The Genetic Variation Landscape of African Swine Fever Virus Reveals Frequent Positive Selection on Amino Acid Replacements

Yun-Juan Bao[1,2,#], Junhui Qiu[2], Fernando Rodríguez[3], Hua-Ji Qiu[4,#]

[1] *State Key Laboratory of Biocatalysis and Enzyme Engineering, Hubei Collaborative Innovation Center for Green Transformation of Bio-Resources, Hubei Key Laboratory of Industrial Biotechnology*.

[2] *School of Life Sciences, Hubei University, Wuhan 430062, China*.

[3] *IRTA, Centre de Recerca en Sanitat Animal (CReSA, IRTA), Campus de la Universitat Autonòma de Barcelona, Bellaterra 08193, Spain*.

[4] *State Key Laboratory of Veterinary Biotechnology, Harbin Veterinary Research Institute, Chinese Academy of Agricultural Sciences, Harbin 150001, China*.

Corresponding authors:
Email: yjbao@hubu.edu.cn (YJ Bao), qiuhuaji@caas.cn (HJ Qiu)

Running title: Variation Landscape of African Swine Fever Virus

Abstract

African swine fever virus (ASFV) is a lethal disease agent that causes high mortality in swine population and devastating loss in swine industries. The development of efficacious vaccines has been hindered by the knowledge gap in genetic properties of ASFV and the interface of virus-host interactions. In this study, we performed a genetic study of ASFV aiming to profile the variation landscape and identify genetic factors with signatures of positive selection and relevance to virus-host interactions. To achieve this goal, we developed a new tool "SweepCluster" for systematic identification of selective sweep. Our data reveals a high level of genetic variability of ASFV shaped by both diversifying selection and selective sweep. The selection signatures are widely distributed across the genome with the diversifying selection falling within 29 genes and selection sweep within 25 genes. Further examination of the structure properties reveals the link of the selection signatures with virus-host interactions. Specifically, we discovered a site at 157th of the antigen protein EP402R under diversifying selection located in the cytotoxic T-cell epitope involved in the serotype-specific T-cell response. Moreover, we reported 24 novel candidate genes with relevance to virus-host interactions. By integrating the candidate genes with selection signatures into a unified framework of interactions between ASFV and hosts, we showed that those genes are involved in multiple processes of host immune evasion and virus life cycles, and may play crucial roles in circumventing host defense systems and enhancing adaptive fitness.

Importance

ASFV causes lethal disease in swine population with up to 100% mortality rates in domestic pigs. The recent outbreak of the disease has spread rapidly worldwide resulting in a large amount of deaths of pigs and tremendous damages to the swine industry. There is no commercially available vaccine against ASFV infection. Current vaccine strategies face the challenges of incomplete protection or deficient cross-protection. The challenges strongly highlight the need for thorough

understanding of genetic properties at the interface of virus-host interactions. In this study, we developed a new bioinformatics tool “SweepCluster” and employed computational approaches to characterize the genetic variation landscape of the virus aiming to identify genetic factors relevant to host interactions. The results we present will allow enhanced understanding of genetic basis of rapid adaptation of ASFV and provide valuable targets for therapeutic intervention. The new analytic tool we offer could be used as a general approach for selection analysis.

Keywords: African swine fever virus; Genetic variation; Virus-host interactions; Positive selection; Selective sweep

Introduction

African swine fever virus (ASFV) is the causative agent of haemorrhagic fever in pigs. ASFV mainly replicates in pig macrophages, causing up to 100% mortality rates in domestic pigs. Conversely, ASFV infects African wild pigs (warthogs) in asymptomatic manners, which act as ASFV reservoirs together with the soft tick (*Ornithodoros spp*.). ASFV is thought to originate from and circulate in wild pigs and soft ticks in Eastern Africa, and the first infection in domestic pigs was reported in Kenya in 1921 (1), coinciding with their first introduction in Europe. From then on, ASFV has spread through Africa and the rest of the world, being the most prominent exportations in 1957 and 1960 to Portugal and in 2007 to Georgia (2). This last introduction led to the expansion of the disease through the Caucasus to European Union and more recently to China and neighboring countries in 2018, affecting hundreds of millions of pigs and threatening the global pork industry (3, 4).

Since there is no commercially available vaccine against ASFV infection, current disease control is based on physical quarantine and animal slaughtering. Numerous pigs have been killed since the spread of infection globally, causing substantial damages on the pork industry (5). Development of efficacious therapeutic and prophylactic tools has been largely hindered by the limited knowledge of genetics and genomics properties of ASFV and the interface of virus-host interactions.

ASFV is a large double-stranded DNA (dsDNA) virus with a genome length of 170~194 kb. Tens of genomes of ASFV strains have been completed by using high-throughput sequencing technologies. It has been shown that the ASFV genome is well conserved in the central part but highly variable at both ends encoding genes of multigene families (MGFs) 505, 360, 300, 110, and 100 (6-10). The genes in each of these families have multiple copies (or paralogs) from 4 to 19, probably induced by gene duplication and gene gain/loss during adaptive evolution(11). Recent studies using engineered deletion mutants investigated the

variation patterns of MGF genes (7, 12), showing that MGF genes are relevant to host interactions and the multiple paralogous copies might be responsible for host tropism (10, 13).

However, there are a limited number of studies for systematic characterization of genetic properties in the genome-wide scale (6, 7). As a dsDNA virus, ASFV has an estimated substitution rate $\mu \sim 6.7 \times 10^{-4}$ (substitutions per site per year) (14), comparable to that of RNA viruses such as the influenza virus with $\mu \sim 10^{-3}$ (15), higher than that of other large dsDNA viruses such as herpes simplex type I with $\mu \sim 10^{-5}$ (16, 17), much higher than that of many bacterial species such as *Streptococcus pneumoniae* with $\mu \sim 10^{-6}$ (18). The high substitution rate indicates a high level of variability among ASFV genomes. It provides genetic foundations for systematic characterization of genetic variations at the genome-wide scale.

In this study, we performed a comparative genomic and genetic study to profile the variation landscape of ASFV and identify the candidate genetic factors with relevance to host interaction. To achieve this goal, we have developed a new tool SweepCluster to capture the genomic regions of mutations under putative selective sweep. It has the advantage of masking the confounding effect of genomic recombination on detection of selective sweep.

Results

Single nucleotide polymorphism (SNP) detection and selection pressure in the core genome of ASFV

We performed comparative genomic study of the ASFV strains by aligning the genomic sequences of the strains to the core genome. The list of ASFV genomes we used is shown in Table S1. Using 27 non-redundant genomes, we identified 18,070 SNPs, of which 6088 are non-synonymous, corresponding to an average of 129 SNPs/kb. In order to examine the influence on variation detection from the five distantly evolved strains from Africa, *i.e.*, Ken05-Tk1, Kenya-1950, Ken06-Bus, UgandaN10-2015, and UgandaR7-2015 (19, 20)

(Figure 1), we excluded the five strains, repeated the comparative analysis and obtained 12,652 SNPs with an average 91 SNPs/kb, again reflecting the high genetic diversity of ASFV. The high mutation rate is in contrast with the previous notion of high conservation of the core genomes of ASFV. Therefore, we further estimate the overall selection pressure exerted on the ASFV population using Tajima's *D* test (21). The calculation of Watterson's estimator θ (22) gives a genome-wide average mutation rate of 0.025, significantly greater than the average pair-wise nucleotide difference of 0.019. It results in a negative Tajima's *D* value of -2.30, indicating evolutionary positive selection of the ASFV population.

Phylogenetic structure of the ASFV population

The genome-wide phylogeny was inferred using the core genome SNPs of the 27 non-redundant strains (Figure 1a). The phylogenetic tree identifies three major distantly related clades (α, β, and γ). The three-clade topology is consistent with that derived from the full-length structural gene p72 (B646L) of the same set of genomes and the partial-length p72 sequences from a broader set of 85 isolates (Figure 1b,c, Figure S1, and Supplementary file 1). The first clade α contains three closely related subgroups, comprising isolates from Europe of genotype I, isolates from Caucasus of genotype II, and isolates from Southern Africa of diverse genotypes, respectively. The second clade β consists of isolates from Eastern Africa of genotype X and IX, which are the predominant genotypes causing outbreaks in this area (23). The third clade γ mainly contains Eastern African isolates of genotype VIII, XI, XII, and XIII, although only one complete genome is available in this clade (Malawi-Lil83 of genotype VIII). The phylogeny topology is consistent with that constructed previously based on different number of ASFV strains (2, 6).

We observed two prominent features of the phylogenetic structure and geographical distribution depicted in Figure 1. First, the tree has a total branch length of 0.25 substitutions per site. The long phylogenetic distance and relatively short separation time between the three

clades, especially α and β indicates that they have accumulated a significant number of genetic differences in a short period of time. Secondly, the virus has recurrently emerged at the same countries at different time points but exhibits significant genomic modifications, such as those isolates from Malawi (Malawi-Tengani62 and Malawi-Lil83 with a genetic distance of 0.09 substitutions per site). It implies that ASFV might be able to rapidly adapt to specific host environments by acquiring a multitude of variations. Next, we will investigate the genetic properties of the variations and the mechanisms they were introduced.

Identification of genes with high frequencies of non-synonymous mutations

The pattern of gene duplication and loss affecting the MGFs at both ends of the ASFV genomes has been intensively studied (12, 13, 24), largely due to the postulated roles of MGF360 and MGF505 in host immune evasion and infection tropism (10, 13). Here, we focus on the whole genome to characterize the genetic variation properties. We at first identify the variations associated with virulent phenotypes of ASFV strains. The low number of non-virulent strains in the currently known data set prevents us from performing a robust statistical association study, we quantified the non-synonymous allelic changes uniquely present in the two natural isolates with low virulence, *i.e.*, Portugal-NHV68 and Portugal-OURT88. A total of 13 non-synonymous mutations from 10 genes were uniquely present in the two Portugal isolates (Table S2). However, none of the genes is enriched with the unique mutations with statistical significance in comparison with the genome-wide average using Hypergeometric tests.

Therefore we further examined the distribution of all 6088 non-synonymous mutations along the genome and identified the gene loci mutated more frequently than the genome-wide average (Figure S2a). The analysis using Hypergeometric test ranked 23 genes to be significantly enriched with non-synonymous mutations (multiple testing corrected p-value $\leq$ 0.001) but not with synonymous mutations (multiple testing corrected p-value $\geq$ 0.05) (Table

1 and Figure S2b). Half of the genes are the members of MGF360, MGF505, and MGF300. The list also includes the genes involved in DNA replication/repair, nucleotide metabolism, redox pathway, host interactions, and others with unknown functions. The non-synonymous mutations in the 23 genes were further laid on each protein domain architecture identified by comparison with the PFAM database (25) (Table S3). We found no significant difference of the mutation distribution between the key functional domains and the neighboring regions. The zoomed-in view of the density distribution of the non-synonymous mutations along the domain architectures for the top genes is shown in Figure 2c.

Identification of genes under positive selection based on the dN/dS method

The high rate of non-synonymous mutations observed prompted us to test the potential occurrence of positive Darwinian selection acting on the ASFV-encoded genes. We test the potential positive selection by measuring the rates of non-synonymous substitution (dN) and synonymous substitutions (dS) and calculating their ratio dN/dS for each gene based on the Nei & Gojobori model (26). The analysis shows that most of the genes have a value of dN/dS < 0.5 and the average value of dN/dS is 0.1, revealing the evolutionary stability of the genes (Table S4). Notably, at the top of the list are six genes with the value of dN/dS ≥ 1 (D1133L, DP63R, 86R, EP153R, EP402R, and MGF505-4R). By removing three genes with deflated values of dS due to increased selection against synonymous substitutions (dS < 0.028, p-value < 0.02, one-tailed t-test), we finally obtained three genes (EP153R, EP402R, and MGF505-4R) with dN/dS > 1, subject to potential positive selection. Among them, the gene MGF505-4R with the value of dN/dS $= 1.2$ was also found to be significantly enriched with non-synonymous mutations in the previous section, implying strong positive selection acting on this gene. The other two genes, the CD2 homolog protein EP402R and C-type lectin-like protein EP153R, were previously shown to be involved in host immune evasion and the hemagglutination ability of ASFV depends on these two genes (27, 28).

Test of selection pressures on individual sites of genes

In most organisms, the genes with dN/dS>1 are rare because non-synonymous mutations are generally detrimental to protein functions and are not preferred. Therefore, the individual sites positively selected are usually masked by the low average value of gene-wide dN/dS. In order to unravel the potential selection acting on specific sites of the genes, we performed likelihood ratio tests (LRTs) using the site-specific model of dN/dS (ω) implemented in PAML (29, 30). We identified 29 genes having been subject to potential positive diversifying selection (p-value ≤ 0.05, Chi-squared test) on an average of 3.1% (±2.4%) of sites (posterior probability ≥ 0.9) (Figure 2a and AdditioTable S5). The list of genes under positive selection covers 11 of the 18 genes with p-value ≤ 0.05 and 8 of 10 genes with p-value ≤ 0.01 identified by a comparative study of 11 complete genomes (6).

The genes here we identified include 17 candidates known to be involved in host cell interactions, such as EP402R, EP153R and MGF genes. Notably, we also discovered twelve novel candidates, which have not been shown to be related with host interactions or investigated thoroughly experimentally, such as the highly divergent proteins B117L and B602L, and the conserved structural protein pp220/CP2475L (Table 2 and Table S5).

In order to ascertain the functional implication of the positively selected sites in the genes, we tabulated the sites under positive selection in each gene with a posterior probability ≥ 0.9 and mapped the sites to the domain architectures of the genes (Figure 2b,c and Supplementary file 2). The positively selected sites are largely located in the variable regions or around the short repeats of the genes, such as EP402R, EP153R, B117L, and B475L. Specifically, the positively selected sites in EP402R are enriched in the extracellular domain (p-value = 0.046, Hypergeometric test), which is highly variable among the ASFV lineages. The extracellular domain has an Ig-like structure resembling to host CD2 protein and is essential for binding of red blood cells to infected cells or extracellular virions (31, 32).

Given the key functions in host infection, EP402R has been described as an important virulence factor and immunogenic target (33, 34). Here we use EP402R as an example to demonstrate the feasibility of using positively selected sites to delineate their links with virus-host interactions. We collected the CD2 homologs of EP402R in animals with known functions and structures, and performed structure-guided comparison with the EP402R extracellular Ig-like domain (Figure 2d and Figure S3). As a CD2 homolog, the extracellular domain of EP402R consists of a constant C-set and a variable V-set Ig-superfamily domain (Figure S3a-d). We then mapped the positively selected sites to the aligned sequences and the tertiary structures. It is remarkable that the sites predominantly reside in the loop regions on the top of the V-set domain of EP402R, in clear contrast with the location of the ligand-binding sites of host CD2 at the side face of the V-set domain (Figure S3a-c) (35). The corresponding loop regions in Ig antibodies are the domains facilitating specificity of antibodies to recognize antigens (36). Therefore, it supports the notion that the loop regions in EP402R and the positively selected sites within the regions are relevant to specificity of ASFV in host cell recognition.

The sites under positive diversifying selection have critical implications for vaccine cross-protection from heterologous viral strains when the subunits containing those sites are used as vaccines. Indeed, one of the positively selected sites E157 is located within the cytotoxic T-cell epitope A6 previously identified (37). The positive diversifying selection on the site E157 and the high variability of the epitope motifs among ASFV strains provide at least partial molecular etiology of the serotype-specific T-cell response against DNA vaccines containing the epitopes in EP402R (Figure S3e). Given the frequent occurrence of positive diversifying selection in a broad set of genes, full evaluation of the sequence variability of the target genes in designing vaccines is warranted.

In addition to the divergent proteins, four highly conserved structural proteins

(J5R/H108R, P11.5/A137R, P10/K78R, and pp220/CP2475L, in Figure 2c) were also found to possess positively selected sites, which have not been shown to be involved in host interactions experimentally. J5R/H108R is a transmembrane protein at the inner envelope and P10 is a DNA-binding protein in the viral nucleoid. The positive selection of the sites in these structural proteins may represent the evolutionary adaptation of ASFV for successful colonization and survival in the host niches. Another two proteins with unknown functions (MGF300-4L and B475L, in Figure 2c), have the positively selected sites distributed across a large proportion of the gene regions. The two proteins are unique that they exhibit high propensity for forming helices through the whole gene region. In spite of being unable to obtain confidently a tertiary structure model for the two proteins, we predicted the secondary structure of MGF300-4L and B475L using PSIPRED (38). It shows that the two proteins predominantly comprise α-helices, indicating their possible roles in protein-protein interactions (Figure S4).

Identification of selective sweeps in the ASFV genomes

A selective sweep is a process where a beneficial allelic change with strong positive selection sweep through the population and the nearby sites will hitchhike. The process leads to specific gene regions with reduced within-population genetic diversity and increased between-population differentiation. Such selective sweeps allow for rapid adaptation and accelerated evolution, and are good indicators for host-pathogen interaction and adaptive evolution (39). The unique mechanism of selective sweeps in causing genetic changes makes it inappropriate to detect them using the dN/dS-based method. Therefore, we developed an open-source tool SweepCluster for detecting the regions with clustered SNPs under selective sweep. The method is also able to generate significance levels for each detected sweep regions based on the spatial distribution model of the SNPs (see Methods). SweepCluster is different from previous methods in that it does not depend on genetic distance between SNPs

and thus is exempted from influences of recombination events. We at first identified 6,054 SNPs associated with between-population subdivision and within-population homogeneity for the clade α and β (Figure 3a). Those SNPs were subsequently subject to detection of selective sweep using SweepCluster. A total of 578 clusters of SNPs were identified encompassing 4,741 SNPs or 2,139 non-synonymous SNPs (Supplementary file 3). That is corresponding to 26% of the total SNPs or 35% of the total non-synonymous SNPs, indicating that a high proportion of the genetic variations among the ASFV population have been likely to be introduced *via* selective sweep. Among them, 32 regions from 25 genes show signatures of selective sweep with high significance (Figure 3b,c and Table 3).

The gene regions with significant selective sweep exhibit higher population differentiation and reduced sequence diversity as shown in the key signature genes (Figure 3d). Among them are a series of known gene factors involved in host cell interactions, including MGF505, MGF360 and I215L, which also harbor sites under positive diversifying selection. Those gene factors exhibit genetic signatures of both diversifying selection and selective sweep (Figure 3d and Figure 2b). Noteworthy are the 15 novel candidate genes showing strong signatures of selective sweep (Table 3). A large proportion of them (60%) are involved in key cellular functions, such as replication, repair, transcription, and metabolism.

We notice that four of the novel candidates (A151R, F1055L, CP312R, and E146L) have been previously demonstrated to induce immune responses in pigs following ASFV challenge (40, 41). Therefore, we proceed to characterize the shared genetic properties of the candidate genes and compare with that of known genes inducing immune responses or involved in host cell interaction.

Sequence variability of the candidate genes with diversifying selection or selective sweep

We ascertain the genetic properties of the genes with positive diversifying selection or selective sweep by calculating population prevalence frequencies and pair-wise amino acid

divergence of the genes and doing comparison with three gene categories cataloged from other studies: (i) the non-antigenic conserved structural proteins without positive selection (32), (ii) the antigen proteins eliciting immunological responses in immunoassay experiments (40-42), (iii) the proteins previously shown to be involved in host cell interactions (10, 43) (Figure 4 and Table S6). A non-uniform population prevalence and higher level of sequence variability are observed in the candidate genes under putative positive diversifying selection in comparison with the category of (i) conserved structural proteins and (ii) antigenic proteins, but not with the gene category (iii) involved in host cell interactions (two-sided Mann-Whitney U-test, Figure 4a,d,i). The overall high divergence in amino acid sequences coupled with the significant positive diversifying selection of those genes suggests that they have mutated frequently during evolution. In contrast, the candidate genes with signatures of selective sweep are relatively more conserved and present a comparable level of sequence variability with that of conserved structural proteins and the known antigenic proteins, supporting their potentiality as generalized immunogenic targets (Figure 4b,e,i).

Patterns of gene loss among the ASFV strains

Gene loss can result from accumulation of truncating mutations affecting the translation of proteins or from removal of the whole gene *via* recombination events. It has been acknowledged as an important vehicle of genetic changes driven by selection pressures for circumventing host defensing systems. In this study, we seek to decipher the genes lost in the non-virulent isolates but intact in virulent isolates or vice versa. We found that two genes EP402R and EP153R, the known virulence factors, are truncated by point mutations in the two non-pathogenic strains Portugal_NHV68 and Portugal_OURT88, consistent with previous findings (7, 44) (Figure 5a). Multiple truncating mutations have occurred in EP402R in the two non-virulent Portugal strains (Figure 5a).

Gene elimination by segmental deletions has been previously investigated based on a

limited number of isolates (12, 13, 24). We also observed a large fragmental deletion at the left end of the genome in all of the three non-pathogenic ASFV strains, Spain-BA71V, Portugal-NHV68, and Portugal-OURT88 in comparison with the virulent strains (Figure 5b). The deletion results in truncation of MGF360-9L in Spain-BA71V, complete removal of MGF360-10L, 11L, 12L, 13L, 14L, and MGF 505-1R in all three strains, full deletion of MGF505-2R and partial deletion of MGF505-3R in the two Portugal strains Portugal-NHV68 and Portugal-OURT88. While the strains Portugal-NHV68 and Portugal-OURT88 are naturally low virulent isolates harboring large fragmental deletions in the genome (45), Spain-BA71V is attenuated by adaptation to cultured cell lines, introducing large fragmental deletions (12). Interestingly, a recent study of the adaptation of a highly virulent strain Georgia-2007 (ASFV-G-ΔMGF) to established cell lines also demonstrated the decreased virulence in swine accompanied by fragmental deletions removing a distinct set of members of MGF360 (13L and 14L) and MGF505 (2R, 3R, 4R, 5R, and 6R) (24). Calculation of the pair-wise amino acid divergence of the eliminated MGF genes reveals comparable sequence variability with the MGF genes with diversifying selection or selective sweep (Figure 4c).

The deletion pattern of EP402R, EP153R, and MGF360/505 in ASFV strains reflects the differential selection pressures among ASFV strains and the link between those genes and virulence (31, 46-48). The links of the genes with virulence, coupled with the functional interactions with host immune systems, make them preferred candidates for immunogenic targets, as demonstrated in previous studies (13, 33, 37).

Genetic diversity and divergent selections among paralogous gene members of MGF360/505

Given that a large number of MGF genes have been identified to be genetically diverse with intensive signatures of positive selection, an natural question is: how about the breath of genetic diversity and selection pressures among the paralogous members of MGF and which

regions are responsible for the genetic and functional diversity? We examine the genetic diversity of MGF genes by evaluating the differential selection between paralogous genes/branches of MGF360 or MGF505. We first constructed the phylogenetic structures of all orthologous and paralogous members of MGF360 and MGF505, respectively (Figure 6a,c and Figure S5), and then chose the phylogenetically close pairs of genes/branches to perform the likelihood ratio test of divergent selection (See Methods). The test identified 10 and 9 pairs showing divergent selection on an average of 8.3% and 9.6% of the sites among MGF360 and MGF505, respectively (p-value ≤ 0.05, Chi-squared test) (Figure 6a,c and Table S7). The divergent selection clearly indicates the distinct evolutionary forces exerted on the array of paralogs of MGF, thus forming a genetic pool for functional diversification. The functional diversification is further supported by the divergent regulation patterns across the paralogous members of MGF (Figure 6b,d). Though the expression data for MGF genes is unavailable, the regulatory divergence is manifested qualitatively in the distinct promoter motifs and their distances to the translation start site (TSS) among paralogous members of MGF. Further profiling the promoter regions 55 nucleotides upstream TSS of MGF genes shows that the promoter divergence is correlated with the evolutionary distances between paralogs of MGF (Figure 6e,f). The regulatory divergence in the promoter regions, coupled with the differentiated selection pressures between paralogous pairs of MGF360 and MGF505 constitutes important genetic basis for functional diversification of MGF genes, providing a wide spectrum of specificity in host tropism and adaptation.

To unveil the genetic properties of the gene regions under divergent selection, we identified the sites under putative divergent selection between the paired genes/branches of MGF360/MGF505, and quantified the site distribution along the predicted secondary structure of MGF360/MGF505, respectively (Figure 7 and Supplementary file 4). Interestingly, the sites exhibit quasi-periodic distribution and are enriched periodically in a

few patches of length ~ 30 residues (p-value ≤ 0.05, Hypergeometric test). This average length of enrichment is close to the length of the ankyrin repeat (49), which is believed to be the building blocks of the MGF protein structures. Actually, the predicted secondary structures of MGF360 and MGF505 display signatures of tandem ankyrin repeats, each consisting of a helix-loop-helix motif followed by another loop region. Protein domains containing tandem ankyrin repeats usually fold into a conserved tertiary concave/convex structure mediating protein-protein interactions. The surface recognition residues are highly variable, affording specific interactions with a broad range of host targets (49). Ankyrin repeats have been described to be the major functional units in host range factors in several poxvirus species (50-52). Here in the absence of the protein structure of MGF proteins, we demonstrated that the periodic patches of residues in ankyrin repeats exhibit differentiated evolutionary selection among paralogous members, thereby representing the motifs facilitating functional diversity of MGF in the multifaceted interactions with host cells. Further studies are required to ascertain the role of the motifs in host interactions.

Discussion

In our pursuit of characterizing the variation landscape of ASFV genomes and unraveling a comprehensive set of candidate genes with potential relevance to host interactions, we developed the new tool SweepCluster, which is able to capture the regions of clustered SNPs caused by selective sweep. It depends on the spatial distribution model and functional properties of the SNPs without confounded by the effects of recombination events, thereby capable of detecting mutated regions under recent and ancient selection in the whole genome range.

In total, we identified 29 candidate genes with positive diversifying selection using PAML (29) and 25 with selective sweep using the newly developed tool SweepCluster. Among them, eight show signatures of both kinds of selection and 24 are novel candidates

that so far, have not been reported to be associated with host interactions. The genes showing selection signatures are widely distributed across the genome highlighting intense adaptive evolution of ASFV. We summarize and present the candidate genes in a unified scheme of interactions between ASFV and hosts in a framework of the virus life cycles and host defense processes (Figure 8) (53).

The proteins in the scheme include those known to be relevant to host immune evasion, such as EP402R for surface adherence of infected cell (31), EP153R for inhibition of MHC expression and host cell apoptosis (32, 46, 54), A238L for production impairment of immune regulator NF-κB and cytokines TNF-α (55, 56), and multiple MGF genes for modulation of interferon (IFN) response (47, 57, 58).

The scheme also contains the proteins critical for the virus life cycles facilitating successful entry and proliferation in host cells, such as the structural proteins pp220, J5R, P11.5, P10, and B602L localizing at distinct layers of the viral particles for virus entry and assembly (32, 59), the basic enzymes P1192R, F1055L, F778R, A240L and EP1242L involved in replication, repair and transcription in host cytoplasm (10). The key roles played by the proteins and the relatively high conservation make them promising candidates for vaccines with cross-activity. We also detected a few novel candidates showing significant selection pressures, such as, MGF300-4L, B117L, B475L, 86R, L60L, DP238L, and I267L with unknown functions. These genes could serve as potential targets for future immunoassays.

The cellular processes the candidate genes are involved in, provide a variety of sources of selective pressures acting at multiple stages of the infection cycles for ASFV to evolve and adapt. In this regard, these genes may constitute an important part of the genetic factors of ASFV in circumventing host defense systems and enhancing fitness in a specific manner.

Our data revealed that the adaptive evolution of ASFV have been shaped by both positive

diversifying selection and selective sweep. However, the characterization of the genetic properties of the genes with selection signatures show that the genes with diversifying selection exhibit a higher level of sequence variability than those with selective sweep. The results provide important implications for vaccine design.

The most prominent are EP402R, EP153R and MGF genes, with the highest genetic variability, the only known proteins so far shown to be both virulence determinants and immunogenic targets. However, the high sequence diversity of EP402R/EP153R and mosaic presence pattern of MGF genes among the ASFV population make it difficult for them to achieve desirable cross-protection (33, 60). The dual role of EP402R, EP153R and MGF genes, as both virulence determinants and immunogenic proteins, may also introduce confounding factors in designing live-attenuated virus vaccines (LAVs). As an encouraging example, elimination of EP402R from the virulent BA71 to obtain the LAV strain BA71ΔCD2, protected pigs against homologous and heterologous virus challenges (34). Similarly, ASFV-Georgia-ΔMGF, a LAV strain lacking a series of MGF genes, protected animals against homologous challenges (13). Unfortunately, sequential gene deletions provoked in occasion the loss of protection by excessive attenuation (61).

The divergent selection between MGF genes further complicates the vaccine design. We identified differentiated selection pressures and regulation patterns between paralogs of MGF genes conferring genetic diversity and functional diversification. The possible scenario is that the antigenic activities and expression levels of paralogs of MGF genes are strain-specific and/or host-dependent. This scenario provides a rationale for the observations that variable deletion patterns and expression profiles of MGF genes have been resulted from different adaptation processes or have induced distinct viral growth outcomes in host niches (12, 24). Up to now, the precise connections between the MGF genes and physiological conditions are still largely unknown. Optimal choices of MGF genes and gene regions remain to be tested

when they are used as immunogenic targets. The specific sites under divergent selection we dissected in MGF360/MGF505 provide important information in aiding for the tests.

Compared to the high divergence of the candidate genes with diversifying selection, the genes with selective sweep display a low level of within-population diversity at sweeping regions and a high degree of average conservation. Many of them (60% of the novel candidates) are involved in the critical events in the life cycles of ASFV infections, such as replication, repair and transcription. Interestingly, an evolutionary study of the influenza A virus H3N2 showed that the emergent severe seasonal flu in 2004/2005 was correlated with mutations in the key ribonucleoprotein (RNP) complex acquired by a circulating lineage *via* selective sweep and the lineage was demonstrated to induce elevated replicative fitness and more severe clinical diseases (62). We argue that the genes with selective sweep are important contributing factors for the rapid adaptation and enhanced fitness of the ASFV population circulating in specific areas. The high conservation and critical roles of the genes make them promising candidates for vaccine molecules or drug targets.

The multifaceted genetic characteristics of ASFV genes imply that the virus may have evolved multiple mechanisms and pertinent genetic factors for successful replication, adaption, and persistence during interaction with continuously changing host environments, including warthogs, ticks, and domestic pigs. Although the methods we used for identifying selection are not perfect due to the small size of the ASFV population, the data here provides novel insights and valuable targets for vaccine development or therapeutic intervention.

Methods and Materials

Comparative genomic study and phylogenetic inference

The genomic sequences and annotations of ASFV used in this study were downloaded from NCBI GenBank (ftp://ftp.ncbi.nlm.nih.gov). A total of 36 strains were obtained and 27 non-redundant were used for downstream analysis by excluding those with close evolutionary

distance (< 0.001 substitutions per site) from other strains, and the same isolation countries and isolation time (see Table S1). The core genome of ASFV was created by aligning the shredded genomes against the reference strain Georgia-2007 and obtaining the genomic regions mapped by all genomes. Finally, the core genome contains 139,677 base pairs and was used for SNP detection. The bases at all allelic loci for each ASFV genome were concatenated for distance estimation and phylogeny construction using MEGA6 (63) and SplitsTree (64). The pair-wise distance was measured by substitutions per site with the model of maximum composite likelihood and the tree topology was inferred by using the Neighboring-Joining method with a bootstrap value of 1,000. The tree was also constructed using the Maximum Likelihood method. The tree topologies are consistent between different methods. Tajima's *D* value was calculated as defined by Tajima (21).

Detection of functional domains

The functional domains of the genes were detected by searching against the HMM (65) profiles of the PFAM database (25). The hits with score ≥ 20 or *E*-value ≤ 0.003 were considered to be significant and tabulated.

Generation of pan-genome and orthologous groups of ASFV.

The pan-genome of 27 non-redundant ASFV strains was generated using Roary yielding 192 pan-genes encoded by at least one strain of ASFV (66). The amino acid translation of the pan-genes were aligned against each ASFV genome using BLAST tblastn in order to determine the 5'- and 3'-end of the pan-genes in each genome and rescue the genes interrupted by point mutations. Only the genes present in more than 70% of the 27 non-redundant genomes were cataloged into orthologous groups and considered for downstream multiple sequence alignment and positive selection detection. The orthologous groups of MGF genes were refined by stratifying the tandem locations of the paralogous members in each genome to avoid mis-classification given the fact some MGF genes have

higher similarities with paralogs than orthologs. The fusion genes were not considered for further analysis.

Analysis of selection pressures on the ASFV genes

Multiple sequence alignment in amino acids was performed using MUSCLE (67) for orthologs of each gene among the 27 non-redundant strains. The alignments in amino acids were back converted to multiple sequence alignment in nucleotides. All the alignments were manually curated to make the coding sequences in frame. The calculation of non-synonymous substitutions dN and synonymous substitutions dS was based on the Nei & Gorojobri model (26). Likelihood ratio tests (LRT) of selection pressures acting on individual sites of ASFV genes were carried out using PAML with the site-specific model (29, 30). Only the genes with gene-wide average value of dN/dS ≥ 0.2 or mean inter-strain similarity ≤ 90% were selected for the LRTs. For each gene, two LRT tests were conducted, *i.e.*, M2 versus M1 and M8 versus M7. The genes with *p*-value ≤ 0.05 for the test between M8 versus M7 were considered to contain signals with significant positive selection. Only the sites showing positive selection with a posterior probability ≥ 0.9 in M8 were tabulated. The posterior probability was calculated using PAML with the Bayes empirical tests (68). Likelihood ratio tests of genetic diversity and divergent selection of MGF genes were performed using the branch-site Model A in PAML (69). A total of 13 pairs of paralogous members from MGF360 (1L:2L, 1L:3L, 2L:3L, 4L:6L, 8L:10L, 8L:13L, 10L:13L, 9L:11L, 9L:12L, 11L:12L, 14L:16R, the ancestral branch of 1L/2L:3L, and the ancestral branch of 4L/6L:16R) and 13 pairs from MGF505 (1R:4R, 1R:5R, 4R:5R, 2R:4R, 2R:5R, 1R:2R, 2R:10R, 9R:10R, 6R:7R, 6R:9R, 7R:9R, 6R:10R, and 7R:10R) were chosen for LRT of Model A. Either member in the pairs was treated as foreground for the Model A test. The sites under positive selection with a posterior probability ≥ 0.8 and ≥ 0.9 using Bayes empirical tests were identified and mapped to the secondary structure of MGF360 and MGF505, respectively.

Multiple sequence alignments of orthologs and paralogs of the MGF genes

Since sequence similarities between orthologs of MGF genes are much higher than that of paralogs (except MGF360-1L and 2L, MGF505-6R and 7R), we performed multiple sequence alignment in amino acids at first for orthologous members of each paralog of MGF and then for paralogous groups of all MGF360 (except 15R, 18R, 19R, 21R and 22R), or MGF505 (except 3R and 11L due to the high divergence with other paralogs and low reliability of alignment). The alignments in amino acids were back converted to multiple alignments in nucleotides. Profiling of the aligned promoter regions of MGF360 and MGF505 was presented as consensus logo using WebLogo (70).

Secondary structure prediction

The secondary structures of B475L and MGF300-4L were predicted using PSIpred (38), and those of MGF360 and MGF505 using PROMALS3D (71).

Tertiary structure prediction and structure-guided sequence alignment

The tertiary structure of EP402R was modeled using PHYRE server with the structure of human CD2 as template (72). Multiple sequence alignment of EP402R and its homologs in animals, including human CD2 (73) (PDB ID: 1hnf), human CD58 (74) (PDB ID: 1ccz), rat CD2 (75) (PDB ID: 1hng), rat CD48 (76) (PDB ID: 2dru), and boar CD2 (modeled with PHYRE server) was guided by the tertiary structures. The graphical presentation of the alignment was prepared using Espript (77). The structures of the proteins were presented and analyzed using PyMOL (78).

Statistical analysis

The statistical tests used in this study including Hypergeometric test, Mann-Whitney U-test, and T-test, Chi-squared test were performed in the R environment.

Identification of regions with selective sweep using SweepCluster

The population size is highly unbalanced between the two subpopulations α (21 strains) and β

(5 strains), therefore we identified the SNPs associated with between-population subdivision and within-population homogeneity for the clade α and β by selecting loci with the major allele frequency > 85% in clade α and alternative allele frequency > 80% in clade β. The selected SNPs were subject to detection of selective sweep using the newly developed tool SweepCluster. The package SweepCluster is a Python implementation and extension of the clustering algorithm described in (79). It aims to detect the regions with clustered SNPs under selective sweep with the advantage of masking the influence of genomic recombination.

Briefly, a non-synonymous SNP is randomly chosen in a specific gene as the initial cluster assuming that non-synonymous SNPs are more likely to be subject to positive selection than synonymous SNPs or intergenic SNPs. The cluster is then iteratively extended until its spanning range approaches the boundary of the gene or gene operon. If the length of the gene or gene operon is shorter than the specified sweep length, the cluster is further extended to merge the neighboring SNPs or clusters by minimizing the root-mean-square of inter-SNP distances. All the clusters are finally examined and split if any inter-SNP distance within the cluster is longer than the specified distance threshold. The significance of the clustering for each cluster with m distinct SNPs spanning a length of L was evaluated using the gamma distribution with the mean SNP rate μ as the rate parameter under the null hypothesis that the SNPs are randomly and independently distributed on the genome (80):

$$p = \int_0^L \frac{\beta^\alpha}{\Gamma(\alpha)} x^{m-1} e^{-\mu x} \mathrm{d}x$$

The key parameters used in the study are "-max_dist 40 -sweep_lg 520 –min_num 2". The optimal parameters were obtained using the simulation program "sweep_lg_simulation.sh" in the package.

Data availability

The source code of SweepCluster is freely available under the GNU license v3.0 and has

been deposited in GitHub (https://github.com/BaoCodeLab/SweepCluster). The multiple sequence alignments used for selection analysis are available through figshare under the MIT license: https://figshare.com/projects/ASFV_alignment/82718.

Acknowledgements

We thank National Supercomputer Center in Guangzhou for providing partial computing resources. The work was supported by the National Key Research and Development Program of China (2018YFC0840401).

Author contributions

Y.J.B. and H.J.Q. conceived the study. Y.J.B. performed the analysis and wrote the manuscript. J.Q. and Y.J.B. developed and evaluated the computational tool. Y.J.B., F.R. and H.J.Q. analyzed the results and interpreted the data. F.R. and H.J.Q. revised the manuscript critically. All authors wrote and reviewed the manuscript carefully.

Competing interests

The authors declare no competing interests.

References

1. Montgomery R. On a form of swine fever occurring in British East Africa. J Comp Pathol. 1921;34:159-91.
2. Rebecca JR, Vincent M, Livio H, Geoff H, Chris O, Wilna V, et al. African swine fever virus isolate, Georgia, 2007. Emerg Infect Dis. 2008;14(12):1870-4.
3. Sánchez-Cordón PJ, Montoya M, Reis AL, Dixon LK. African swine fever: A re-emerging viral disease threatening the global pig industry. Vet J. 2018;233:41-8.
4. Stokstad E. Deadly virus threatens European pigs and boar. Science. 2017;358(6370):1516-7.
5. Halasa T, Botner A, Mortensen S, Christensen H, Toft N, Boklund A. Simulating the epidemiological and economic effects of an African swine fever epidemic in industrialized swine populations. Vet Microbiol. 2016;193:7-16.
6. de Villiers EP, Gallardo C, Arias M, da Silva M, Upton C, Martin R, et al. Phylogenomic analysis of 11 complete African swine fever virus genome sequences. Virology. 2010;400(1):128-36.
7. Chapman DAG, Tcherepanov V, Upton C, Dixon LK. Comparison of the genome sequences of non-pathogenic and pathogenic African swine fever virus isolates. J Gen Virol. 2008;89(2):397-408.
8. Farlow J, Donduashvili M, Kokhreidze M, Kotorashvili A, Vepkhvadze NG, Kotaria N, et al. Intra-epidemic genome variation in highly pathogenic African swine fever virus (ASFV) from the country of Georgia. Virol J. 2018;15(1):190.
9. Bacciu D, Deligios M, Sanna G, Madrau MP, Sanna ML, Dei Giudici S, et al. Genomic analysis of Sardinian 26544/OG10 isolate of African swine fever virus. Virol Rep. 2016;6:81-9.
10. Dixon LK, Chapman DA, Netherton CL, Upton C. African swine fever virus replication and genomics. Virus Res. 2013;173(1):3-14.
11. Nei M, Gu X, Sitnikova T. Evolution by the birth-and-death process in multigene families of the vertebrate immune system. Proc Natl Acad Sci U S A. 1997;94(15):7799.
12. Rodríguez JM, Moreno LT, Alejo A, Lacasta A, Rodríguez F, Salas ML. Genome sequence of African swine fever virus BA71, the virulent parental strain of the nonpathogenic and tissue-culture adapted BA71V. PLoS One. 2015;10(11):e0142889.
13. Donnell V, Holinka LG, Gladue DP, Sanford B, Krug PW, Lu X, et al. 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. 2015;89(11):6048-56.
14. Michaud V, Randriamparany T, Albina E. Comprehensive phylogenetic reconstructions of African swine fever virus: proposal for a new classification and molecular dating of the virus. PLoS One. 2013;8(7):e69662.
15. Hanada K, Gojobori T, Suzuki Y. A large variation in the rates of synonymous substitution for RNA viruses and its relationship to a diversity of viral infection and transmission modes. Mol Biol Evol. 2004;21(6):1074-80.
16. Drake JW, Hwang CB. On the mutation rate of herpes simplex virus type 1. Genetics. 2005;170(2):969-70.
17. Duffy S, Shackelton LA, Holmes EC. Rates of evolutionary change in viruses: patterns and determinants. Nature reviews Genetics. 2008;9(4):267-76.
18. Croucher NJ, Finkelstein JA, Pelton SI, Mitchell PK, Lee GM, Parkhill J, et al. Population genomics of post-vaccine changes in pneumococcal epidemiology. Nat Genet. 2013;45(6):656-63.
19. Bishop RP, Fleischauer C, de Villiers EP, Okoth EA, Arias M, Gallardo C, et al. Comparative analysis of the complete genome sequences of Kenyan African swine fever virus isolates within p72 genotypes IX and X. Virus Genes. 2015;50(2):303-9.
20. Masembe C, Sreenu VB, Da Silva Filipe A, Wilkie GS, Ogweng P, Mayega FJ, et al. Genome sequences of five African swine fever virus genotype IX isolates from domestic pigs in Uganda. Microbiol Resour Announc. 2018;7(13):e01018-18.
21. Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics. 1989;123(3):585-95.
22. Watterson GA. On the number of segregating sites in genetical models without recombination. Theor Popul Biol. 1975;7(2):256-76.
23. Atuhaire DK, Afayoa M, Ochwo S, Mwesigwa S, Okuni JB, Olaho-Mukani W, et al. Molecular characterization and phylogenetic study of African swine fever virus isolates from recent outbreaks in Uganda (2010-2013). Virol J. 2013;10:247.
24. Krug PW, Holinka LG, Donnell V, Reese B, Sanford B, Fernandez-Sainz I, et al. The Progressive adaptation of a Georgian isolate of African swine fever virus to Vero cells leads to a gradual attenuation of virulence in swine corresponding to major modifications of the viral genome. J Virol. 2015;89(4):2324-32.
25. Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, et al. The Pfam protein families database. Nucleic Acids Res. 2012;40(Database issue):D290-D301.

26. Nei M, Gojobori T. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol. 1986;3(5):418-26.
27. Ruiz-Gonzalvo F, Rodriguez F, Escribano JM. Functional and immunological properties of the baculovirus-expressed hemagglutinin of African swine fever virus. Virology. 1996;218(1):285-9.
28. Galindo I, Almazán F, Bustos MJ, Viñuela E, Carrascosa AL. African swine fever virus EP153R open reading frame encodes a glycoprotein involved in the hemadsorption of infected cells. Virology. 2000;266(2):340-51.
29. Yang Z. PAML 4: Phylogenetic analysis by maximum likelihood. Mol Biol Evol. 2007;24(8):1586-91.
30. Yang Z, Nielsen R. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Mol Biol Evol. 2002;19(6):908-17.
31. Borca MV, Carrillo C, Zsak L, Laegreid WW, Kutish GF, Neilan JG, et al. Deletion of a CD2-like gene, 8-DR, from African swine fever virus affects viral infection in domestic swine. J Virol. 1998;72(4):2881-9.
32. Alejo A, Matamoros T, Guerra M, Andrés G. A proteomic atlas of the African swine fever virus particle. J Virol. 2018;92(23):e01293-18.
33. Burmakina G, Malogolovkin A, Tulman ER, Zsak L, Delhon G, Diel DG, et al. African swine fever virus serotype-specific proteins are significant protective antigens for African swine fever. J Gen Virol. 2016;97(7):1670-5.
34. Monteagudo PL, Lacasta A, López E, Bosch L, Collado J, Pina-Pedrero S, et al. BA71ΔCD2: a new recombinant live attenuated African swine fever virus with cross-protective capabilities. J Virol. 2017;91(21):e01058-17.
35. Davis SJ, Ikemizu S, Wild MK, van der Merwe PA. CD2 and the nature of protein interactions mediating cell-cell recognition. Immunological reviews. 1998;163:217-36.
36. Morea V, Lesk AM, Tramontano A. Antibody modeling: implications for engineering and design. Methods. 2000;20(3):267-79.
37. Argilaguet JM, Pérez-Martín E, Nofrarías M, Gallardo C, Accensi F, Lacasta A, et al. DNA vaccination partially protects against African swine fever virus lethal challenge in the absence of antibodies. PloS ONE. 2012;7(9):e40942.
38. Buchan DWA, Jones DT. The PSIPRED Protein Analysis Workbench: 20 years on. Nucleic Acids Res. 2019;47(W1):W402-w7.
39. Stephan W. Selective Sweeps. Genetics. 2019;211(1):5.
40. Jancovich JK, Chapman D, Hansen DT, Robida MD, Loskutov A, Craciunescu F, et al. Immunization of pigs by DNA prime and recombinant vaccinia virus boost to identify and rank African swine fever virus immunogenic and protective proteins. J Virol. 2018;92(8):e02219-17.
41. Netherton CL, Goatley LC, Reis AL, Portugal R, Nash RH, Morgan SB, et al. Identification and immunogenicity of African swine fever virus antigens. Front Immunol. 2019;10:1318.
42. Lopera-Madrid J, Osorio JE, He Y, Xiang Z, Adams LG, Laughlin RC, et al. Safety and immunogenicity of mammalian cell derived and Modified Vaccinia Ankara vectored African swine fever subunit antigens in swine. Vet Immunol Immunopathol. 2017;185:20-33.
43. Dixon LK, Islam M, Nash R, Reis AL. African swine fever virus evasion of host defences. Virus Res. 2019;266:25-33.
44. Leitao A, Cartaxeiro C, Coelho R, Cruz B, Parkhouse RM, Portugal F, et al. The non-haemadsorbing African swine fever virus isolate ASFV/NH/P68 provides a model for defining the protective anti-virus immune response. J Gen Virol. 2001;82(Pt 3):513-23.
45. Portugal R, Coelho J, Höper D, Little NS, Smithson C, Upton C, et al. Related strains of African swine fever virus with different virulence: genome comparison and analysis. J Gen Virol. 2015;96(2):408-19.
46. Hurtado C, Granja AG, Bustos MaJ, Nogal MaL, González de Buitrago G, de Yébenes VG, et al. The C-type lectin homologue gene (EP153R) of African swine fever virus inhibits apoptosis both in virus infection and in heterologous expression. Virology. 2004;326(1):160-70.
47. Afonso CL, Piccone ME, Zaffuto KM, Neilan J, Kutish GF, Lu Z, et al. African swine fever virus multigene family 360 and 530 genes affect host interferon response. J Virol. 2004;78(4):1858-64.
48. Zsak L, Lu Z, Burrage TG, Neilan JG, Kutish GF, Moore DM, et al. African swine fever virus multigene family 360 and 530 genes are novel macrophage host range determinants. J Virol. 2001;75(7):3066-76.
49. Mosavi LK, Cammett TJ, Desrosiers DC, Peng Z-Y. The ankyrin repeat as molecular architecture for protein recognition. Prot Sci. 2004;13(6):1435-48.
50. Herbert MH, Squire CJ, Mercer AA. Poxviral ankyrin proteins. Viruses. 2015;7(2):709-38.
51. Bradley RR, Terajima M. Vaccinia virus K1L protein mediates host-range function in RK-13 cells via ankyrin repeat and may interact with a cellular GTPase-activating protein. Virus Res. 2005;114(1-2):104-12.
52. Li Y, Meng X, Xiang Y, Deng J. Structure function studies of vaccinia virus host range protein k1 reveal a novel functional surface for ankyrin repeat proteins. J Virol. 2010;84(7):3331-8.
53. Rodriguez JM, Salas ML. African swine fever virus transcription. Virus Res. 2013;173(1):15-28.
54. Hurtado C, Bustos MJ, Granja AG, de Leon P, Sabina P, Lopez-Vinas E, et al. The African swine fever virus

lectin EP153R modulates the surface membrane expression of MHC class I antigens. Arch Virol. 2011;156(2):219-34.
55. Powell PP, Dixon LK, Parkhouse RM. An IkappaB homolog encoded by African swine fever virus provides a novel mechanism for downregulation of proinflammatory cytokine responses in host macrophages. J Virol. 1996;70(12):8527-33.
56. Salguero FJ, Gil S, Revilla Y, Gallardo C, Arias M, Martins C. Cytokine mRNA expression and pathological findings in pigs inoculated with African swine fever virus (E-70) deleted on A238L. Vet Immunol Immunopathol. 2008;124(1-2):107-19.
57. Zhang F, Hopwood P, Abrams CC, Downing A, Murray F, Talbot R, et al. Macrophage transcriptional responses following in vitro infection with a highly virulent African swine fever virus isolate. J Virol. 2006;80(21):10514-21.
58. Correia S, Ventura S, Parkhouse RM. Identification and utility of innate immune system evasion mechanisms of ASFV. Virus Res. 2013;173(1):87-100.
59. Alcami A, Angulo A, Vinuela E. Mapping and sequence of the gene encoding the African swine fever virion protein of M(r) 11500. J Gen Virol. 1993;74 (Pt 11):2317-24.
60. Malogolovkin A, Burmakina G, Tulman ER, Delhon G, Diel DG, Salnikov N, et al. African swine fever virus CD2v and C-type lectin gene loci mediate serological specificity. The Journal of general virology. 2015;96(Pt 4):866-73.
61. O'Donnell V, Holinka LG, Sanford B, Krug PW, Carlson J, Pacheco JM, et al. African swine fever virus Georgia isolate harboring deletions of 9GL and MGF360/505 genes is highly attenuated in swine but does not confer protection against parental virus challenge. Virus Res. 2016;221:8-14.
62. Memoli MJ, Jagger BW, Dugan VG, Qi L, Jackson JP, Taubenberger JK. Recent human influenza A/H3N2 virus evolution driven by novel selection factors in addition to antigenic drift. J Infect Dis. 2009;200(8):1232-41.
63. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. Mol Biol Evol. 2013;30(12):2725-9.
64. Huson DH, Bryant D. Application of phylogenetic networks in evolutionary studies. Mol Biol Evol. 2006;23(2):254-67.
65. Eddy SR. Accelerated Profile HMM Searches. PLoS Comput Biol. 2011;7(10):e1002195.
66. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG, et al. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics (Oxford, England). 2015;31(22):3691-3.
67. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32(5):1792-7.
68. Yang Z, Wong WSW, Nielsen R. Bayes empirical bayes inference of amino acid sites under positive selection. Mol Biol Evol. 2005;22(4):1107-18.
69. Zhang J, Nielsen R, Yang Z. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol Biol Evol. 2005;22(12):2472-9.
70. Crooks GE, Hon G, Chandonia J-M, Brenner SE. WebLogo: a sequence logo generator. Genome Res. 2004;14(6):1188-90.
71. Pei J, Kim BH, Grishin NV. PROMALS3D: a tool for multiple protein sequence and structure alignments. Nucleic Acids Res. 2008;36(7):2295-300.
72. Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJE. The Phyre2 web portal for protein modeling, prediction and analysis. Nat Protoc. 2015;10(6):845-58.
73. Bodian DL, Jones EY, Harlos K, Stuart DI, Davis SJ. Crystal structure of the extracellular region of the human cell adhesion molecule CD2 at 2.5 A resolution. Structure. 1994;2(8):755-66.
74. Ikemizu S, Sparks LM, van der Merwe PA, Harlos K, Stuart DI, Jones EY, et al. Crystal structure of the CD2-binding domain of CD58 (lymphocyte function-associated antigen 3) at 1.8-A resolution. Proc Natl Acad Sci U S A. 1999;96(8):4289-94.
75. Jones EY, Davis SJ, Williams AF, Harlos K, Stuart DI. Crystal structure at 2.8 A resolution of a soluble form of the cell adhesion molecule CD2. Nature. 1992;360(6401):232-9.
76. Evans EJ, Castro MA, O'Brien R, Kearney A, Walsh H, Sparks LM, et al. Crystal structure and binding properties of the CD2 and CD244 (2B4)-binding protein, CD48. J Biol Chem. 2006;281(39):29309-20.
77. Robert X, Gouet P. Deciphering key features in protein structures with the new ENDscript server. Nucleic Acids Res. 2014;42(Web Server issue):W320-4.
78. Benoit M, Desnues B, Mege JL. Macrophage Polarization in Bacterial Infections. J Immunol. 2008;181:3733-9.
79. Bao Y-J, Shapiro BJ, Lee SW, Ploplis VA, Castellino FJ. Phenotypic differentiation of Streptococcus pyogenes populations is induced by recombination-driven gene-specific sweeps. Sci Rep. 2016;6:36644.
80. Sinnett D, Beaulieu P, Belanger H, Lefebvre JF, Langlois S, Theberge MC, et al. Detection and characterization of DNA variants in the promoter regions of hundreds of human disease candidate genes. Genomics. 2006;87(6):704-10.

Table 1. ASFV genes enriched with non-synonymous mutations.

Gene name	Non-synonymous mutation counts	Gene length	*p*-value	*p*-value corrected	Gene function	Functional category
MGF300-4L	116	993	<1E-20	<1E-20	MGF300-4L	Multigene family
MGF300-1L	74	807	2.57E-09	2.35E-08	MGF300-1L	Multigene family
MGF505-4R	274	1521	<1E-20	<1E-20	MGF505-4R	Multigene family
MGF505-5R	186	1497	<1E-20	<1E-20	MGF505-5R	Multigene family
MGF505-6R	99	1578	0.0002	0.00115	MGF505-6R	Multigene family
MGF505-9R	192	1521	<1E-20	<1E-20	MGF505-9R	Multigene family
MGF505-10R	145	1629	<1E-20	<1E-20	MGF505-10R	Multigene family
MGF505-11L	128	1629	1.99E-10	2.13E-9	MGF505-11L	Multigene family
MGF360-8L	118	960	<1E-20	<1E-20	MGF360-8L	Multigene family
MGF360-15R	75	870	2.71E-08	1.92E-07	MGF360-15R	Multigene family
MGF360-16R	93	930	2.08E-13	2.67E-12	MGF360-16R	Multigene family
A151R	85	477	<1E-20	<1E-20	CXXC-motif containing protein	Involved in redox pathway
I215L	78	639	<1E-20	<1E-20	Ubiquitin-conjugation enzyme	Shuttles between the nucleus and cytoplasm
I196L	72	609	3.51E-14	4.99E-13	Uncharacterized protein	
I177L	31	201	1.14E-09	1.12E-08	Uncharacterized protein	
DP238L	68	717	2.88E-09	2.46E-08	Uncharacterized protein	
H240R	68	726	4.76E-09	3.81E-08	Uncharacterized protein	
K205R	59	618	2.42E-08	1.82E-07	Uncharacterized protein	
E183L/P54	49	555	3.25E-06	2.08E-05	Structural protein p54	Structural protein
A240L	58	711	5.17E-06	5.15E-05	Thymidylate kinase	Nucleotide metabolism
EP364R	79	1110	1.94E-05	1.13E-04	ERCC4 domain	DNA replication and repair
I267L	61	840	9.21E-05	5.13E-04	RING finger containing protein	
CP312R	65	924	1.40E-04	7.33E-04	Uncharacterized protein	
A137R/P11.5	35	414	1.80E-04	8.98E-04	Structural protein P11.5	Structural protein
I329L	68	990	1.90E-04	9.54E-4	Transmembrane protein	Host-cell interactions

Note: The enrichment *p*-value for each gene was calculated with Hypergeometric test and the multiple testing correction was determined using the Benjamini-Hochberg procedure. The enrichment with corrected *p*-value < 0.001 is considered to be significant.

Table 2. Novel candidates with positive selection signals at a fraction of sites with ω (dN/dS) >1 based on the likelihood ratio tests.

Gene	*p*-value	# of sites	Function
pp220/CP2475L	< 1E-20	25	Structural polyprotein precursor (core shell)
B602L	2.2E-05	4	Chaperone protein of P72
MGF300-4L	0.004	9	Multigene family 300
J5R/H108R	0.001	1	Structural protein (inner envelop)
P11.5/A137R	0.006	4	Structural protein (virus factories)
p10/K78R	0.022	4	DNA-binding structural protein (viral nucleoid)
A240L	0.002	2	Thymidylate kinase
Q706L	0.034	1	Helicase superfamily II
B117L	1.1E-05	3	Uncharacterized protein
86R	4.6E-05	8	Uncharacterized protein
B475L	0.005	14	Uncharacterized protein
L60L	0.023	3	Uncharacterized protein

Note: # of sites indicates the number of sites in the specific gene under positive selection with a posterior probability ≥ 0.9 using Bayes empirical tests.

Table 3. Gene regions with significant selective sweep.

Genomic location		# of SNPs	Sweep length	Gene location			*p*-value corrected	Function
Start	End			Gene	Start	End		
Genes known to be involved in host cell interaction								
176588	177082	56	494	MGF360-16R	-1	493	<1E-20	Multigene family 360
42644	42992	55	349	MGF505-9R	12	360	<1E-20	Multigene family 505
43166	43394	28	229	MGF505-9R	534	762	<1E-20	Multigene family 505
43580	43775	24	196	MGF505-9R	948	1143	<1E-20	Multigene family 505
44899	45466	50	568	MGF505-10R	336	903	<1E-20	Multigene family 505
37872	38331	46	460	MGF505-5R	528	987	<1E-20	Multigene family 505
38379	38643	29	265	MGF505-5R	1035	1299	<1E-20	Multigene family 505
37356	37560	25	205	MGF505-5R	12	216	<1E-20	Multigene family 505
49855	50151	30	297	MGF360-15R	490	786	<1E-20	Multigene family 360
178236	178537	29	302	MGF505-11L	819	1120	<1E-20	Multigene family 505
36697	36997	25	301	MGF505-4R	902	1202	0.0150	Multigene family 505
37041	37193	17	153	MGF505-4R	1246	1398	0.0087	Multigene family 505
23397	23606	20	210	MGF360-8L	396	605	0.0145	Multigene family 360
173990	174197	21	208	I215L	234	441	0.0040	Ubiquitin conjugating enzyme
46308	46684	29	377	A224L	300	676	0.0145	IAP apoptosis inhibitor
Novel candidate genes								
150420	150855	46	436	P1192R	2890	3325	<1E-20	DNA topoisomerase type II
150185	150371	18	187	P1192R	2655	2841	0.0318	DNA topoisomerase type II
22021	22360	35	340	MGF300-4L	570	909	<1E-20	Multigene family 300
48674	49031	34	358	A151R	24	381	<1E-20	Involved in redox pathway
58166	58389	32	224	F778R	1167	1390	<1E-20	Ribonucleotide reductase
175802	176124	31	323	DP238L	290	612	<1E-20	Uncharacterized protein
156642	156932	30	291	R298L	22	312	<1E-20	Serine protein kinase
63391	63588	27	198	K205R	199	396	<1E-20	Uncharacterized protein
119386	119642	27	257	CP2475L	5049	5305	<1E-20	Structural polyprotein precursor
160977	161318	31	342	QP383R	453	794	0.0006	Nif S-like protein
161389	161625	23	237	QP383R	865	1101	0.0029	Nif S-like protein
62145	62415	25	271	F1055L	585	855	0.0029	Helicase superfamily II
165252	165489	23	238	E146L	120	357	0.0029	Uncharacterized protein
170094	170377	24	284	I267L	68	351	0.0168	RING finger containing protein
127731	127897	20	167	CP312R	447	613	0.0006	Uncharacterized protein
67870	68014	19	145	EP1242L	2229	2373	<1E-20	RNA polymerase subunit 2
47935	48092	18	158	A240L	273	430	0.0035	Thymidylate kinase

Note: The significant sweeping regions should satisfy two thresholds: multiple testing corrected *p*-value ≤ 0.05 and the number of SNPs ≥ 18 in each region. The multiple testing corrected *p*-value was determined using the Bonferroni procedure.

Figure legends

Figure 1. Phylogenetic tree and geographical distribution of ASFV strains.

(a) Phylogeny built from the core genome of 27 non-redundant ASFV strains. (b) Phylogeny built from the full-length structural gene p72 (B646L) of the 27 non-redundant ASFV genomes. The subtypes are shown on the right. (c) Geographical distribution of 85 non-redundant ASFV isolates and the phylogeny constructed using the C-terminal 414 bp of p72 sequences available from public databases. The partial p72 sequences of the 85 non-redundant ASFV isolates with unique geographical location and isolate time were compiled from the NCBI database https://www.ncbi.nlm.nih.gov/ and mapped to the geographical locations. The trees were inferred using the Neighboring-Joining method with 1000 bootstrap. The trees built from all three datasets forms three major clades α, β, and γ indicated on the corresponding branches.

Figure 2. Genetic and functional properties of genes with positive diversifying selection signals.

(a) The genes containing sites under positive diversifying selection (p-value ≤ 0.05). Top panel: the genomic locations of the genes. Bottom panel: histogram representation of the number of sites with significant selection in each gene (posterior probability ≥ 0.9). (b,c) Layout of the positively selected sites on the domain architectures of the key genes known to be relevant to host interactions (b) and of novel candidate genes with unknown host interactions (c). The positively selected sites (in black triangles) of EP402R, EP153R, MGF505-4R, B475L, and B117L are largely located in the variable regions or near around short repeat-rich regions (arrows, with blue ones for putative *N*-linked glycosylation sites). The functional domains are represented as colored bars and the transmembrane domains as directed frames pointing towards outside of the membrane. The active sites are shown as diamonds. The red bars show overlapping regions with signatures of selective sweep. The lengths of the proteins might be longer than the actual length due to gaps induced by multiple alignments. The length of the protein CP2475L is in a shrunk scale due to its exceptionally large size. Abbreviations: DXQNT: DXQNT repeats; TM: transmembrane domain; P-rich repeat: proline-rich repeats; ANK: ankyrin repeat; UQ_con: ubiquitin-conjugating enzyme; H-rep: histidine-rich repeats; Colicin-V: Colicin-V production domain; SP-like: signal peptide-like domain; Thymidylate_kin: thymidylate kinase domain; bZIP_1: basic leucine zipper domain; Viral polyN: viral polyprotein N-terminal domain. (d) Multiple sequence alignment of the extracellular Ig-like domain of EP402R and its homologs in rat (CD2, CD48), human (CD2, CD58), and boar (CD2). The secondary structure of rat CD2 is displayed on the top of the alignment with β strands in arrows and β turns in TT. The known ligand-binding sites of CD2, CD48, and CD58 are highlighted in yellow and the positively selected sites in EP402R are in green (posterior probability ≥ 0.9) or light green (posterior probability ≥ 0.8). Two known epitopes F3 and A6 in ASFV strain Spain-E75 are framed in cyan boxes.

Figure 3. Genomic distribution and genetic properties of genes with signatures of selective sweep.

(a) Distribution of population differentiation F_{st} and diversity π of a series of 100-loci sliding windows from three groups of SNPs: associated with between-population subdivision, not associated with between-population subdivision, and all detected SNPs. The between-group differences were evaluated using wilcoxon rank sum test and the *p*-values were indicated for the comparison between associated SNPs and the other two groups. (b) Venn diagram of number of genes with putative diversifying selection and selective sweep. (c) Significance of regions with signatures of selective sweep as shown with gradient colors. The height of bars shows the number of SNPs in the sweeping regions and the width shows the spanning length of the sweeping regions. (d) Between-population differentiation F_{st} (in magenta) and within-population diversity π (in blue for the clade α and cyan for the clade β) of six representative genes containing regions with putative selective sweep as shown with red bars. Only the sweeping regions longer than 135bp and Bonferroni-corrected *p*-value ≤ 0.05 were considered significant and indicated. The regions show higher between-population differentiation and reduced within-population diversity in comparison with the nearby regions. The scale for the between-population differentiation is shown on the left axis and the within-population diversity on the right axis.

Figure 4. Presence frequencies and sequence divergence of the genes with signatures of diversifying selection and those with selective sweep.

(a) The genes with signals of diversifying selection in this study and known to be involved in host interactions. (b) The genes with selective sweep in this study and known to be involved in host interactions. (c) The genes lost in avirulent strains without significant diversifying selection or selective sweep. (d) The novel candidate genes with diversifying selection signals. (e) The novel candidate genes with selective sweep signals (f) The non-antigenic conserved structural proteins. (g) The antigen proteins eliciting immunological responses in immunoassay experiments. (h) The genes known to be involved in interactions with host cell components. (i) Mann-Whitney U-test of amino acid divergence between any two groups of genes above. For each gene, the mean amino acid divergence among the ASFV strains was used as the proxy for the test. The presence frequency was calculated as the percentage of presence of each gene within the 27 non-redundant ASFV strains and represented as colored bars. The sequence divergence was evaluated as pair-wise amino acid differences displayed as jitter plots. The average of pair-wise divergence for each gene is indicated with grey diamond. The names of MGF genes ignore “MGF” for figure compactness.

Figure 5. Schematic representation of gene presence/absence among virulent and avirulent strains.

Three representative virulent strains and all three avirulent strains are shown. Genes are indicated as colored frames. (a) Gene organization of the locus of EP153R and EP402R. The two genes are interrupted by single nucleotide indels in the avirulent strains Portugal-NHV68 and Portugal-OURT88. EP153R is interrupted by

a nucleotide deletion of "A" at the 47th bp of the gene, whereas EP402R is repeatedly interrupted by three distinct nucleotide deletions (one "T" at 32th bp, one "T" at 744th bp, and one "A" at 908th bp). The sites with indels are indicated with red stars. The kinked C-terminus depicts the truncation of the genes by indels. (b) Pattern of gene presence/absence at the locus of MGF. There is a large fragmental deletion in all three avirulent strains Spain-BA71V, Portugal-NHV68 and Portugal-OURT88, removing multiple genes in MGF360 and MGF505. Abbreviations for the strain names are used for figure compactness.

Figure 6. Genetic diversity among paralogs of MGF360 and MGF505.

(a,c) Divergent selection between paralogous pairs of genes/branches of MGF360 and MGF505 mapping to the phylogenetic structure. The phylogenetic trees were inferred using Neighbor-Joining method with 1000 bootstraps. The branches containing orthologous members of each paralog are collapsed indicated with triangle. The exceptions are three isolates of MGF360-1L (Kenya-1950, Ken05-Tk1, and Spain-E75), which cluster together with MGF360-2L, and five isolates of MGF505-7R (Malawi-Lil83, Kenya-1950, Ken05-Tk1, Ken06-Bus, and UgandaN10-2015), which cluster together with MGF505-6R. The pairs of genes/branches used for LRTs are connected by frame lines with blue arrows indicating the gene/branch under positive selection at a fraction of sites and grey lines indicating no significant positive selection detected in either of the gene/branch. (b,d) Divergent promoter regions from -55 to -1 upstream translational start sites of MGF360 and MGF505. Profiling of promoter regions is presented as consensus logo for each orthologous group. The sequences with common signatures are highlighted with underline and the potential 5-nucleotide promoter motifs with double underline. (e,f) Divergence of the promoter regions (y axis) against the synonymous substitution rate (x axis) for each pair of genes/branches in MGF360 (e) and MGF505 (f). The fitted lines of linear regression are shown in red and the fitting equation and Pearson correlations R^2 are indicated.

Figure 7. Distribution and enrichment of sites identified as being under divergent selection in LRTs between paralogous pairs of genes/branches of MGF360 (a) and MGF505 (b).

Only the sites with a posterior probability ≥ 0.8 in MGF360 and ≥ 0.9 in MGF505 are shown (colored pentagons). Either of the partners in the pairs was treated as foreground in LRTs (indicated in the parentheses). The sites are mapped to the predicted secondary structure of MGF360 and MGF505, respectively. The secondary structures are represented as α-helices (cylinders), β-strands (arrows), or coiled loops (lines). A 25-codon sliding window plot of the site density (sites per window) is shown as dotted grey lines. The *p*-value of enrichment was calculated with the Hypergeometric test for each 25-codon window and the consecutive windows with *p*-value ≤ 0.05 were merged to a single enriched region indicated with horizontal bars.

Figure 8. The integrated scheme of interactions between ASFV genes with signatures of diversifying selection/selective sweep and host components.

The interactions are depicted in the framework of the virus life cycles and host defense processes. The ASFV-encoded proteins are associated with different parts of the viral particle or released at different stages of the infection cycle (purple ovals). They interact with host cells *via* DNA-binding, surface adhesion, inhibition, or activation. The host cell is bounded with membrane indicated with the round soft edge. Host-encoded proteins are shown as aqua squares. ASFV-encoded proteins with unknown function or expression time are shown as grey ovals outside of the membrane. Not all members of MGF360 or MGF505 are involved in the interactions. Key host molecules affected by ASFV, such as NF-κB, IFN, TNF-α, and ISGs are shown in red. Other abbreviations: TNFR: TNF receptor; IFNR: IFN receptor; Viral DNA PRR: viral DNA pattern recognition receptor; ISGF: IFN-stimulated gene factor; ISGs: IFN-stimulated genes; ISRE: IFN-stimulated response elements; RBCs: red blood cells.

Supplementary materials

Table S1. Information of ASFV isolates with complete genomic sequences.

Table S2. List of unique non-synonymous mutations in the non-pathogenic strains.

Table S3. Functional domain identification of the genes enriched with non-synonymous mutations.

Table S4. Genes with the value of dN/dS lower than the average (dN/dS < 0.1) using the Nei & Gojobori method.

Table S5. Genes with positive selection signals at a fraction of sites with ω (dN/dS) >1 based on the likelihood ratio tests.

Table S6. Three categories of proteins used for comparison of sequence variability.

Table S7. Pairs of paralogous genes/branches of MGF360 and MGF505 showing divergent selection at a fraction of sites based on the likelihood ratio tests of Model A of PAML.

Figure S1. Phylogenetic structure constructed from the C-terminal 414 bp of the structural gene p72 and presented in a dendrogram tree

The isolates were compiled from the NCBI database https://www.ncbi.nlm.nih.gov. A total of 85 non-redundant isolates were obtained with unique geographical location and isolate time and were used for tree construction. The tree was inferred using the Neighboring-Joining method with 1000 bootstrap. The isolate names were presented as the combination of accession number, location, time and genotype.

Figure S2. Profiling of the distribution of non-synonymous mutations along the ASFV genomes.

(a) The density distribution (number of mutations per kb) of non-synonymous mutations along the genome of the representative strain Georgia-2007. The top genes with the highest density of non-synonymous mutations are indicated. (b) All genes enriched with non-synonymous mutations (q-value ≤ 0.001) but not with synonymous mutations (q-value > 0.05) are shown in blue dots. The genes enriched with synonymous

mutations (q-value ≤ 0.05) but not with non-synonymous mutations (q-value > 0.05) are shown in red dots. The genes are not enriched with either mutations are in black dots. The q-value is defined as the multiple testing corrected p-value using the Benjamini-Hochberg procedure. The p-value was calculated with the Hypergeometric test. (c) A detailed view of the density distribution of non-synonymous mutations for three top genes is depicted along the domain architecture of the genes. There is no significant difference of the mutation distribution between different functional domains.

Figure S3. The structural mapping of the positively selected sites of EP402R and comparison with key sites in CD2 homologs.

(a) The positively selected sites in EP402R mapped to the modeled structure of EP402R. Both C-set and V-set domain are shown. (b) The ligand-binding sites of human CD2 mapped to the V-set domain in the structure (PDB ID: 1hnf). (c) The ligand-binding sites of rat CD2 mapped to the V-set domain in the structure (PDB ID: 1hng). The sites are shown as colored sticks with positive-charged residues in blue, negative-charged residues in red, polar residues in magenta, and hydrophobic residues in yellow. (d) Superposition of the V-set domain of the structure of EP402R, human CD2, and rat CD2. Three proteins share a similar V-set domain structure forming a globular fold with two β-sheets. (e) Two known epitopes F3 and A6 in EP402R showing high divergence among ASFV strains. The positively selected site E157 in A6 is indicated in black triangle. The strain Portugal-L60 has a deletion at the location of A6. The truncation of EP402R by deleted nucleotides in Portugal-OURT88 and Portugal-NHV68 was recovered to obtain the normally translated epitope sequences.

Figure S4. The predicted secondary structures of B475L (a) and MGF300-4L (b).

The secondary structures are represented as α-helices (cylinders), β-strands (arrows), or coiled loops (lines). Both proteins are predominated by α-helices.

Figure S5. The phylogeny and heatmap of pair-wise nucleotide similarities of the orthologous and paralogous genes of MGF360 (a) and MGF505 (b).

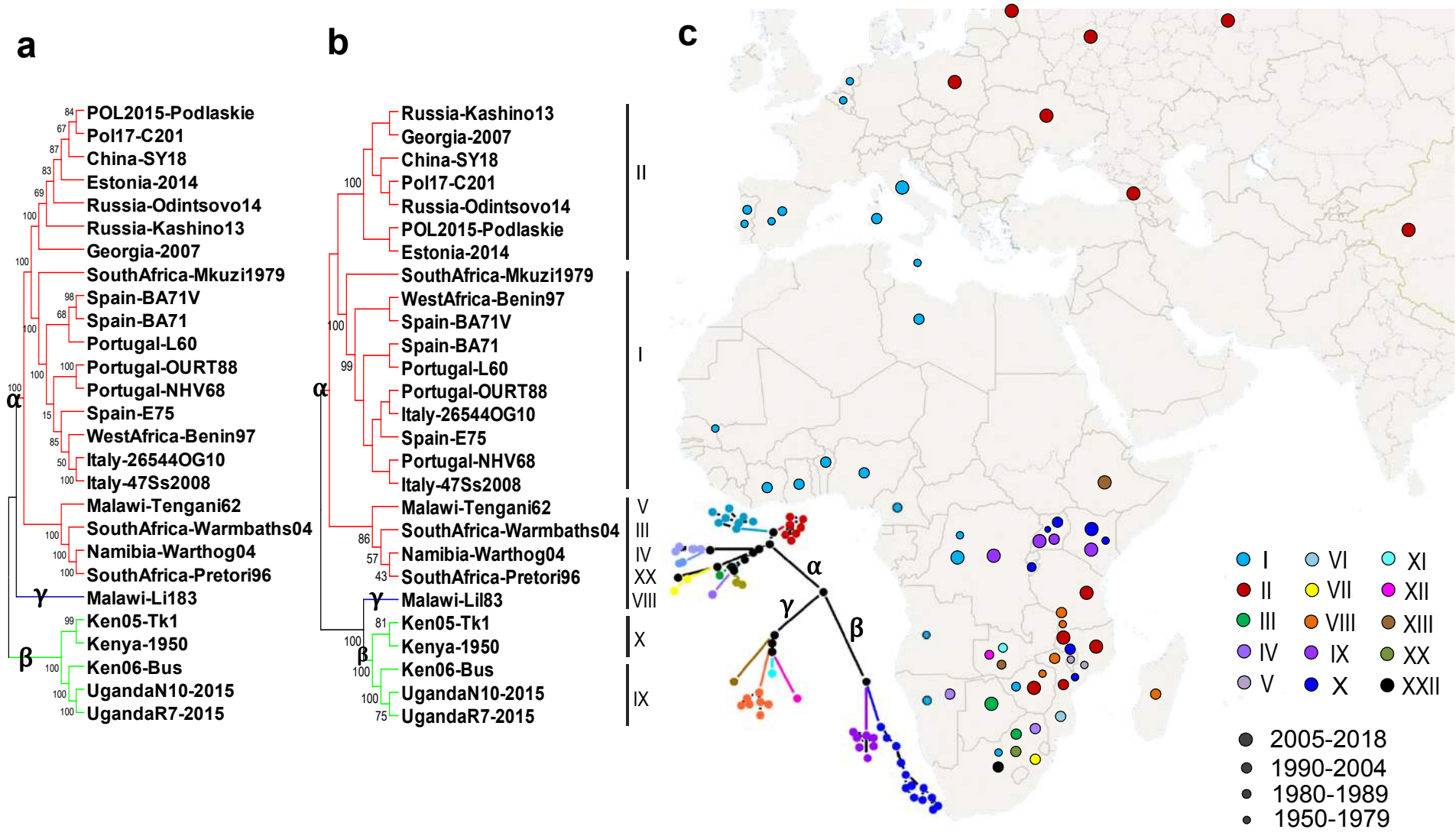
The phylogenetic structure was inferred using Neighbor-Joining method with 1000 bootstraps. Only nodes with the support value > 30 are shown. A colored scale for the nucleotide similarities is given on the right side of the heatmap. The similarities between orthologous genes are much higher than that for paralogous genes, and therefore the former cluster together in the trees, except three isolates of MGF360-1L (Kenya-1950, Ken05-Tk1, and Spain-E75), which cluster together with MGF360-2L, and five isolate of MGF505-7R (Malawi-Lil83, Kenya-1950, Ken05-Tk1, Ken06-Bus, and UgandaN10-2015), which cluster together with MGF505-6R.

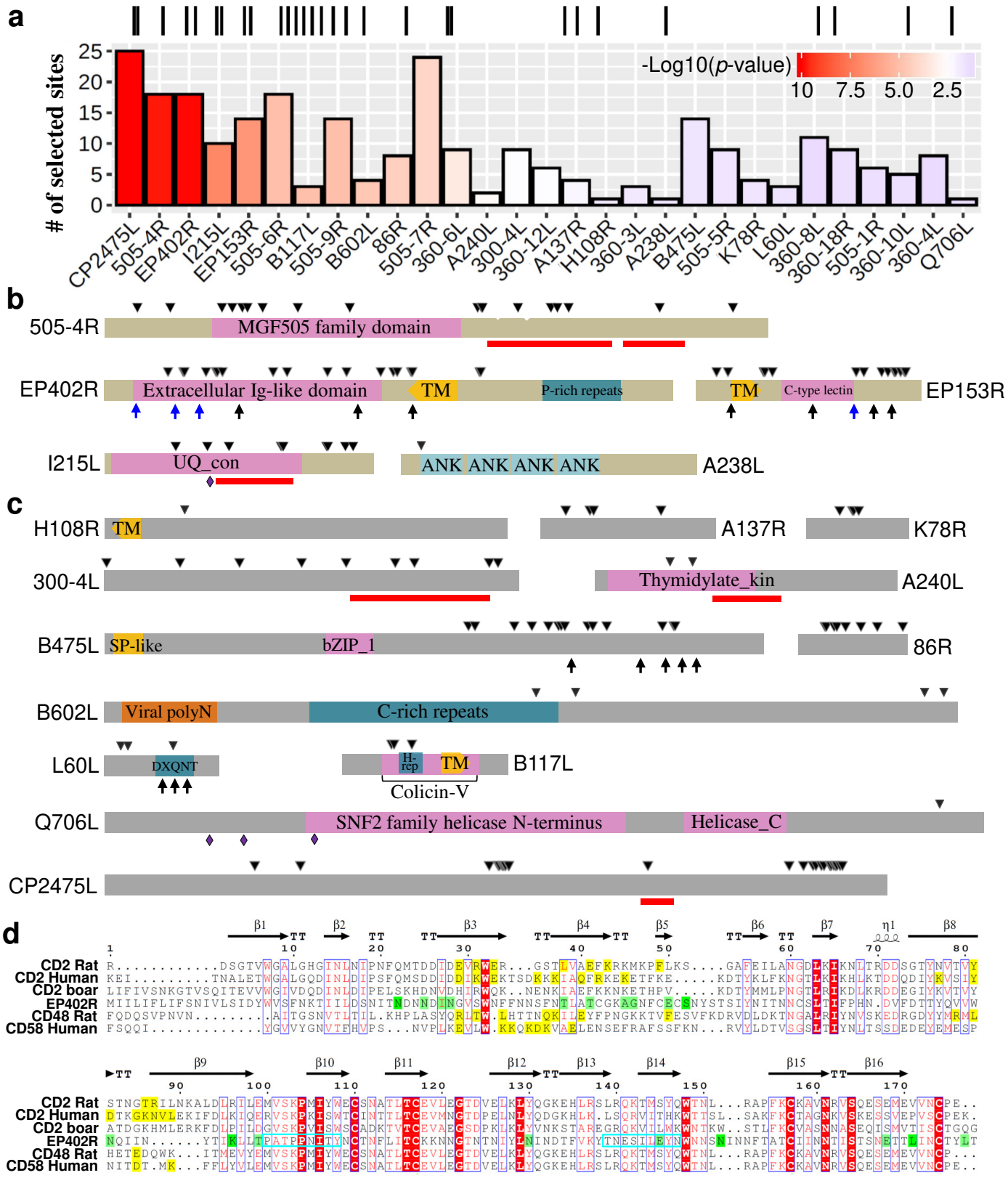
Supplementary file 1: Related to Figure 1c. Information of the 85 ASFV strains with available p72 sequences. The C-terminal 414 bp of p72 was used for phylogeny construction.

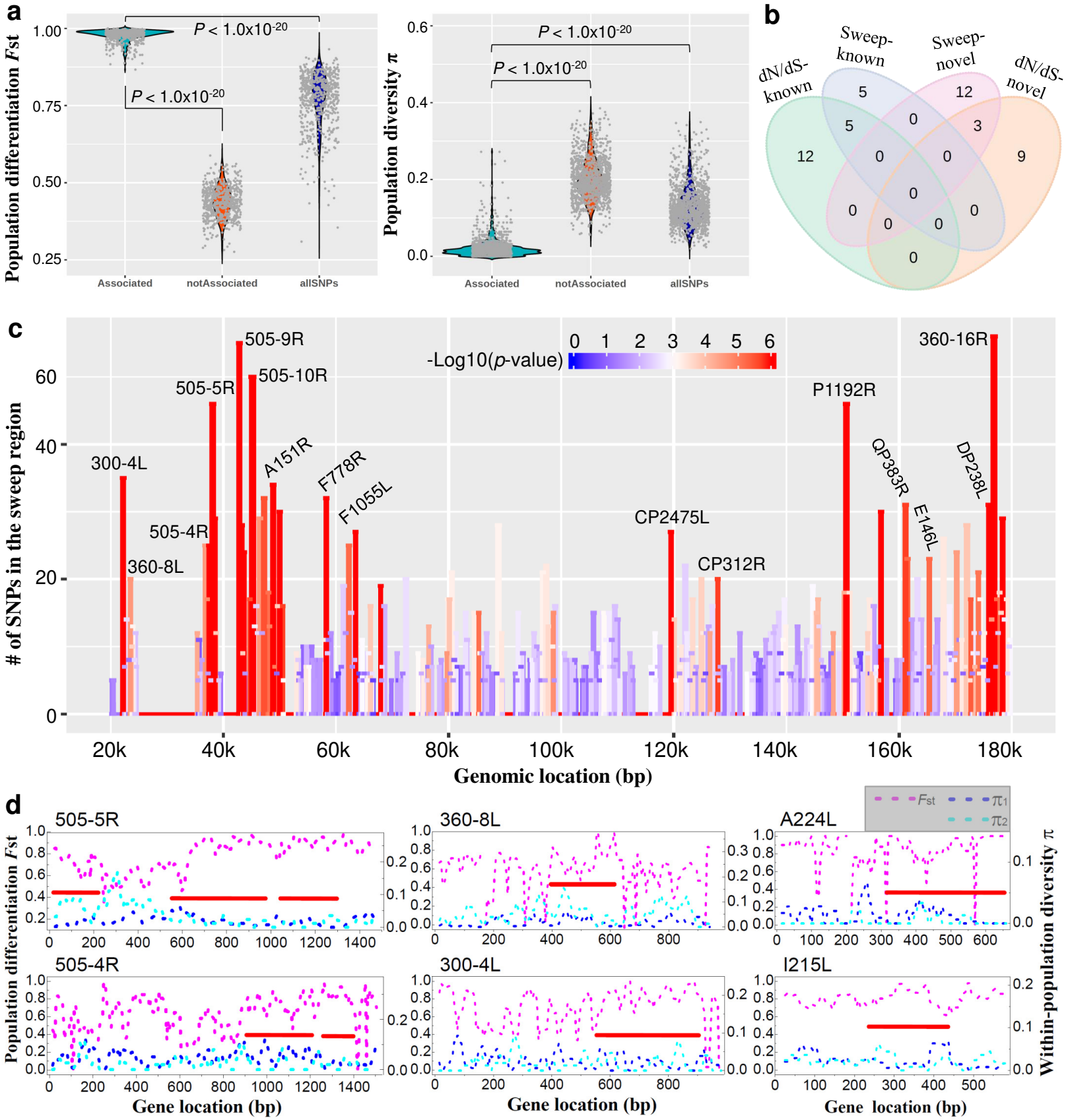
Supplementary file 2: The sites of genes under potential positive selection with ω (dN/dS) >1. The sites with the posterior probability $p \geq 0.9$ are shown in MGF505 genes and sites with the posterior probability $p \geq 0.8$ in MGF360 genes are shown (due to the overall low level of p-values for MGF360 genes). The sites with $p \geq 0.8$ in genes other than MGF360 are only shown if it is close or neighboring to the sites $p \geq 0.9$. The sites containing multiple gaps were not included in the list. The positions refer to those in the multiple alignment of each protein.

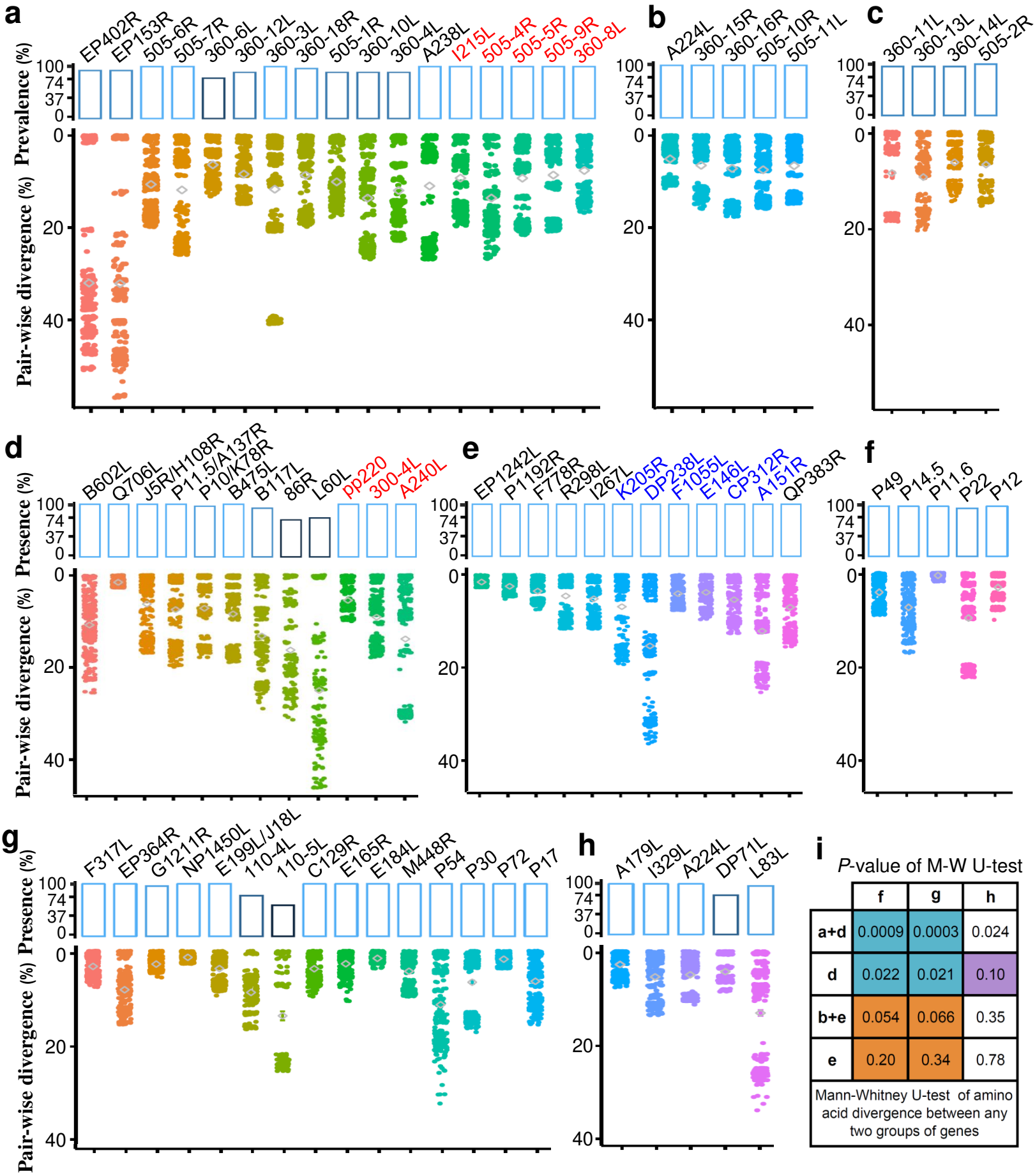
Supplementary file 3: The complete list of SNPs used for detection of selective sweep and the full list of regions of clusters identified by SweepCluster.

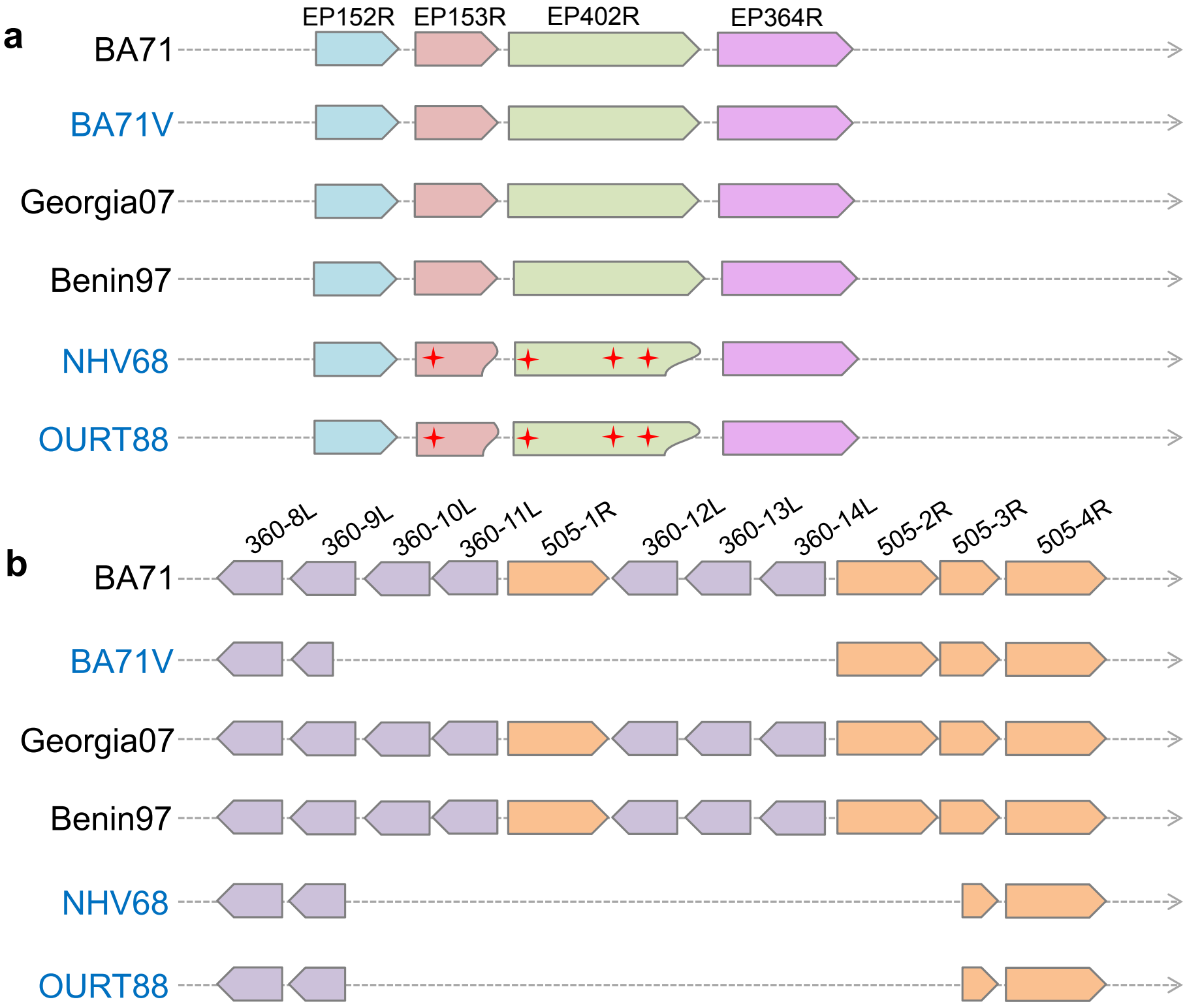
Supplementary file 4: The sites under positive selection using LRTs for pairs of MGF505 with posterior probability ≥ 0.9 and of MGF360 with posterior probability ≥ 0.8 in Bayes empirical tests.

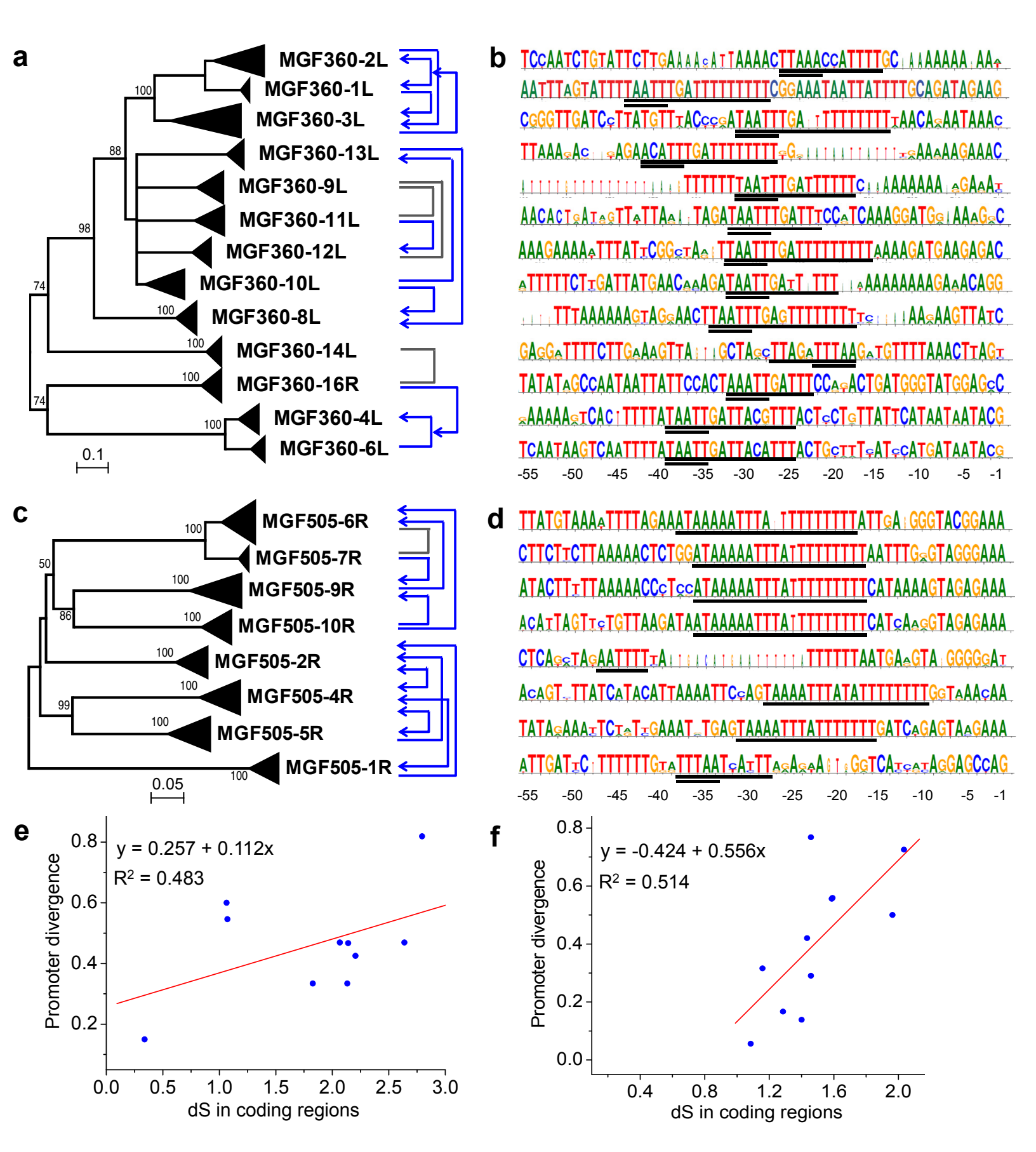

a
POL2015-Podlaskie
Pol17-C201
China-SY18
Estonia-2014
Russia-Odintsovo14
Russia-Kashino13
Georgia-2007
SouthAfrica-Mkuzi1979
Spain-BA71V
Spain-BA71
Portugal-L60
Portugal-OURT88
Portugal-NHV68
Spain-E75
WestAfrica-Benin97
Italy-26544OG10
Italy-47Ss2008
Malawi-Tengani62
SouthAfrica-Warmbaths04
Namibia-Warthog04
SouthAfrica-Pretori96
Malawi-Li183
Ken05-Tk1
Kenya-1950
Ken06-Bus
UgandaN10-2015
UgandaR7-2015
α
β
γ
b
Russia-Kashino13
Georgia-2007
China-SY18
Pol17-C201
Russia-Odintsovo14
POL2015-Podlaskie
Estonia-2014
SouthAfrica-Mkuzi1979
WestAfrica-Benin97
Spain-BA71V
Spain-BA71
Portugal-L60
Portugal-OURT88
Italy-26544OG10
Spain-E75
Portugal-NHV68
Italy-47Ss2008
Malawi-Tengani62
SouthAfrica-Warmbaths04
Namibia-Warthog04
SouthAfrica-Pretori96
Malawi-Lil83
Ken05-Tk1
Kenya-1950
Ken06-Bus
UgandaN10-2015
UgandaR7-2015
II
I
V
III
IV
XX
VIII
X
IX
c
I
II
III
IV
V
VI
VII
VIII
IX
X
XI
XII
XIII
XX
XXII
2005-2018
1990-2004
1980-1989
1950-1979

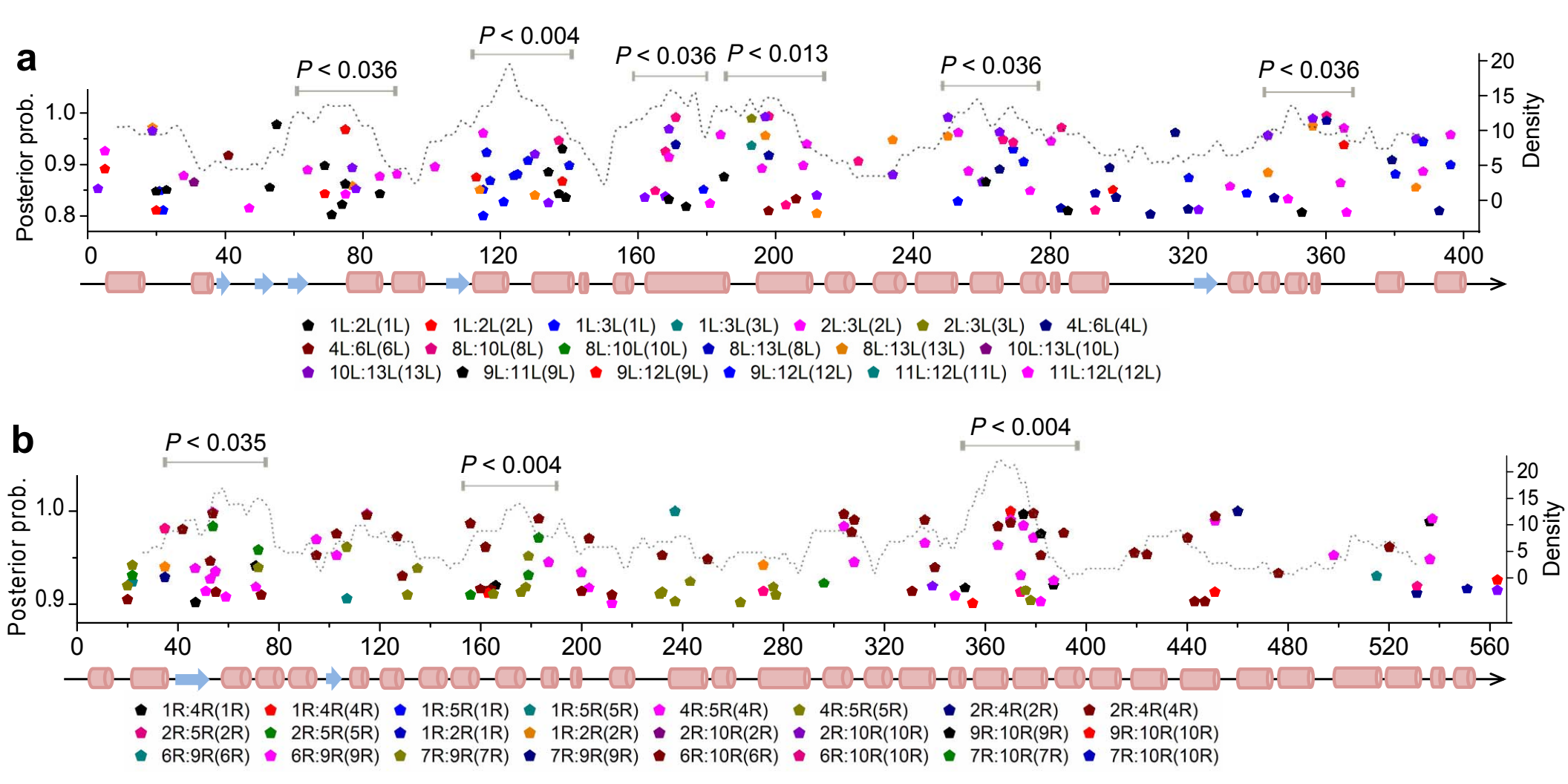

a
# of selected sites
-Log10(p-value)
10 7.5 5.0 2.5
CP2475L 505-4R EP402R I215L EP153R 505-6R B117L 505-9R B602L 86R 505-7R 360-6L A240L 300-4L 360-12L A137R H108R 360-3L A238L B475L 505-5R K78R L60L 360-8L 360-18R 505-1R 360-10L 360-4L Q706L
b
505-4R
MGF505 family domain
EP402R
Extracellular Ig-like domain
TM
P-rich repeats
TM
C-type lectin
EP153R
I215L
UQ_con
ANK ANK ANK ANK
A238L
c
H108R
TM
A137R
K78R
300-4L
Thymidylate_kin
A240L
B475L
SP-like
bZIP_1
86R
B602L
Viral polyN
C-rich repeats
L60L
DXQNT
H-rep
TM
B117L
Colicin-V
Q706L
SNF2 family helicase N-terminus
Helicase_C
CP2475L
d
CD2 Rat
CD2 Human
CD2 boar
EP402R
CD48 Rat
CD58 Human

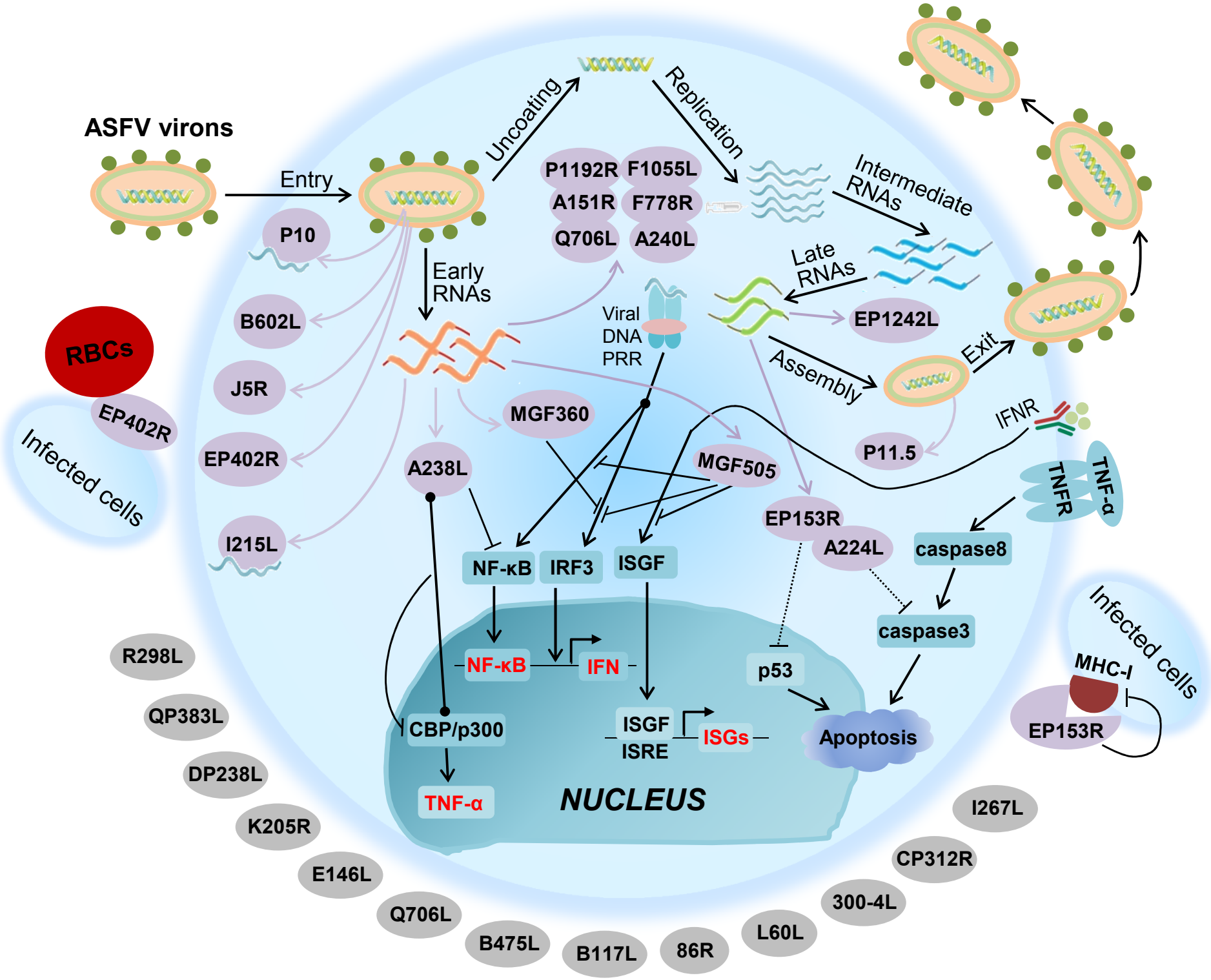

a
Population differentiation Fst
P < 1.0x10-20
P < 1.0x10-20
Associated
notAssociated
allSNPs
Population diversity π
P < 1.0x10-20
P < 1.0x10-20
b
dN/dS-known
Sweep-known
Sweep-novel
dN/dS-novel
c
-Log10(p-value)
# of SNPs in the sweep region
Genomic location (bp)
505-9R
505-10R
505-5R
A151R
F778R
F1055L
300-4L
505-4R
360-8L
CP2475L
CP312R
P1192R
QP383R
E146L
DP238L
360-16R
d
505-5R
360-8L
A224L
505-4R
300-4L
I215L
Fst
π1
π2
Population differentiation Fst
Within-population diversity π
Gene location (bp)


a
EP402R
EP153R
505-6R
505-7R
360-6L
360-12L
360-3L
360-18R
505-1R
360-10L
360-4L
A238L
I215L
505-4R
505-5R
505-9R
360-8L
b
A224L
360-15R
360-16R
505-10R
505-11L
c
360-11L
360-13L
360-14L
505-2R
Prevalence (%)
Pair-wise divergence (%)
d
B602L
Q706L
J5R/H108R
P11.5/A137R
P10/K78R
B475L
B117L
86R
L60L
pp220
300-4L
A240L
e
EP1242L
P1192R
F778R
R298L
I267L
K205R
DP238L
F1055L
E146L
CP312R
A151R
QP383R
f
P49
P14.5
P11.6
P22
P12
Presence (%)
g
F317L
EP364R
G1211R
NP1450L
E199L/J18L
110-4L
110-5L
C129R
E165R
E184L
M448R
P54
P30
P72
P17
h
A179L
I329L
A224L
DP71L
L83L
i
P-value of M-W U-test
f g h
a+d 0.0009 0.0003 0.024
d 0.022 0.021 0.10
b+e 0.054 0.066 0.35
e 0.20 0.34 0.78
Mann-Whitney U-test of amino acid divergence between any two groups of genes


a
EP152R
EP153R
EP402R
EP364R
BA71
BA71V
Georgia07
Benin97
NHV68
OURT88
b
360-8L
360-9L
360-10L
360-11L
505-1R
360-12L
360-13L
360-14L
505-2R
505-3R
505-4R
BA71
BA71V
Georgia07
Benin97
NHV68
OURT88


a
MGF360-2L
MGF360-1L
MGF360-3L
MGF360-13L
MGF360-9L
MGF360-11L
MGF360-12L
MGF360-10L
MGF360-8L
MGF360-14L
MGF360-16R
MGF360-4L
MGF360-6L
100
88
98
74
100
100
100
74
100
0.1
b
-55 -50 -45 -40 -35 -30 -25 -20 -15 -10 -5 -1
c
MGF505-6R
MGF505-7R
MGF505-9R
MGF505-10R
MGF505-2R
MGF505-4R
MGF505-5R
MGF505-1R
100
50
100
86
100
100
100
99
100
100
0.05
d
-55 -50 -45 -40 -35 -30 -25 -20 -15 -10 -5 -1
e
y = 0.257 + 0.112x
R² = 0.483
Promoter divergence
dS in coding regions
0.0 0.5 1.0 1.5 2.0 2.5 3.0
0.2 0.4 0.6 0.8
f
y = -0.424 + 0.556x
R² = 0.514
Promoter divergence
dS in coding regions
0.4 0.8 1.2 1.6 2.0
0.0 0.2 0.4 0.6 0.8


a
P < 0.036
P < 0.004
P < 0.036
P < 0.013
P < 0.036
P < 0.036
Posterior prob.
Density
1L:2L(1L)
1L:2L(2L)
1L:3L(1L)
1L:3L(3L)
2L:3L(2L)
2L:3L(3L)
4L:6L(4L)
4L:6L(6L)
8L:10L(8L)
8L:10L(10L)
8L:13L(8L)
8L:13L(13L)
10L:13L(10L)
10L:13L(13L)
9L:11L(9L)
9L:12L(9L)
9L:12L(12L)
11L:12L(11L)
11L:12L(12L)
b
P < 0.035
P < 0.004
P < 0.004
Posterior prob.
Density
1R:4R(1R)
1R:4R(4R)
1R:5R(1R)
1R:5R(5R)
4R:5R(4R)
4R:5R(5R)
2R:4R(2R)
2R:4R(4R)
2R:5R(2R)
2R:5R(5R)
1R:2R(1R)
1R:2R(2R)
2R:10R(2R)
2R:10R(10R)
9R:10R(9R)
9R:10R(10R)
6R:9R(6R)
6R:9R(9R)
7R:9R(7R)
7R:9R(9R)
6R:10R(6R)
6R:10R(10R)
7R:10R(7R)
7R:10R(10R)


ASFV virons
Entry
Uncoating
Replication
Intermediate RNAs
Late RNAs
Early RNAs
Assembly
Exit
P1192R
F1055L
A151R
F778R
Q706L
A240L
P10
B602L
J5R
EP402R
I215L
EP1242L
P11.5
Viral DNA PRR
MGF360
MGF505
A238L
EP153R
A224L
IFNR
TNFR
TNF-α
caspase8
caspase3
p53
Apoptosis
NF-κB
IRF3
ISGF
IFN
ISGF
ISRE
ISGs
CBP/p300
TNF-α
NUCLEUS
RBCs
EP402R
Infected cells
Infected cells
MHC-I
EP153R
R298L
QP383L
DP238L
K205R
E146L
Q706L
B475L
B117L
86R
L60L
300-4L
CP312R
I267L