWAV (swav.popgenetics.org), which is an open source web application for population genetics visualization and analysis. We used (Yachdav et al., 2016) to show the multiple alignments of orthologous CDS or proteins and phylotree (Shank et al., 2018) to display the phylogenetic trees of genomes or proteins. We made changes to the source code of the web application to fit the development of ASFVdb, such as adding links within the diagram. The search interfaces of the sequence alignments were constructed based on PHP parsing of the results from BLAT (Kent, 2002) and NCBI

BLAST (Johnson et al., 2008). The workflow of the other search blocks, such as searching by the molecular structure or gene names, are mostly SQL query pipelines.

Data availability

All ASFVdb data are publicly and freely accessible at http://asfvdb.popgenetics.net. Feedback on any aspect of the ASFVdb database and discussions of ASFV gene annotations are welcome by email to zhuzl@cqu.edu.cn or mg@cau.edu.cn.

Acknowledgments

This work was supported by grants from the National Natural Science Foundation of China (31200941) and the Fundamental Research Funds for the Central Universities (106112016CDJXY290002).

Figure Legends

Figure 1. Number of annotated genes in all the ASFV strains. Genes annotated in NCBI, genes newly annotated and genetic remains are marked in dark red, red and gray, respectively.

Figure 2. A shows a diagram of the infection process of ASFV with the MSLVP subcellular locations marked. B shows the percentages of ASFV gene clusters in each subcellular location group.

Figure 3. Enrichment of the gene ontology (GO) of ASFV gene clusters. This figure was drawn by WEGO (Ye et al., 2006).

Figure 4. A, Phylogenetic tree of the ASFV strains in ASFVdb with historical events marked. The numbers marked in red are the marginal likelihood of the tree. B. The distribution of strains in ASFVdb according to country. The blue arrows with dotted lines indicate the possible infection paths of ASFV.

Figure 5. Gnome browser view to display ASFV genes and population genetics test statistics.

References

Anderson EC, Hutchings GH, Mukarati N, Wilkinson PJ. 1998. African swine fever virus infection of the bushpig (Potamochoerus porcus) and its significance in the epidemiology of the disease. *Vet Microbiol* 62:1-15.

Angiuoli SV, Salzberg SL. 2011. Mugsy: fast multiple alignment of closely related whole genomes. *Bioinformatics* 27:334-342.

Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M et al. 2004. UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res* 32:D115-119.

Bao J, Wang Q, Lin P, Liu C, Li L, Wu X, Chi T, Xu T, Ge S, Liu Y et al. 2019. Genome comparison of African swine fever virus China/2018/AnhuiXCGQ strain and related European p72 Genotype II strains. *Transbound Emerg Dis*.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. 2000. The Protein Data Bank. *Nucleic Acids Res* 28:235-242.

Bishop RP, Fleischauer C, de Villiers EP, Okoth EA, Arias M, Gallardo C, Upton C. 2015. Comparative analysis of the complete genome sequences of Kenyan African swine fever virus isolates within p72 genotypes IX and X. *Virus Genes* 50:303-309.

Blome S, Gabriel C, Beer M. 2014. Modern adjuvants do not enhance the efficacy of an inactivated African swine fever virus vaccine preparation. *Vaccine* 32:3879-3882.

Burley SK, Berman HM, Bhikadiya C, Bi C, Chen L, Di Costanzo L, Christie C, Dalenberg K, Duarte JM, Dutta S et al. 2019. RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. *Nucleic Acids Res* 47:D464-D474.

Chapman DA, Darby AC, Da Silva M, Upton C, Radford AD, Dixon LK. 2011. Genomic analysis of highly virulent Georgia 2007/1 isolate of African swine fever virus. *Emerg Infect Dis* 17:599-605.

Chapman DA, 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.

Costard S, Wieland B, de Glanville W, Jori F, Rowlands R, Vosloo W, Roger F, Pfeiffer DU, Dixon LK. 2009. African swine fever: how can global spread be prevented? *Philos Trans R Soc Lond B Biol Sci* 364:2683-2696.

de Villiers EP, Gallardo C, Arias M, da Silva M, Upton C, Martin R, Bishop RP. 2010. Phylogenomic analysis of 11 complete African swine fever virus genome sequences. *Virology* 400:128-136.

DeGiorgio M, Huber CD, Hubisz MJ, Hellmann I, Nielsen R. 2016. SweepFinder2: increased sensitivity, robustness and flexibility. *Bioinformatics* 32:1895-1897.

Detray DE. 1963. African Swine Fever. *Adv Vet Sci* 8:299-333.

Edgar RC. 2004a. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32:1792-1797.

Edgar RC. 2004b. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5:113.

Escribano JM, Galindo I, Alonso C. 2013. Antibody-mediated neutralization of African swine fever virus: myths and facts. *Virus Res* 173:101-109.

Galindo I, Alonso C. 2017. African Swine Fever Virus: A Review. *Viruses* 9.

Harris RS. 2007. Improved pairwise alignment of genomic DNA. *Ph.D. thesis, Pennsylvania State University*.

Hulo C, de Castro E, Masson P, Bougueleret L, Bairoch A, Xenarios I, Le Mercier P. 2011. ViralZone: a knowledge resource to understand virus diversity. *Nucleic Acids Res* 39:D576-582.

Hutter S, Vilella AJ, Rozas J. 2006. Genome-wide DNA polymorphism analyses using VariScan. *BMC Bioinformatics* 7:409.

Johnson M, Zaretskaya I, Raytselis Y, Merezhuk Y, McGinnis S, Madden TL. 2008. NCBI BLAST: a better web interface. *Nucleic Acids Res* 36:W5-9.

Jones EY, Davis SJ, Williams AF, Harlos K, Stuart DI. 1992. Crystal structure at 2.8 A resolution of a soluble form of the cell adhesion molecule CD2. *Nature* 360:232-239.

Karolchik D, Hinrichs AS, Kent WJ. 2007. The UCSC Genome Browser. *Curr Protoc Bioinformatics* Chapter 1:Unit 1 4.

Kent WJ. 2002. BLAT--the BLAST-like alignment tool. *Genome Res* 12:656-664.

Krogh A, Larsson B, von Heijne G, Sonnhammer EL. 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol* 305:567-580.

Li W, Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22:1658-1659.

Mebus CA. 1988. African swine fever. *Adv Virus Res* 35:251-269.

Monteagudo PL, Lacasta A, Lopez E, Bosch L, Collado J, Pina-Pedrero S, Correa-Fiz F, Accensi F, Navas MJ, Vidal E et al. 2017. BA71DeltaCD2: a New Recombinant Live Attenuated African Swine Fever Virus with Cross-Protective Capabilities. *J Virol* 91.

Montgomery RE. 1921. On a form of swine fever occurring in british east Africa (Kenya Colony). *J. Comp. Pathol. Ther* 1921:159-191.

Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG, Bustamante C. 2005. Genomic scans for selective sweeps using SNP data. *Genome Res* 15:1566-1575.

OIE. 2019. African Swine Fever (ASF). *Word Animal Health Information and Analysis Department. Report Nº11* February 15-March 01.

Olesen AS, Lohse L, Dalgaard MD, Wozniakowski G, Belsham GJ, Botner A, Rasmussen TB. 2018. Complete genome sequence of an African swine fever virus (ASFV POL/2015/Podlaskie) determined directly from pig erythrocyte-associated nucleic acid. *J Virol Methods* 261:14-16.

Parker J, Plowright W, Pierce MA. 1969. The epizootiology of African swine fever in Africa. *Vet Rec* 85:668-674.

Patient S, Wieser D, Kleen M, Kretschmann E, Jesus Martin M, Apweiler R. 2008. UniProtJAPI: a remote API for accessing UniProt data. *Bioinformatics* 24:1321-1322.

Portugal R, Coelho J, Hoper D, Little NS, Smithson C, Upton C, Martins C, Leitao A, Keil GM. 2015. Related strains of African swine fever virus with different virulence: genome comparison and analysis. *J Gen Virol* 96:408-419.

Price MN, Dehal PS, Arkin AP. 2010. FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One* 5:e9490.

Rodriguez JM, Yanez RJ, Almazan F, Vinuela E, Rodriguez JF. 1993. African swine fever virus encodes a CD2 homolog responsible for the adhesion of erythrocytes to infected cells. *J Virol* 67:5312-5320.

Rowlands RJ, Michaud V, Heath L, Hutchings G, Oura C, Vosloo W, Dwarka R, Onashvili T, Albina E,

Dixon LK. 2008. African swine fever virus isolate, Georgia, 2007. *Emerg Infect Dis* 14:1870-1874.

Ruiz-Gonzalvo F, Rodriguez F, Escribano JM. 1996. Functional and immunological properties of the baculovirus-expressed hemagglutinin of African swine fever virus. *Virology* 218:285-289.

Sanchez-Botija A. 1963. Reservorios del virus de la peste porcina Africana. Investigacion del virus la P.P.A en los artropodos mediante la prueba de la hemoadsorcion. Bull. *Bull. Off. Int. Epizoot.* 60:895-899.

Sanchez EG, Perez-Nunez D, Revilla Y. 2019. Development of vaccines against African swine fever virus. *Virus Res*.

Shank SD, Weaver S, Kosakovsky Pond SL. 2018. phylotree.js - a JavaScript library for application development and interactive data visualization in phylogenetics. *BMC Bioinformatics* 19:276.

She R, Chu JS, Wang K, Pei J, Chen N. 2009. GenBlastA: enabling BLAST to identify homologous gene sequences. *Genome Res* 19:143-149.

Stone SS, Hess WR. 1967. Antibody response to inactivated preparations of African swine fever virus in pigs. *Am J Vet Res* 28:475-481.

Sunwoo SY, Perez-Nunez D, Morozov I, Sanchez EG, Gaudreault NN, Trujillo JD, Mur L, Nogal M, Madden D, Urbaniak K et al. 2019. DNA-Protein Vaccination Strategy Does Not Protect from Challenge with African Swine Fever Virus Armenia 2007 Strain. *Vaccines (Basel)* 7.

Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123:585-595.

Thakur A, Rajput A, Kumar M. 2016. MSLVP: prediction of multiple subcellular localization of viral proteins using a support vector machine. *Mol Biosyst* 12:2572-2586.

Thomson G. R. BADS. 2004. African swine fever. In Infectious diseases of livestock. Oxford, UK: Oxford University Press.

Thomson GR, Gainaru MD, Van Dellen AF. 1980. Experimental infection of warthos (Phacochoerus aethiopicus) with African swine fever virus. *Onderstepoort J Vet Res* 47:19-22.

Vilella AJ, Blanco-Garcia A, Hutter S, Rozas J. 2005. VariScan: Analysis of evolutionary patterns from large-scale DNA sequence polymorphism data. *Bioinformatics* 21:2791-2793.

Watterson GA. 1975. On the number of segregating sites in genetical models without recombination. *Theor Popul Biol* 7:256-276.

Wilgenbusch JC, Huang W, Gallivan KA. 2017. Visualizing phylogenetic tree landscapes. *BMC Bioinformatics* 18:85.

Yachdav G, Wilzbach S, Rauscher B, Sheridan R, Sillitoe I, Procter J, Lewis SE, Rost B, Goldberg T. 2016. MSAViewer: interactive JavaScript visualization of multiple sequence alignments. *Bioinformatics* 32:3501-3503.

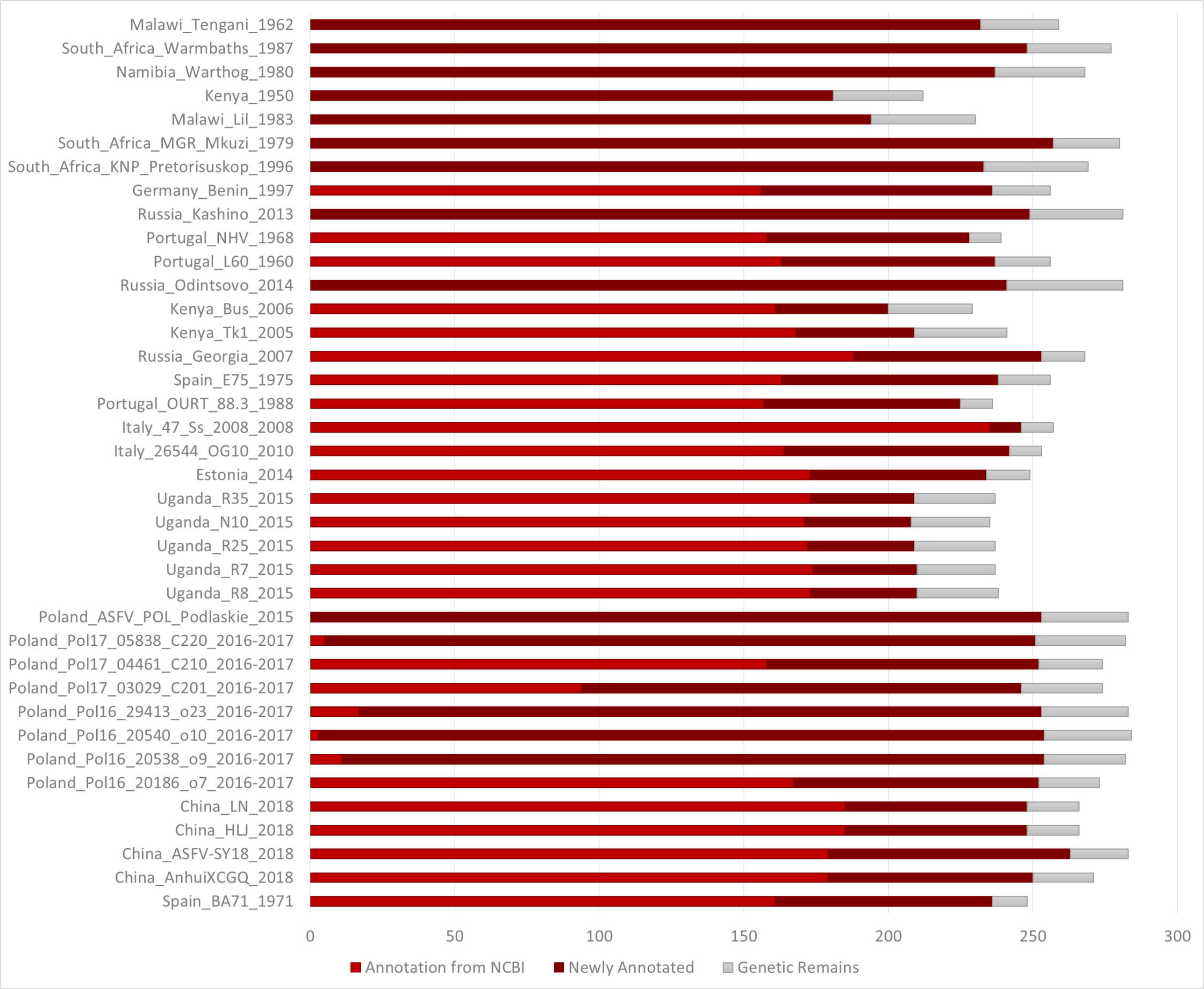
Ye J, Fang L, Zheng H, Zhang Y, Chen J, Zhang Z, Wang J, Li S, Li R, Bolund L. 2006. WEGO: a web tool for plotting GO annotations. *Nucleic Acids Res* 34:W293-297.

Yue H, Shu D, Wang M, Xing G, Zhan H, Du X, Song W, Nie X. 2018. Genome-Wide Identification and Expression Analysis of the HD-Zip Gene Family in Wheat (Triticum aestivum L.). *Genes (Basel)* 9.

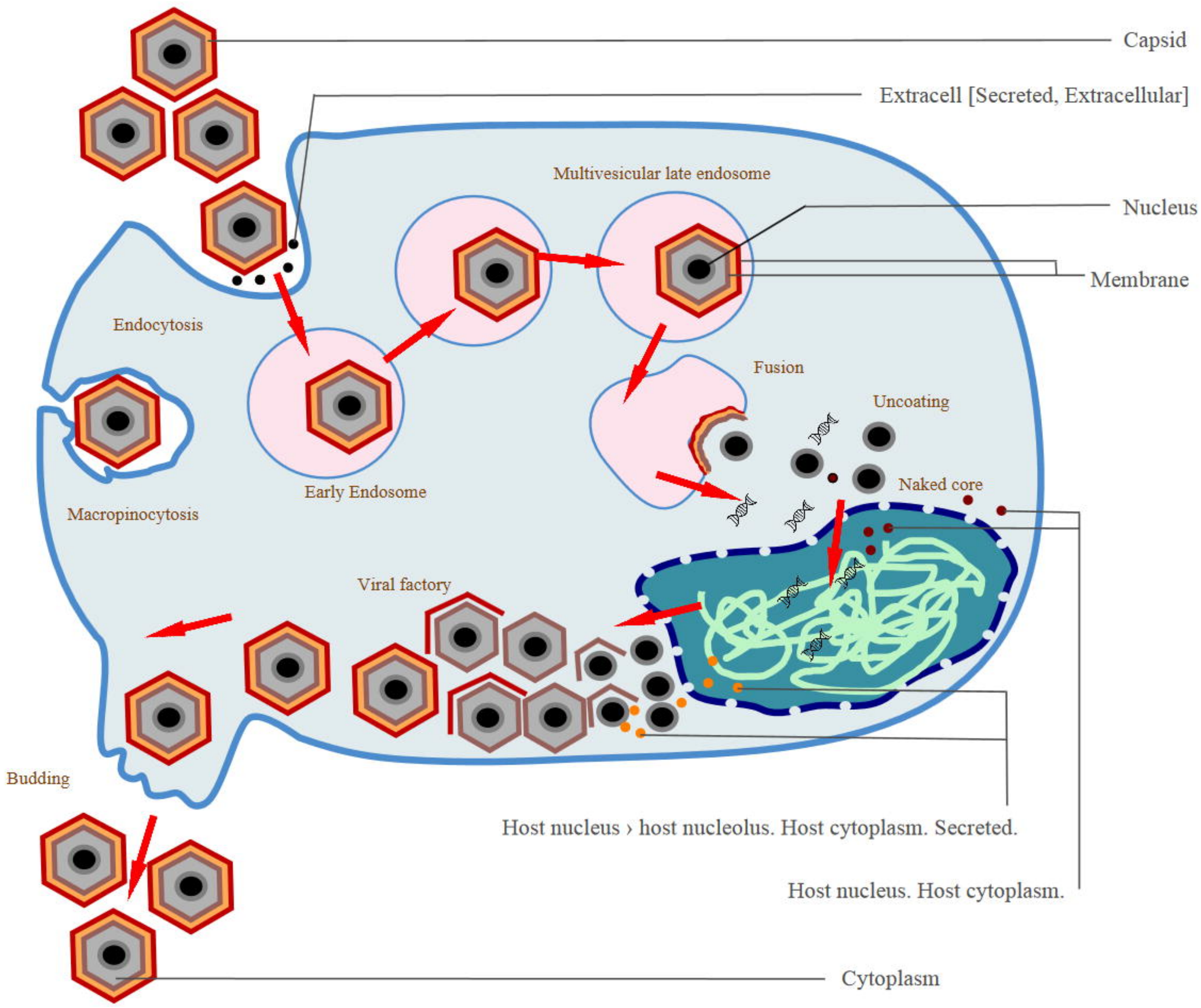
Zani L, Forth JH, Forth L, Nurmoja I, Leidenberger S, Henke J, Carlson J, Breidenstein C, Viltrop A, Hoper D et al. 2018. Deletion at the 5'-end of Estonian ASFV strains associated with an attenuated phenotype. *Sci Rep* 8:6510.

Zhu L, Bustamante CD. 2005. A composite-likelihood approach for detecting directional selection from

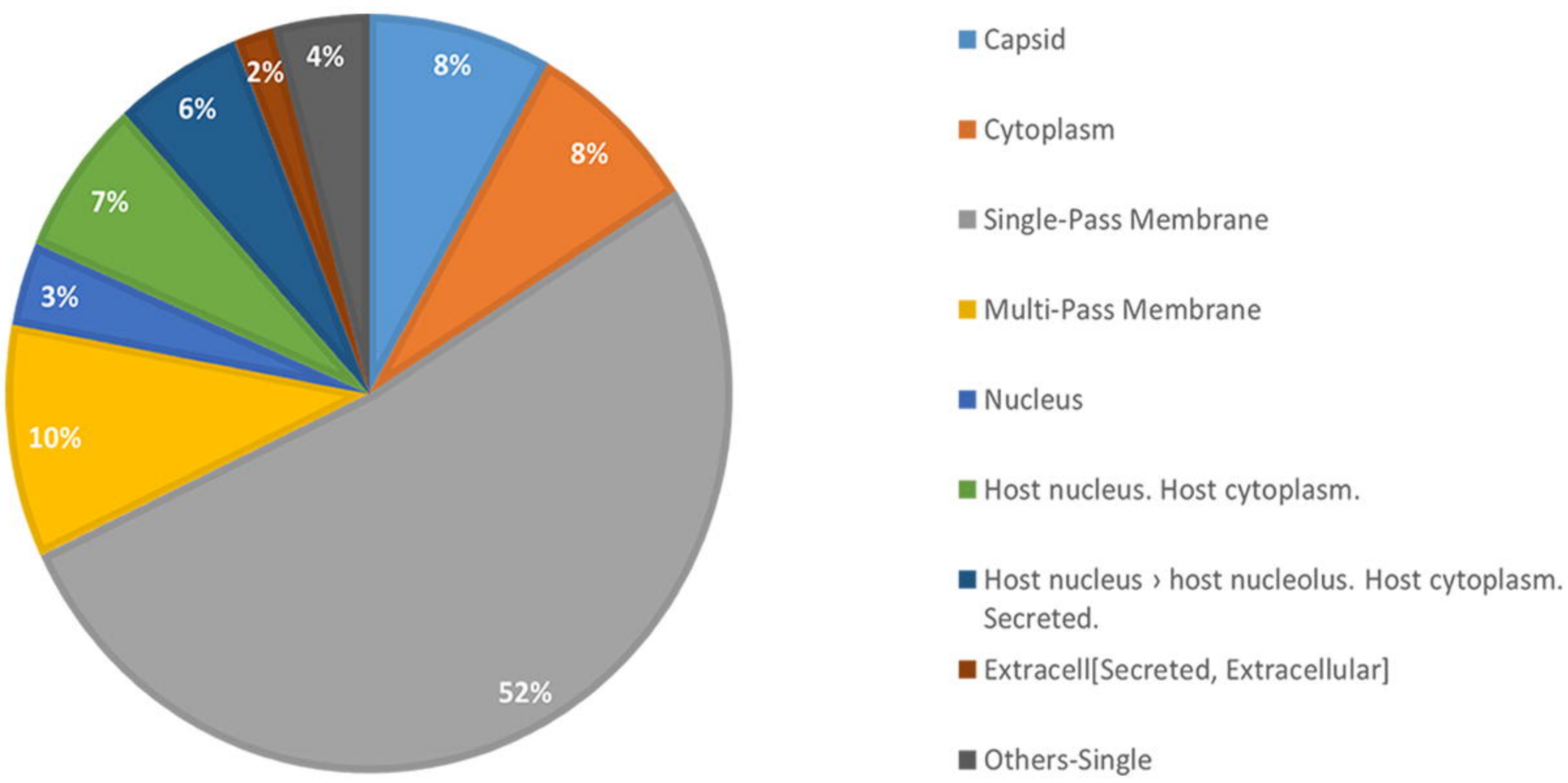
DNA sequence data. *Genetics* 170:1411-1421.

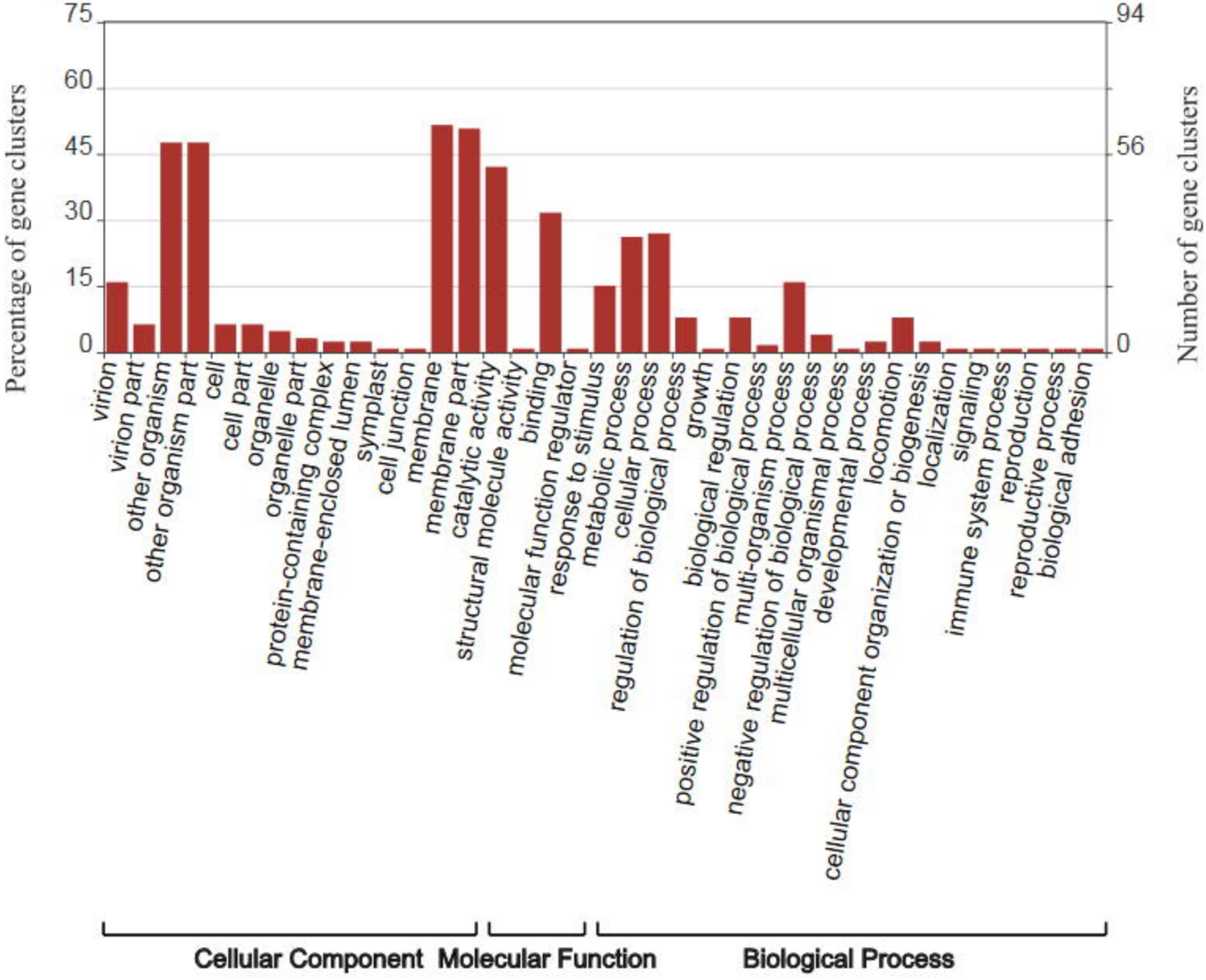

Malawi_Tengani_1962
South_Africa_Warmbaths_1987
Namibia_Warthog_1980
Kenya_1950
Malawi_Lil_1983
South_Africa_MGR_Mkuzi_1979
South_Africa_KNP_Pretorisuskop_1996
Germany_Benin_1997
Russia_Kashino_2013
Portugal_NHV_1968
Portugal_L60_1960
Russia_Odintsovo_2014
Kenya_Bus_2006
Kenya_Tk1_2005
Russia_Georgia_2007
Spain_E75_1975
Portugal_OURT_88.3_1988
Italy_47_Ss_2008_2008
Italy_26544_OG10_2010
Estonia_2014
Uganda_R35_2015
Uganda_N10_2015
Uganda_R25_2015
Uganda_R7_2015
Uganda_R8_2015
Poland_ASFV_POL_Podlaskie_2015
Poland_Pol17_05838_C220_2016-2017
Poland_Pol17_04461_C210_2016-2017
Poland_Pol17_03029_C201_2016-2017
Poland_Pol16_29413_o23_2016-2017
Poland_Pol16_20540_o10_2016-2017
Poland_Pol16_20538_o9_2016-2017
Poland_Pol16_20186_o7_2016-2017
China_LN_2018
China_HLJ_2018
China_ASFV-SY18_2018
China_AnhuiXCGQ_2018
Spain_BA71_1971
0
50
100
150
200
250
300
Annotation from NCBI
Newly Annotated
Genetic Remains

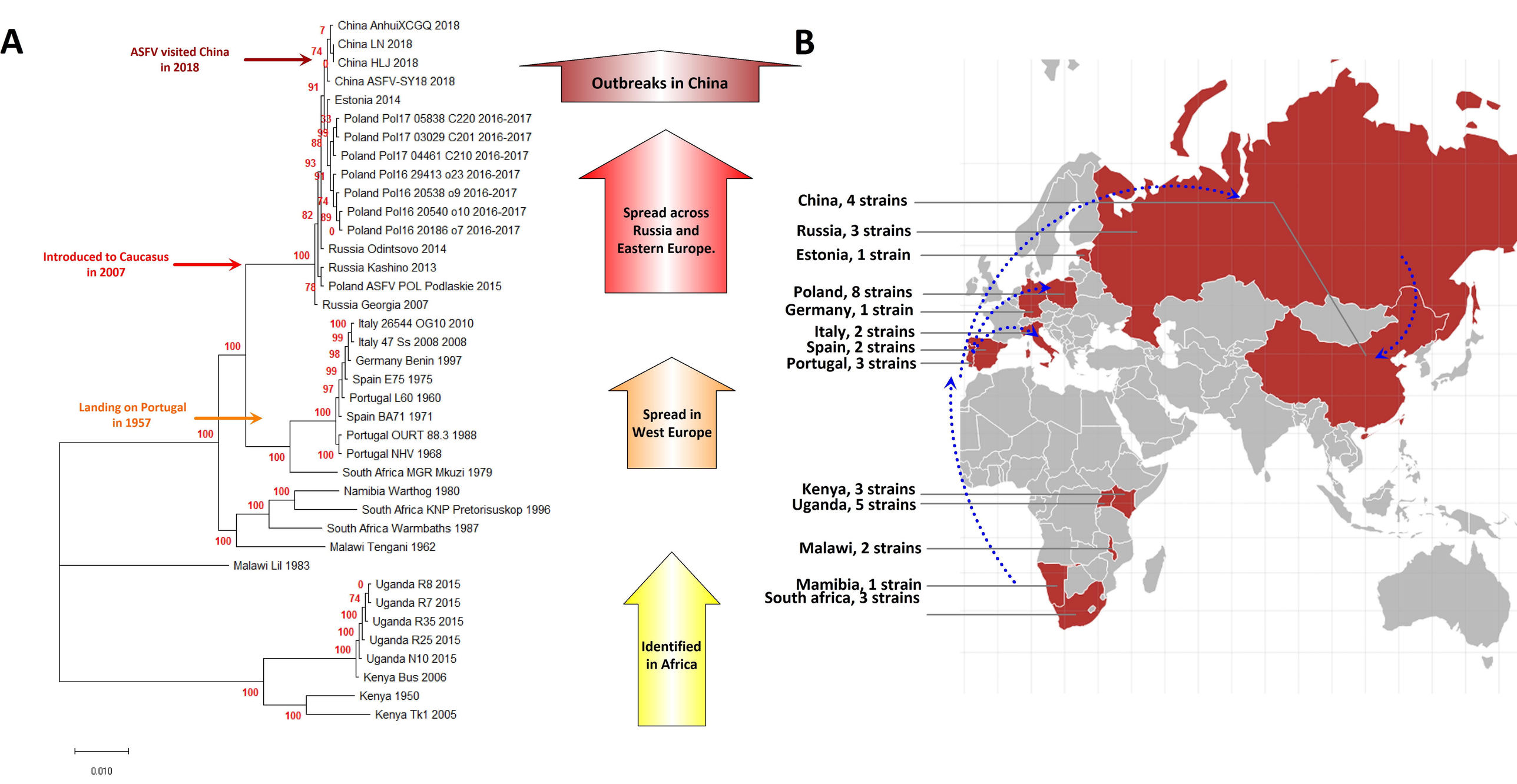
A

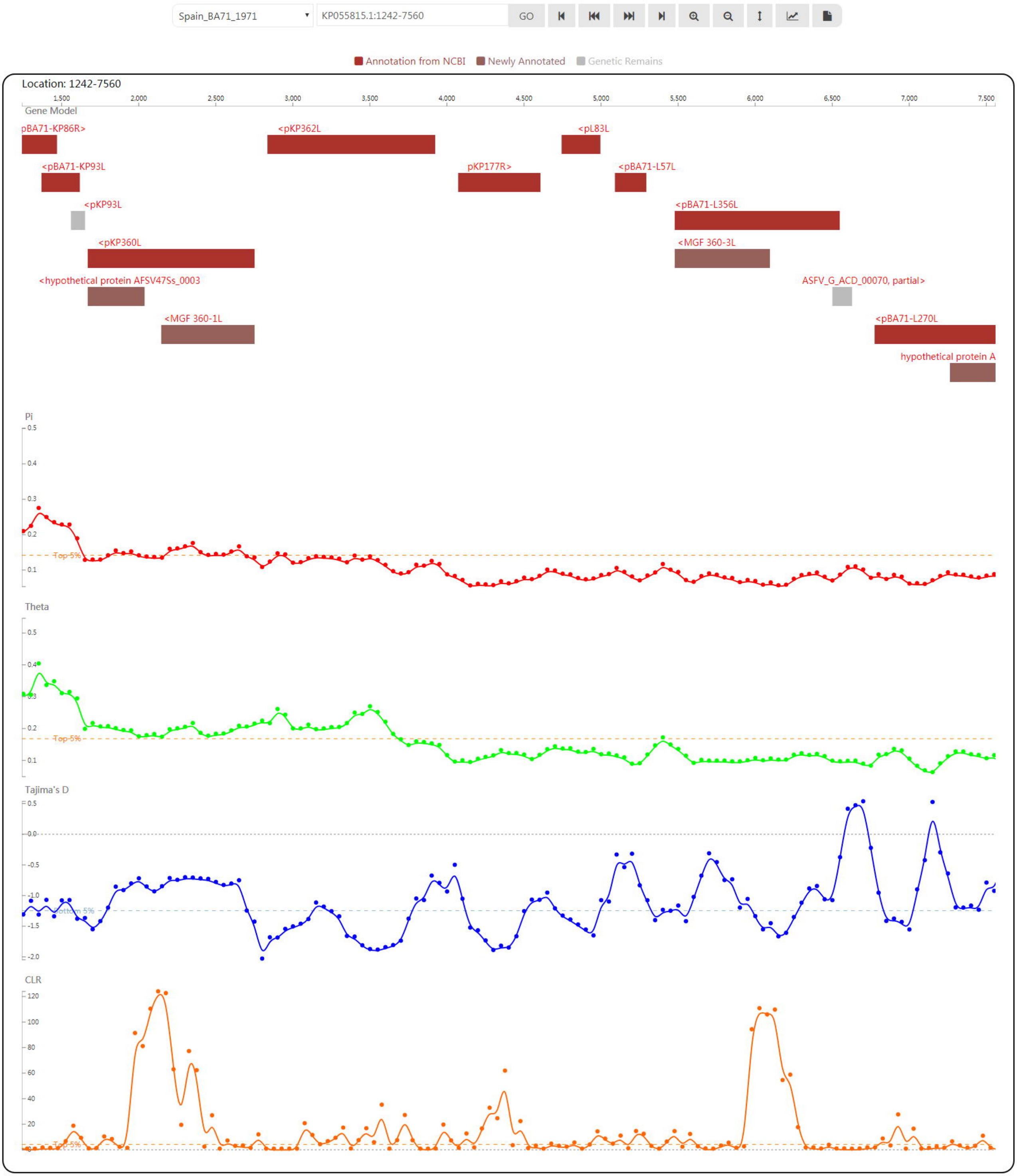

Capsid
Extracell [Secreted, Extracellular]
Multivesicular late endosome
Nucleus
Membrane
Endocytosis
Fusion
Uncoating
Naked core
Early Endosome
Macropinocytosis
Viral factory
Budding
Host nucleus › host nucleolus. Host cytoplasm. Secreted.
Host nucleus. Host cytoplasm.
Cytoplasm

B


8%
8%
52%
10%
3%
7%
6%
2%
4%
Capsid
Cytoplasm
Single-Pass Membrane
Multi-Pass Membrane
Nucleus
Host nucleus. Host cytoplasm.
Host nucleus › host nucleolus. Host cytoplasm. Secreted.
Extracell[Secreted, Extracellular]
Others-Single


Percentage of gene clusters
Number of gene clusters
75
60
45
30
15
0
94
56
0
virion
virion part
other organism
other organism part
cell
cell part
organelle
organelle part
protein-containing complex
membrane-enclosed lumen
symplast
cell junction
membrane
membrane part
catalytic activity
structural molecule activity
binding
molecular function regulator
response to stimulus
metabolic process
cellular process
regulation of biological process
growth
biological regulation
positive regulation of biological process
multi-organism process
negative regulation of biological process
multicellular organismal process
developmental process
locomotion
cellular component organization or biogenesis
localization
signaling
immune system process
reproduction
reproductive process
biological adhesion
Cellular Component
Molecular Function
Biological Process


A
ASFV visited China in 2018
Introduced to Caucasus in 2007
Landing on Portugal in 1957
China AnhuiXCGQ 2018
China LN 2018
China HLJ 2018
China ASFV-SY18 2018
Estonia 2014
Poland Pol17 05838 C220 2016-2017
Poland Pol17 03029 C201 2016-2017
Poland Pol17 04461 C210 2016-2017
Poland Pol16 29413 o23 2016-2017
Poland Pol16 20538 o9 2016-2017
Poland Pol16 20540 o10 2016-2017
Poland Pol16 20186 o7 2016-2017
Russia Odintsovo 2014
Russia Kashino 2013
Poland ASFV POL Podlaskie 2015
Russia Georgia 2007
Italy 26544 OG10 2010
Italy 47 Ss 2008 2008
Germany Benin 1997
Spain E75 1975
Portugal L60 1960
Spain BA71 1971
Portugal OURT 88.3 1988
Portugal NHV 1968
South Africa MGR Mkuzi 1979
Namibia Warthog 1980
South Africa KNP Pretorisuskop 1996
South Africa Warmbaths 1987
Malawi Tengani 1962
Malawi Lil 1983
Uganda R8 2015
Uganda R7 2015
Uganda R35 2015
Uganda R25 2015
Uganda N10 2015
Kenya Bus 2006
Kenya 1950
Kenya Tk1 2005
0.010
Outbreaks in China
Spread across Russia and Eastern Europe.
Spread in West Europe
Identified in Africa
B
China, 4 strains
Russia, 3 strains
Estonia, 1 strain
Poland, 8 strains
Germany, 1 strain
Italy, 2 strains
Spain, 2 strains
Portugal, 3 strains
Kenya, 3 strains
Uganda, 5 strains
Malawi, 2 strains
Mamibia, 1 strain
South africa, 3 strains


Spain_BA71_1971
KP055815.1:1242-7560
GO
Annotation from NCBI
Newly Annotated
Genetic Remains
Location: 1242-7560
1,500
2,000
2,500
3,000
3,500
4,000
4,500
5,000
5,500
6,000
6,500
7,000
7,500
Gene Model
pBA71-KP86R>
<pKP362L
<pL83L
<pBA71-KP93L
pKP177R>
<pBA71-L57L
<pKP93L
<pBA71-L356L
<pKP360L
<MGF 360-3L
<hypothetical protein AFSV47Ss_0003
ASFV_G_ACD_00070, partial>
<MGF 360-1L
<pBA71-L270L
hypothetical protein A
Pi
0.5
0.4
0.3
0.2
0.1
Top 5%
Theta
0.5
0.4
0.3
0.2
0.1
Top 5%
Tajima's D
0.5
0.0
-0.5
-1.0
-1.5
-2.0
Bottom 5%
CLR
120
100
80
60
40
20
0
Top 5%