The genetic variation landscape of African swine fever virus reveals frequent positive selection and adaptive flexibility

Yun-Juan Bao[1*], Junhui Qiu[1], Yuzi Luo[2], Fernando Rodr ǵuez[3], Hua-Ji Qiu[2,*]

[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, School of Life Sciences, Hubei University, Wuhan 430062, China

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

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

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

Running head: Genetic variation landscape of African swine fever virus

Keywords: African swine fever virus, Positive selection, Genetic variation, Selective sweep

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 gap in knowledge concerning genetic variation of ASFV and the genetic factors involved in host adaptation and virus-host interactions. In this study, we performed a meta-genetic study of ASFV aiming to profile the variation landscape and identify genetic factors with signatures of positive selection and relevance to host adaptation. 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, highlighting strong signals of adaptive evolution of ASFV. Further examination of the sequence properties reveals the link of the selection signatures with virus-host interactions and adaptive flexibility. Specifically, we discovered a site at 157th of the key antigen protein EP402R under diversifying selection, which is located in the cytotoxic T-cell epitope related with the low level of cross-reaction in T-cell response. Importantly, two multigene families MGF360 and MGF505, the host range factors of ASFV, exhibit divergent selection among the paralogous members, conferring sequence pools for genetic diversification and adaptive capability. By integrating the genes with selection signatures into a unified framework of interactions between ASFV and hosts, we showed that the genes are involved in multiple processes of host immune interaction and virus life cycles, and may play crucial roles in circumventing host defense systems and enhancing adaptive fitness. Our findings will allow enhanced understanding of genetic basis of rapid spreading and adaptation of ASFV among the hosts.

Introduction

African swine fever virus (ASFV) is the causative agent of haemorrhagic fever in swine. ASFV mainly replicates in swine macrophages, causing up to 100% mortality rates in domestic pigs. However, the transmission pathway of ASFV is highly intertwined through the sylvatic cycle and domestic cycle involving multiple intermediate points, such as warthogs, soft ticks, wild boars, domestic pigs and human activities (Sánchez-Cordón et al., 2018). ASFV is thought to originate from and circulate in wild swine and soft ticks in Eastern Africa, and the first infection in domestic pigs was reported in Kenya in 1921 (Montgomery, 1921). From Africa, ASFV has spread to Europe in 1957 and 1960 (Portugal) and to Georgia in 2007 (Rebecca et al., 2008). This last introduction led to the expansion of the disease through the Caucasus and Russia (Oganesyan et al., 2013) to European Union countries, such as Estonia, Latvia, and Poland (Gallardo et al., 2014; Stokstad, 2017). More recently, ASFV was detected in wild boars in Belgium and an outbreak in a pig farm was reported in China in mid-2018 (Garigliany et al., 2019; Ge et al., 2018). Since then, the virus has continuously spread to neighboring countries in Asia, affecting hundreds of millions of swine population (FAO, 2019).

Since there is no commercially available vaccine against ASFV infection, current disease control is based on physical quarantine and animal slaughtering. A large amount of them have been killed since the spread of infection globally, causing substantial damages on the swine population. Development of efficacious therapeutic and prophylactic tools has been largely hindered by the limited knowledge of genetics properties and the evolutionary adaptation of this highly pathogenic virus.

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. Previous studies mainly focused on a limited number of virulence determinants or host-range factors, such as EP402R (Borca et al., 1998), EP153R (Hurtado et al., 2011), A238L (Powell et al., 1996), and the highly variable multigene families (MGFs) at both ends of the genome (Chapman et al., 2008; de Villiers et al., 2010; Dixon et al., 2013). Specifically, studies using engineered deletion mutants investigated the variation patterns of MGF genes (Chapman et al., 2008; Rodr ǵuez et al., 2015), showing that MGF genes are relevant to host interactions and might be responsible for host range functions (Dixon et al., 2013; Zsak et al., 2001).

However, there are a limited number of studies on systematic characterization of genetic properties for the whole gene set in the genome-wide scale (Chapman et al., 2008; de Villiers et al., 2010). As a dsDNA virus, ASFV has an estimated substitution rate $\mu \sim 6.7x10^{-4}$ (substitutions per site per year) (Michaud et al., 2013), roughly between that of RNA viruses such as the influenza virus with $\mu \sim 10^{-3}$ (Hanada et al., 2004) and that of other large dsDNA viruses such as herpes simplex type I virus with $\mu \sim 10^{-5}$ (Duffy et al., 2008). This substitution level is much higher than that of many bacterial species such as *Streptococcus pneumoniae* with $\mu \sim 10^{-6}$ (Croucher et al., 2013). The high substitution rate indicates a high level of variability in the seemingly conserved central regions previously thought. This high level of genetic variability may have important implication for understanding the puzzling adaptive capability and host range of ASFV. For instance, investigation of the breath of the variability on the genome-wide scale will help to understand the molecular basis underlying the evolutionary adaptation; identifying the selection pressures acting on the genes will reveal the genetic factors exposed to host-virus interfaces; dissecting the genetic diversity of MGF genes will help to clarify the driving force of sequence divergence and functional diversification of MGF. During the decades of study on ASFV, those critical issues are still largely unknown.

In this study, we performed a meta-genetic study by using sophisticated computational methods to profile the variation landscape of ASFV and identify the genetic factors under positive selection aiming to characterize the genetic factors relevant to the versatile adaptation and host interactions for ASFV.

Materials and Methods

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). The non-redundant genomes were identified and used for downstream analysis by excluding those with close evolutionary distance (< 0.001 substitutions per site), the same isolation countries and isolation time with other strains (see Table S1). The core genome represents the genes or genomic locations present in all studied strains of a species and we created the core genome of ASFV by aligning the shredded genomes against the reference strain Georgia-2007 and extracting the genomic regions mapped by all other genomes. Finally, the core genome contains 139,677 base pairs and was used for single nucleotide polymorphism (SNP) detection. The bases at the variant loci for each ASFV genome were concatenated for distance estimation and phylogeny construction using MEGA6 (Tamura et al., 2013) and SplitsTree (Huson and Bryant, 2006). The pair-wise distance was measured by substitutions per site with the model of maximum composite likelihood and the tree topology was inferred 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* is a statistic for testing the neutrality of the mutations on the overall scale by computing the difference between the average number of pairwise nucleotide differences and the number of segregating sites and was calculated as defined by Tajima (Tajima, 1989).

Detection of functional domains. The functional domains of the genes were detected by comparison with the PFAM database (Punta et al., 2012). The hits with score $\geq$ 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 a species is the whole set of genes encoded by all studied strains. The genes present in different strains facilitating similar functions form orthologous groups and are key components of the pan-genome. On the other hand, paralogous genes are those duplicated in the same strain from a common ancestor with related but divergent functions (Jensen, 2001). Derivation of a new paralogous member in a gene family will lead to emergence of new functions and expansion of the pan-genome size. The pan-genome of the 27 non-redundant ASFV genomes was generated using Roary yielding 192 pan-genes encoded by at least one strain of ASFV (Page et al., 2015). 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 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 was performed at first in amino acids for each orthologous gene group and then were back converted to 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 (Nei and Gojobori, 1986). Likelihood ratio tests (LRT) of selection pressures acting on individual sites of ASFV genes were carried out using PAML with the site-specific model (Yang, 2007). 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 (Yang et al., 2005). Likelihood ratio tests of divergent selection of MGF genes were performed using the branch-site Model A in PAML (Zhang et al., 2005). 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 for MGF360 and ≥ 0.9 for MGF505 using Bayes empirical tests were tabulated and mapped to the respective secondary structures.

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.

Secondary structure prediction. The secondary structures of B475L and MGF300-4L were predicted using PSIpred (Buchan and Jones, 2019), and those of MGF360 and MGF505 using PROMALS3D (Pei et al., 2008).

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 (Kelley et al., 2015). Multiple sequence alignment of EP402R and its homologs in animals, including human CD2 (Bodian et al., 1994) (PDB ID: 1hnf), human CD58 (Ikemizu et al., 1999) (PDB ID: 1ccz), rat CD2 (Jones et al., 1992) (PDB ID: 1hng), rat CD48 (Evans et al., 2006) (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 (Robert and Gouet, 2014). The structures of the proteins were presented and analyzed using PyMOL (Benoit et al., 2008).

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

Identification of regions with selective sweep. The population size is highly unbalanced between the two subpopulations α (21 strains) and β (5 strains), therefore we at first 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 clustering algorithm described in (Bao et al., 2016). Briefly, a non-synonymous SNP is randomly chosen in a specific gene as the initial cluster and each initial cluster is then iteratively extended until its spanning range approaches the specified sweep length or the boundary of the gene or gene operon. The cluster is further extended to merge the neighboring SNPs or clusters by minimizing the root-mean-square of inter-SNP distances. 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:

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

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 (Fig. 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 (Tajima, 1989). The calculation of Watterson's estimator θ (Watterson, 1975) 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 (Fig. 1a and Supplementary file 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 (Fig. 1b,c, Fig. S1, and Supplementary file 1b). 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 (Atuhaire et al., 2013). 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 (de Villiers et al., 2010; Rebecca et al., 2008).

We observed two prominent features of the phylogenetic structure and geographical distribution depicted in Fig. 1. First, the tree has a total branch length of 2.1 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 re-elaborates the rapid adaptation of ASFV to host environments and the complexity of the transmission pathways of ASFV. Third, no temporal-spatial dynamics pattern can be inferred from the phylogenetic structure except the recent spreading of genotype II strains. Next, we will investigate in details the genetic variation profile of the whole population of ASFV, but without focusing on specific genotypes.

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 (Donnell et al., 2015; Krug et al., 2015; Rodr ǵuez et al., 2015), largely due to the postulated roles of MGF360 and MGF505 in host immune evasion and infection tropism (Dixon et al., 2013; Donnell et al., 2015). 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 (Fig. S2a). The analysis using Hypergeometric test ranked 23 genes to be significantly enriched with non-synonymous mutations (multiple testing corrected *p*-value ≤ 0.001) but not with synonymous mutations (multiple testing corrected *p*-value ≥ 0.05) (Table S3 and Fig. 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 (Punta et al., 2012) (Table S4). 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 Fig. 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 diversifying selection acting on the ASFV-encoded genes. Positive diversifying selection is represented as elevated amino acid diversity within or across the populations resulting in selection of multiple phenotypes. It can be detected by measuring the rates of non-synonymous substitution (dN) and synonymous substitutions (dS) and calculating their ratio dN/dS. We at first calculated the dN/dS for each gene based on the Nei & Gojobori model (Nei and Gojobori, 1986). 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 S5). 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 (Galindo et al., 2000; Ruiz-Gonzalvo et al., 1996).

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) in PAML (Yang, 2007). 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) (Fig. 2a and Table S6). 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 (de Villiers et al., 2010).

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 1 and Table S6).

Functional implication of the positively selected sites. 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 (Fig. 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, eighteen positively selected sites are identified in EP402R and significantly 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 (Alejo et al., 2018; Borca et al., 1998; Rodr ǵuez et al., 1993). 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 (Fig. 2d and Fig. 3). As a CD2 homolog, the extracellular domain of EP402R consists of a constant C-set and a variable V-set Ig-superfamily domain (Fig. 3a-d). We then mapped the positively selected sites to the aligned sequences and the tertiary structures. It is remarkable that the sites under positive selection 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 (Fig. 3a-c) (Davis et al., 1998). The orthologous loop regions in Ig antibodies are responsible for facilitating specificity of antibodies to recognize antigens (Morea et al., 2000). It indicates the potential roles of the positively selected sites in the loop regions of EP402R in determining specificity of ASFV for host cell recognition and enhancing adaptability.

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 (Argilaguet et al., 2012). 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 (Fig. 3e). 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.

A recent study of EP402R and EP153R also investigated the high sequence variability and positively selected sites in the two proteins in detail, but detected less sites under positive selection (Nefedeva et al., 2020). However, the recombination, as demonstrated in that study, might provide an alternative manifestation of positive selection exerted on ASFV, although the explicit quantification of their link might pose a new challenge.

In addition to the divergent proteins, four highly conserved structural proteins (J5R/H108R, P11.5/A137R, P10/K78R, and pp220/CP2475L, in Fig. 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 Fig. 2c), have the positively selected sites distributed across a large proportion of the gene regions. The two proteins are unique in 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 (Buchan and Jones, 2019). It shows that the two proteins predominantly comprise tandem α-helices. The tandem α-helix structural units have been shown to be able to stack side-by-side arranged in a specific three-dimensional conformation to create protein-binding interfaces and are commonly found in binding proteins (Groves and Barford, 1999). The presence of the tandem α-helices in the two proteins MGF300-4L and B475L indicates their possible roles in protein-protein interactions (Fig. S3).

Identification of selective sweeps in the ASFV genomes. A selective sweep is a process where a beneficial allelic change sweeps through the population and becomes fixed in a specific population, and the nearby linked sites will hitchhike together and also become fixed. The process leads to reduced within-population genetic diversity and increased between-population differentiation in the sweeping region. Such selective sweeps allow for rapid adaptation and accelerated evolution, and are good indicators for host-pathogen interaction and adaptive evolution (Stephan, 2019). The unique mechanism of selective sweeps in causing genetic changes makes it inappropriate to detect them using the dN/dS-based method. Therefore, we detect the regions of clustered SNPs with gamma distribution, which is characteristic of SNPs under selective sweep (See Materials and Methods). We at first identified 6,054 SNPs associated with between-population subdivision and within-population homogeneity for the clade α and β (Fig. 4a). Those SNPs were subsequently subject to detection of selective sweep. 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 high significance in the signatures of selective sweep (Fig. 4b,c and Table 2).

The gene regions with significant selective sweep exhibit higher population differentiation and reduced sequence diversity as shown in the key signature genes (Fig. 4d). 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 (Fig. 4d and Fig. 2b). Noteworthy are the 15 novel candidate genes showing strong signatures of selective sweep (Table 2). A large proportion of them (60%) are involved in key cellular functions, such as replication, repair, transcription, and metabolism (Table 2).

We notice that four of the novel candidates (A151R, F1055L, CP312R, and E146L) have been previously demonstrated to induce immune responses in swine following ASFV challenge (Jancovich et al., 2018; Netherton et al., 2019). 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 (Alejo et al., 2018), (ii) the antigen proteins eliciting immunological responses in immunoassay experiments (Jancovich et al., 2018; Lopera-Madrid et al., 2017; Netherton et al., 2019), (iii) the proteins previously shown to be involved in host cell interactions (Dixon et al., 2013; Dixon et al., 2019) (Fig. 5 and Table S7). 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, Fig. 5a,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 (Fig. 5b,e,i).

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 significant signatures of positive selection, a 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 the two families MGF360 or MGF505. We first constructed the phylogenetic structures of all orthologous and paralogous members of MGF360 and MGF505, respectively (Fig. 6a,c and Fig. S4), and then chose the phylogenetically close pairs of genes/branches to perform the likelihood ratio test of divergent selection. 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) (Fig. 6a,c and

Table S8). 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 (Fig. 6b,d). 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 (Fig. S5). 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. Interestingly, we found that the consensus motif patterns we obtained specifically for MGF360/MGF505 are very similar to that profiled experimentally in a recent study for a set of early transcribed genes including MGF genes in the ASFV strain Spain-BA71V (Cackett et al., 2020). The consistency between the two studies provides further support for our results.

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 (Fig. 6e,f 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 (Mosavi et al., 2004), 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 (Mosavi et al., 2004). Ankyrin repeats have been described to be the major functional units in host range factors in several poxvirus species (Bradley and Terajima, 2005; Herbert et al., 2015; Li et al., 2010). 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 genetic and 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 positive selection signatures and relevance to host adaptation and interaction, we identified 29 candidate genes with positive diversifying selection and 25 with selective sweep. 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 adaptive evolution at multiple genomic regions of ASFV during the interactions with hosts. 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 (Fig. 7) (Rodriguez and Salas, 2013).

The proteins in the scheme include those known to be relevant to host immune evasion, such as EP402R for surface adherence of infected cell (Borca et al., 1998), EP153R for inhibition of MHC expression and host cell apoptosis (Alejo et al., 2018; Hurtado et al., 2011), A238L for production impairment of immune regulator NF-κB and cytokines TNF-α (Powell et al., 1996), and multiple MGF genes for modulation of interferon (IFN) response (Afonso et al., 2004; Correia et al., 2013).

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 (Alejo et al., 2018), the basic enzymes P1192R, F1055L, F778R, A240L and EP1242L involved in replication, repair and transcription in host cytoplasm (Dixon et al., 2013). The key roles played by the proteins and the relatively high conservation make them promising candidates for vaccines with cross-activity.

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 reveals that the adaptive evolution of ASFV has been shaped by both positive diversifying selection and selective sweep. The results show that the genes with diversifying selection exhibit a higher level of sequence variability than those with selective sweep and provide important implications for vaccine design. The most prominent are EP402R, EP153R and MGF360/MGF505 with the highest genetic variability, the only known proteins so far shown to be both virulence determinants and immunogenic targets (Boinas et al., 2004; Burmakina et al., 2016). However, the high sequence diversity of EP402R/EP153R and mosaic presence pattern of MGF360/MGF505 among the ASFV population make it difficult for them to achieve desirable cross-protection (Malogolovkin et al., 2015). The dual role of EP402R, EP153R and MGF360/MGF505, as both potential virulence determinants and immunogenic proteins, may also introduce confounding factors in designing live-attenuated virus vaccines (LAVs). Recently, 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 (Monteagudo et al., 2017). Similarly, ASFV-Georgia-ΔMGF, a LAV strain lacking a series of MGF genes, protected animals against homologous challenges (Donnell et al., 2015). Unfortunately, sequential deletion of multiple genes provoked in occasion the loss of protection due to excessive attenuation (O'Donnell et al., 2016). It is worth mentioning that the role of EP402R as a virulence factor of ASFV has not yet been explicitly determined due to differential virulence outcomes from disruption of EP402R in distinct isolates. A few studies have shown that abrogation of EP402R function does not significantly alter the virulence of the mutants (Borca et al., 1998; Borca et al., 2020). The isolate-dependent functional effect of EP402R will pose additional challenges for designing LAVs.

The divergent selection between paralogous genes of MGF360/MGF505 further complicates the vaccine design. We identified differentiated selection pressures and regulation patterns between paralogs of MGF360/MGF505 conferring genetic diversity and functional diversification. The possible scenario is that the antigenic activities and expression levels of paralogs of MGF360/MGF505 are strain-specific and/or host-dependent. This scenario provides a rationale for the observations that variable deletion patterns and expression profiles of paralogous members of MGF have been resulted from different adaptation processes or have induced distinct viral growth outcomes in host niches (Krug et al., 2015; Rodr ǵuez et al., 2015). Up to now, the precise connections between the MGF genes and physiological conditions are still largely unknown. Optimal choices of paralogous 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 relatively 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 (Memoli et al., 2009). 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 relatively high conservation and critical roles of the genes make them promising candidates for vaccine molecules or drug targets.

We highlight the importance of our findings in the two following two aspects. (1) Our data provides novel insights into the adaptation and fitness of ASFV. The multifaceted genetic characteristics of ASFV genes imply that the virus may have utilizing multiple mechanisms (such as genetic diversification, selective sweep and divergent selection) and pertinent genetic factors for successful replication, adaption, and persistence during interaction with continuously changing host environments, including warthogs, ticks, and domestic pigs. The plethora of variable genetic factors may act as a genetic pool for adaptation to new hosts by functional diversification or immune escape, though the alternative new hosts beyond the currently known have not yet been reported. (2) The candidate genes we identified in the study could serve as valuable targets for vaccine molecules or therapeutic agents, and the sites with signatures of positive selection will be valuable for precise design and engineering. We also understand that the methods we used for identifying selection are not perfect and the genetic variability might have been underestimated due to the limited size of the ASFV population or the conservation of the selection analysis methods. We believe that the availability of more genomic information in the future will be of great help for overcoming the limitation.

Data availability

The multiple sequence alignments used for selection analysis and supplementary files are available through the links: https://figshare.com/projects/ASFV_alignment/82718 and https://figshare.com/projects/ASFV_supplementary_files/90335, respectively under the MIT license.

Acknowledgements

The work was supported by the National Key Research and Development Program of China (2018YFC0840401).

Conflict of interest

The authors declare no competing interests.

Ethics statement

The authors confirm that the ethical policies of the journal, as noted on the journal's author guidelines page, have been adhered to. No ethical approval was required as this is a meta-analysis article.

References

Afonso C.L., Piccone M.E., Zaffuto K.M., Neilan J., Kutish G.F., Lu Z., *et al.* (2004) African swine fever virus multigene family 360 and 530 genes affect host interferon response. J. Virol. 78:1858-1864.

Alejo A., Matamoros T., Guerra M., Andrés G. (2018) A proteomic atlas of the African swine fever virus particle. J. Virol. 92:e01293-18.

Argilaguet J.M., Pérez-Martín E., Nofrarías M., Gallardo C., Accensi F., Lacasta A., *et al.* (2012) DNA vaccination partially protects against African swine fever virus lethal challenge in the absence of antibodies. PloS ONE 7:e40942.

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

Bao Y.-J., Shapiro B.J., Lee S.W., Ploplis V.A., Castellino F.J. (2016) Phenotypic differentiation of Streptococcus pyogenes populations is induced by recombination-driven gene-specific sweeps. Sci Rep. 6:36644.

Benoit M., Desnues B., Mege J.L. (2008) Macrophage Polarization in Bacterial Infections. J. Immunol. 181:3733-3739.

Bodian D.L., Jones E.Y., Harlos K., Stuart D.I., Davis S.J. (1994) Crystal structure of the extracellular region of the human cell adhesion molecule CD2 at 2.5 A resolution. Structure 2:755-766.

Boinas F.S., Hutchings G.H., Dixon L.K., Wilkinson P.J. (2004) Characterization of pathogenic and non-pathogenic African swine fever virus isolates from Ornithodoros erraticus inhabiting pig premises in Portugal, J. Gen. Virol. pp. 2177-2187.

Borca M.V., Carrillo C., Zsak L., Laegreid W.W., Kutish G.F., Neilan J.G., Burrage T.G., Rock D.L. (1998) Deletion of a CD2-like gene, 8-DR, from African swine fever virus affects viral infection in domestic swine. J. Virol. 72:2881-2889.

Borca M.V., O'Donnell V., Holinka L.G., Risatti G.R., Ramirez-Medina E., Vuono E.A., *et al.* (2020) Deletion of CD2-like gene from the genome of African swine fever virus strain Georgia does not attenuate virulence in swine. Sci. Rep. 10:494.

Bradley R.R., Terajima M. (2005) 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. 114:104-12.

Buchan D.W.A., Jones D.T. (2019) The PSIPRED Protein Analysis Workbench: 20 years on. Nucleic Acids Res 47:W402-w407.

Burmakina G., Malogolovkin A., Tulman E.R., Zsak L., Delhon G., Diel D.G., *et al.* (2016) African swine fever virus serotype-specific proteins are significant protective antigens for African swine fever. J. Gen. Virol. 97:1670-1675.

Cackett G., Matelska D., Sýkora M., Portugal R., Malecki M., Bähler J., Dixon L., Werner F. (2020) The African Swine Fever Virus Transcriptome. J. Virol. 94.

Chapman D.A.G., Tcherepanov V., Upton C., Dixon L.K. (2008) Comparison of the genome sequences of non-pathogenic and pathogenic African swine fever virus isolates. J. Gen. Virol. 89:397-408.

Correia S., Ventura S., Parkhouse R.M. (2013) Identification and utility of innate immune system evasion mechanisms of ASFV. Virus Res. 173:87-100.

Croucher N.J., Finkelstein J.A., Pelton S.I., Mitchell P.K., Lee G.M., Parkhill J., Bentley S.D., Hanage W.P., Lipsitch M. (2013) Population genomics of post-vaccine changes in pneumococcal epidemiology. Nat. Genet. 45:656-663.

Davis S.J., Ikemizu S., Wild M.K., van der Merwe P.A. (1998) CD2 and the nature of protein interactions mediating cell-cell recognition. Immunol Rev。 163:217-236.

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

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

Dixon L.K., Islam M., Nash R., Reis A.L. (2019) African swine fever virus evasion of host defences. Virus Res. 266:25-33.

Donnell V., Holinka L.G., Gladue D.P., Sanford B., Krug P.W., Lu X., *et al.* (2015) African swine fever virus Georgia isolate harboring deletions of MGF360 and MGF505 genes is attenuated in swine and confers protection against challenge with virulent parental virus. J. Virol. 89:6048-6056.

Duffy S., Shackelton L.A., Holmes E.C. (2008) Rates of evolutionary change in viruses: patterns and determinants. Nat. Rev. Genet. 9:267-276.

Evans E.J., Castro M.A., O'Brien R., Kearney A., Walsh H., Sparks L.M., *et al.* (2006) Crystal structure and binding properties of the CD2 and CD244 (2B4)-binding protein, CD48. J. Biol. Chem. 281:29309-29320.

FAO. (2019) ASF situation update - African Swine Fever (ASF) - FAO Emergency Prevention System for Animal Health (EMPRES-AH). www.fao.org 2019-11-03.

Galindo I., Almazán F., Bustos M.J., Viñuela E., Carrascosa A.L. (2000) African swine fever virus EP153R open reading frame encodes a glycoprotein involved in the hemadsorption of infected cells. Virology 266:340-351.

Gallardo C., Fernández-Pinero J., Pelayo V., Gazaev I., Markowska-Daniel I., Pridotkas G., *et al.* (2014) Genetic variation among African swine fever genotype II viruses, eastern and central Europe. Emerg. Infect. Dis. 20:1544-1547.

Garigliany M., Desmecht D., Tignon M., Cassart D., Lesenfant C., Paternostre J., *et al.* (2019) Phylogeographic Analysis of African Swine Fever Virus, Western Europe, 2018. Emerg. Infect. Dis. 25:184-186.

Ge S., Li J., Fan X., Liu F., Li L., Wang Q., *et al.* (2018) Molecular Characterization of African Swine Fever Virus, China, 2018. Emerg. Infect. Dis. 24:2131-2133.

Groves M.R., Barford D. (1999) Topological characteristics of helical repeat protein. Curr. Opin. Struct. Biol. 9:383-389.

Hanada K., Gojobori T., Suzuki Y. (2004) 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. 21:1074-1080.

Herbert M.H., Squire C.J., Mercer A.A. (2015) Poxviral ankyrin proteins. Viruses 7:709-738.

Hurtado C., Bustos M.J., Granja A.G., de Leon P., Sabina P., Lopez-Vinas E., Gomez-Puertas P., Revilla Y., Carrascosa A.L. (2011) The African swine fever virus lectin EP153R modulates the surface membrane expression of MHC class I antigens. Arch. Virol. 156:219-234.

Huson D.H., Bryant D. (2006) Application of phylogenetic networks in evolutionary studies. Mol. Biol. Evol. 23:254-267.

Ikemizu S., Sparks L.M., van der Merwe P.A., Harlos K., Stuart D.I., Jones E.Y., Davis S.J. (1999) 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 96:4289-4294.

Jancovich J.K., Chapman D., Hansen D.T., Robida M.D., Loskutov A., Craciunescu F., *et al.* (2018) 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. 92:e02219-17.

Jensen R.A. (2001) Orthologs and paralogs - we need to get it right. Genome Biol. 2:INTERACTIONS1002.

Jones E.Y., Davis S.J., Williams A.F., Harlos K., Stuart D.I. (1992) Crystal structure at 2.8 A resolution of a soluble form of the cell adhesion molecule CD2. Nature 360:232-239.

Kelley L.A., Mezulis S., Yates C.M., Wass M.N., Sternberg M.J.E. (2015) The Phyre2 web portal for protein modeling, prediction and analysis. Nat. Protoc. 10:845-858.

Krug P.W., Holinka L.G., Donnell V., Reese B., Sanford B., Fernandez-Sainz I., *et al.* (2015) 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. 89:2324-2332.

Li Y., Meng X., Xiang Y., Deng J. (2010) Structure function studies of vaccinia virus host range protein k1 reveal a novel functional surface for ankyrin repeat proteins. J. Virol. 84:3331-3338.

Lopera-Madrid J., Osorio J.E., He Y., Xiang Z., Adams L.G., Laughlin R.C., *et al.* (2017) Safety and immunogenicity of mammalian cell derived and Modified Vaccinia Ankara vectored African swine fever subunit antigens in swine. Vet. Immunol. Immunopathol. 185:20-33.

Malogolovkin A., Burmakina G., Tulman E.R., Delhon G., Diel D.G., Salnikov N., Kutish G.F., Kolbasov D., Rock D.L. (2015) African swine fever virus CD2v and C-type lectin gene loci mediate serological specificity. J Gen Virol 96:866-873.

Memoli M.J., Jagger B.W., Dugan V.G., Qi L., Jackson J.P., Taubenberger J.K. (2009) Recent human influenza A/H3N2 virus evolution driven by novel selection factors in addition to antigenic drift. J. Infect. Dis. 200:1232-1241.

Michaud V., Randriamparany T., Albina E. (2013) Comprehensive phylogenetic reconstructions of African swine fever virus: proposal for a new classification and molecular dating of the virus. PLoS One 8:e69662.

Monteagudo P.L., Lacasta A., López E., Bosch L., Collado J., Pina-Pedrero S., *et al.* (2017) BA71ΔCD2: a new recombinant live attenuated African swine fever virus with cross-protective capabilities. J. Virol. 91:e01058-17.

Montgomery R. (1921) On a form of swine fever occurring in British East Africa. J. Comp. Pathol. 34:159-191.

Morea V., Lesk A.M., Tramontano A. (2000) Antibody modeling: implications for engineering and design. Methods 20:267-279.

Mosavi L.K., Cammett T.J., Desrosiers D.C., Peng Z.-Y. (2004) The ankyrin repeat as molecular architecture for protein recognition. Prot. Sci. 13:1435-1448.

Nefedeva M., Titov I., Tsybanov S., Malogolovkin A. (2020) Recombination shapes African swine fever virus serotype-specific locus evolution. Sci. Rep. 10:18474.

Nei M., Gojobori T. (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. 3:418-426.

Netherton C.L., Goatley L.C., Reis A.L., Portugal R., Nash R.H., Morgan S.B., *et al.* (2019) Identification and immunogenicity of African swine fever virus antigens. Front. Immunol. 10:1318.

O'Donnell V., Holinka L.G., Sanford B., Krug P.W., Carlson J., Pacheco J.M., *et al.* (2016) 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. 221:8-14.

Oganesyan A.S., Petrova O.N., Korennoy F.I., Bardina N.S., Gogin A.E., Dudnikov S.A. (2013) African swine fever in the Russian Federation: Spatio-temporal analysis and epidemiological overview. Virus Res. 173:204-211.

Page A.J., Cummins C.A., Hunt M., Wong V.K., Reuter S., Holden M.T.G., *et al.* (2015) Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics (Oxford, England) 31:3691-3693.

Pei J., Kim B.H., Grishin N.V. (2008) PROMALS3D: a tool for multiple protein sequence and structure alignments. Nucleic Acids Res. 36:2295-300.

Powell P.P., Dixon L.K., Parkhouse R.M. (1996) An IkappaB homolog encoded by African swine fever virus provides a novel mechanism for downregulation of proinflammatory cytokine responses in host macrophages. J. Virol. 70:8527-8533.

Punta M., Coggill P.C., Eberhardt R.Y., Mistry J., Tate J., Boursnell C., *et al.* (2012) The Pfam protein families database. Nucleic Acids Res. 40:D290-D301.

Rebecca J.R., Vincent M., Livio H., Geoff H., Chris O., Wilna V., *et al.* (2008) African swine fever virus isolate, Georgia, 2007. Emerg. Infect. Dis. 14:1870-1874.

Robert X., Gouet P. (2014) Deciphering key features in protein structures with the new ENDscript server. Nucleic Acids Res. 42:W320-4.

Rodríguez J.M., Moreno L.T., Alejo A., Lacasta A., Rodríguez F., Salas M.L. (2015) Genome sequence of African swine fever virus BA71, the virulent parental strain of the nonpathogenic and tissue-culture adapted BA71V. PLoS One 10:e0142889.

Rodriguez J.M., Salas M.L. (2013) African swine fever virus transcription. Virus Res. 173:15-28.

Rodríguez J.M., Yáñez R.J., Almazán F., Viñuela E., Rodriguez J.F. (1993) African swine fever virus encodes a CD2 homolog responsible for the adhesion of erythrocytes to infected cells. J. Virol. 67:5312-5320.

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

Sánchez-Cordón P.J., Montoya M., Reis A.L., Dixon L.K. (2018) African swine fever: A re-emerging viral disease threatening the global pig industry. Vet. J. 233:41-48.

Stephan W. (2019) Selective Sweeps. Genetics 211:5.

Stokstad E. (2017) Deadly virus threatens European pigs and boar. Science 358:1516-1517.

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

Tamura K., Stecher G., Peterson D., Filipski A., Kumar S. (2013) MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. Mol. Biol. Evol. 30:2725-2729.

Watterson G.A. (1975) On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. 7:256-276.

Yang Z. (2007) PAML 4: Phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24:1586-1591.

Yang Z., Wong W.S.W., Nielsen R. (2005) Bayes empirical bayes inference of amino acid sites under positive selection. Mol. Biol. Evol. 22:1107-1118.

Zhang J., Nielsen R., Yang Z. (2005) Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol. Biol. Evol. 22:2472-2479.

Zsak L., Lu Z., Burrage T.G., Neilan J.G., Kutish G.F., Moore D.M., Rock D.L. (2001) African swine fever virus multigene family 360 and 530 genes are novel macrophage host range determinants. J. Virol. 75:3066-3076.

Table 1. 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 2. 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

Fig. 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.

Fig. 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.

Fig. 3. 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.

Fig. 4. 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.

Fig. 5. 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.

Fig. 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 in either of the gene/branch. (b,d) Divergent promoter regions from -55 to -1 upstream translational start sites of MGF360 and MGF505. The sequences with common signatures are highlighted with underline and the potential 5-nucleotide promoter motifs with double underline. (e,f) Distribution and enrichment of sites under divergent selection between paralogous pairs of genes/branches of MGF360 (e) and MGF505 (f). 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 (cylinders for α-helices, arrows for β-strands, and lines for coiled loops. A 25-codon sliding window plot of the site density 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 region indicated with horizontal bars.

Fig. 7. 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.

Fig. S1. Phylogenetic structure constructed for the core genome of ASFV (a) and the C-terminal 414 bp of the structural gene p72 presented in a dendrogram tree (b). The tree for the p72 was built from the isolates 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 without consensus. The isolate names in the p72 tree were presented as the combination of accession number, location, time and genotype.

Fig. 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.

Fig. S3. 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 tandem α-helices.

Fig. S4. The heatmap of pair-wise nucleotide similarities of the orthologous and paralogous genes of MGF360 (a) and MGF505 (b) along with the phylogenetic structure. The phylogenetic structure was inferred using Neighbor-Joining method with 1000 bootstraps for the orthologs and paralogs for each member of MGF360 and MGF505. The sequence alignments used for the phylogeny are provided in the public link https://figshare.com/projects/ASFV_alignment/82718 (see “Data availability”). Only nodes with the support value > 30 in the phylogeny 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 members, and therefore the former cluster together in the trees, except three isolates of MGF360-1L (from Kenya-1950, Ken05-Tk1, and Spain-E75), which cluster together with MGF360-2L, and five isolate of MGF505-7R (from Malawi-Lil83, Kenya-1950, Ken05-Tk1, Ken06-Bus, and

UgandaN10-2015), which cluster together with MGF505-6R.

Fig. S5. Correlation between the promoter divergence (y axis) and the synonymous substitution rate for each pair of genes/branches (x axis) in MGF360 (a) and MGF505 (b). The fitted lines of linear regression are shown in red and the fitting equation and Pearson correlations R^2 are indicated.

Fig. 1

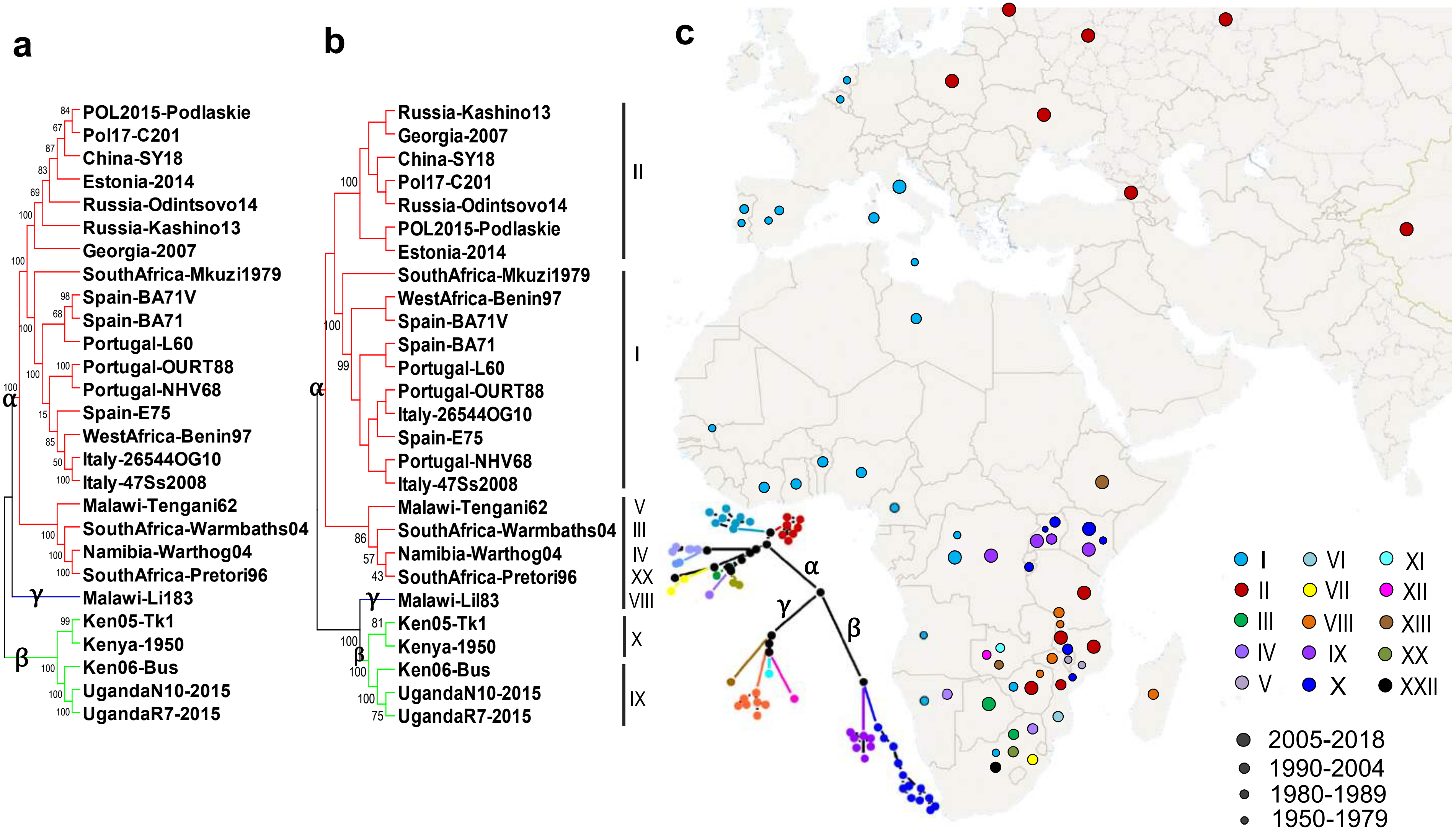

a
84
POL2015-Podlaskie
67
Pol17-C201
87
China-SY18
83
Estonia-2014
69
Russia-Odintsovo14
100
Russia-Kashino13
Georgia-2007
100
SouthAfrica-Mkuzi1979
98
Spain-BA71V
68
Spain-BA71
100
Portugal-L60
100
Portugal-OURT88
100
Portugal-NHV68
100
α
15
Spain-E75
85
WestAfrica-Benin97
50
Italy-26544OG10
100
Italy-47Ss2008
Malawi-Tengani62
100
SouthAfrica-Warmbaths04
100
Namibia-Warthog04
100
SouthAfrica-Pretori96
γ
Malawi-Li183
99
Ken05-Tk1
Kenya-1950
β
100
Ken06-Bus
100
UgandaN10-2015
100
UgandaR7-2015
b
Russia-Kashino13
Georgia-2007
China-SY18
100
Pol17-C201
Russia-Odintsovo14
POL2015-Podlaskie
Estonia-2014
II
SouthAfrica-Mkuzi1979
WestAfrica-Benin97
100
Spain-BA71V
99
Spain-BA71
Portugal-L60
α
Portugal-OURT88
Italy-26544OG10
Spain-E75
Portugal-NHV68
Italy-47Ss2008
I
Malawi-Tengani62
V
86
SouthAfrica-Warmbaths04
III
57
Namibia-Warthog04
IV
43
SouthAfrica-Pretori96
XX
γ
Malawi-Lil83
VIII
81
Ken05-Tk1
X
100
β
Kenya-1950
100
Ken06-Bus
100
UgandaN10-2015
75
UgandaR7-2015
IX
c
α
γ
β
● I
● II
● III
● IV
● V
● VI
● VII
● VIII
● IX
● X
● XI
● XII
● XIII
● XX
● XXII
● 2005-2018
● 1990-2004
● 1980-1989
● 1950-1979

Fig. 2

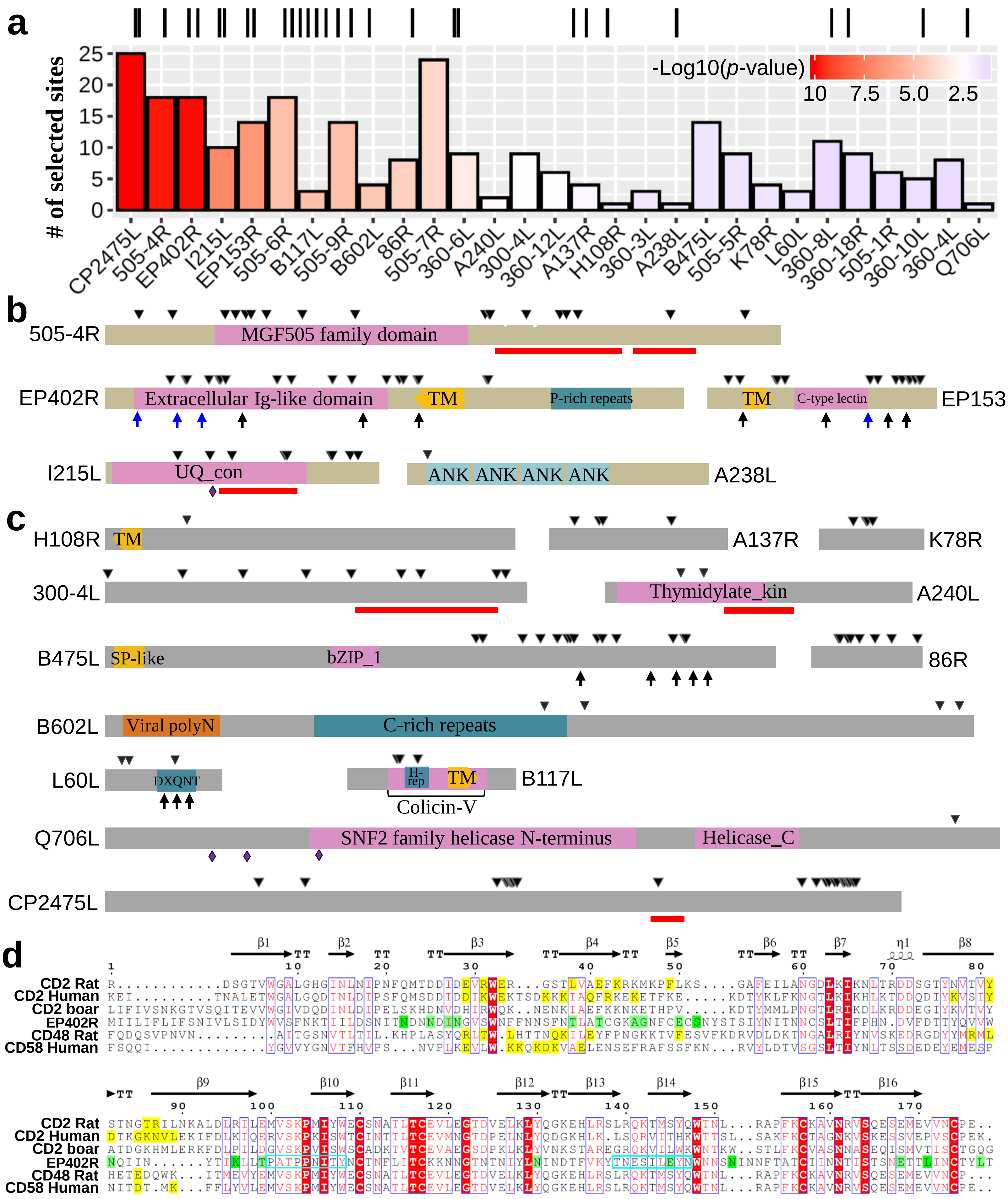
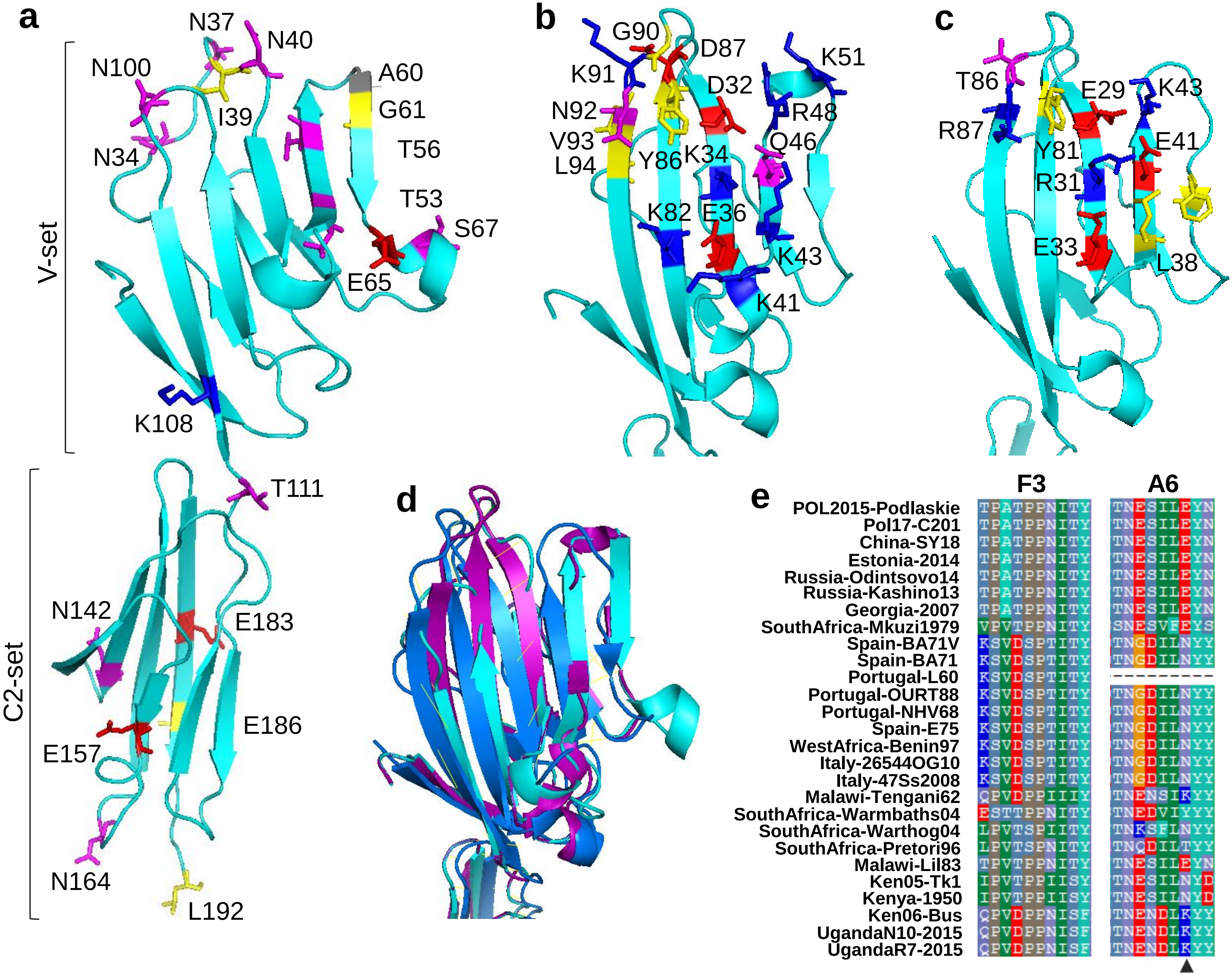

a
# of selected sites
25
20
15
10
5
0
-Log10(p-value)
10
7.5
5.0
2.5
CP2475L
505-4R
EP402R
I215L
EP153R
505-6R
B117L
505-9R
B602L
86R
505-7R
360-6L
A240L
300-4L
360-12L
A137R
H108R
360-3L
A238L
B475L
505-5R
K78R
L60L
360-8L
360-18R
505-1R
360-10L
360-4L
Q706L
b
505-4R
MGF505 family domain
EP402R
Extracellular 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
β1
β2
β3
β4
β5
β6
β7
η1
β8
TT
1
10
20
30
40
50
60
70
80
CD2 Rat R............DSGTVWGALGHGINLNIPNFQMTDDIDEVRWER...GSTLVAEFKRKMKPFLKS....GAFEILANGDLKIKNLTRDDSGTYNVTVY
CD2 Human KEI.........TNALETWGALGQDINLDIPSFQMSDDIDDIKWEKTSDKKKIAQFRKEKETFKE.....KDTYKLFKNGTLKIKHLKTDDQDIYKVSIY
CD2 boar LIFIVSNKGTVSQITEVVWGIVDQDINLDIPELSKHDNVDHIRWQK..NENKIAEFKKNKETHPV.....KDTYMMLPNGTLRIKDLKRDDEGIYKVTVY
EP402R MIILIFLIFSNIVLSIDYWVSFNKTIILDSNITNDNNDINGVSWNFFNNSFNTLATCGKAGNFCECSNYSTSIYNITNNCSLTIFPHN.DVFDTTYQVVW
CD48 Rat FQDQSVPNVN.........AITGSNVTLTIL.KHPLASYQRLTW.LHTTNQKILEYFPNGKKTVFESVFKDRVDLDKTNGALRIYNVSKEDRGDYYMRML
CD58 Human FSQQI.............YGVVYGNVTFHVPS...NVPLKEVLW.KKQKDKVAELENSEFRAFSSFKN...RVYLDTVSGSLTIYNLTSSDEDEYEMESP
β9
β10
β11
β12
β13
β14
β15
β16
90
100
110
120
130
140
150
160
170
CD2 Rat STNGTRILNKALDLRILEMVSKPMIYWECSNATLTCEVLEGTDVELKLYQGKEHLRSLRQKTMSYQWTNL...RAPFKCKAVNRVSQESEMEVVNCPE..
CD2 Human DTKGKNVLEKIFDLKIQERVSKPKISWTCINTTLTCEVMNGTDPELNLYQDGKHLK.LSQRVITHKWTTSL..SAKFKCTAGNKVSKESSVEPVSCPEK.
CD2 boar ATDGKHMLERKFDLPILDGVSKPVISWSCADKTVTCEVAEGSDPKLKLYVNKSTAREGRQKVILWKWNTKW..STLFKCVASNNASEQISMVTISCTGQG
EP402R NQIIN......YTIKLLTPATPPNITYNCTNFLITCKKNNGTNTNIYLNINDTFVKYTNESILEYNWNNSNINNFTATCIINNTISTSNETTLINCTYLT
CD48 Rat HETEDQWK...ITMEVYEMVSKPMIYWECSNATLTCEVLEGTDVELKLYQGKEHLRSLRQKTMSYQWTNL...RAPFKCKAVNRVSQESEMEVVNCP...
CD58 Human NITDT.MK...FFLYVLEMVSKPMIYWECSNATLTCEVLEGTDVELKLYQGKEHLRSLRQKTMSYQWTNL...RAPFKCKAVNRVSQESEMEVVNCPE..

Fig. 3



	F3	A6
POL2015-Podlaskie	TPATPPNITY	TNESILEYN
Pol17-C201	TPATPPNITY	TNESILEYN
China-SY18	TPATPPNITY	TNESILEYN
Estonia-2014	TPATPPNITY	TNESILEYN
Russia-Odintsovo14	TPATPPNITY	TNESILEYN
Russia-Kashino13	TPATPPNITY	TNESILEYN
Georgia-2007	TPATPPNITY	TNESILEYN
SouthAfrica-Mkuzi1979	VPVTPPNITY	SNESVFEYS
Spain-BA71V	KSVDSPTITY	TNGDILNYY
Spain-BA71	KSVDSPTITY	TNGDILNYY
Portugal-L60	KSVDSPTITY	---------
Portugal-OURT88	KSVDSPTITY	TNGDILNYY
Portugal-NHV68	KSVDSPTITY	TNGDILNYY
Spain-E75	KSVDSPTITY	TNGDILNYY
WestAfrica-Benin97	KSVDSPTITY	TNGDILNYY
Italy-26544OG10	KSVDSPTITY	TNGDILNYY
Italy-47Ss2008	KSVDSPTITY	TNGDILNYY
Malawi-Tengani62	QPVDPPIIIY	TNENSIKYY
SouthAfrica-Warmbaths04	ESTTPPNITY	TNEDVIYYY
SouthAfrica-Warthog04	LPVTSPIITY	TNKSFLNYY
SouthAfrica-Pretori96	LPVTSPNITY	TNQDILTYY
Malawi-Lil83	TPVTPPNITY	TNESILEYN
Ken05-Tk1	IPVTPPIISY	TNESILNYD
Kenya-1950	IPVTPPIISY	TNESILNYD
Ken06-Bus	QPVDPPNISF	TNENDLKYY
UgandaN10-2015	QPVDPPNISF	TNENDLKYY
UgandaR7-2015	QPVDPPNISF	TNENDLKYY

Fig. 4

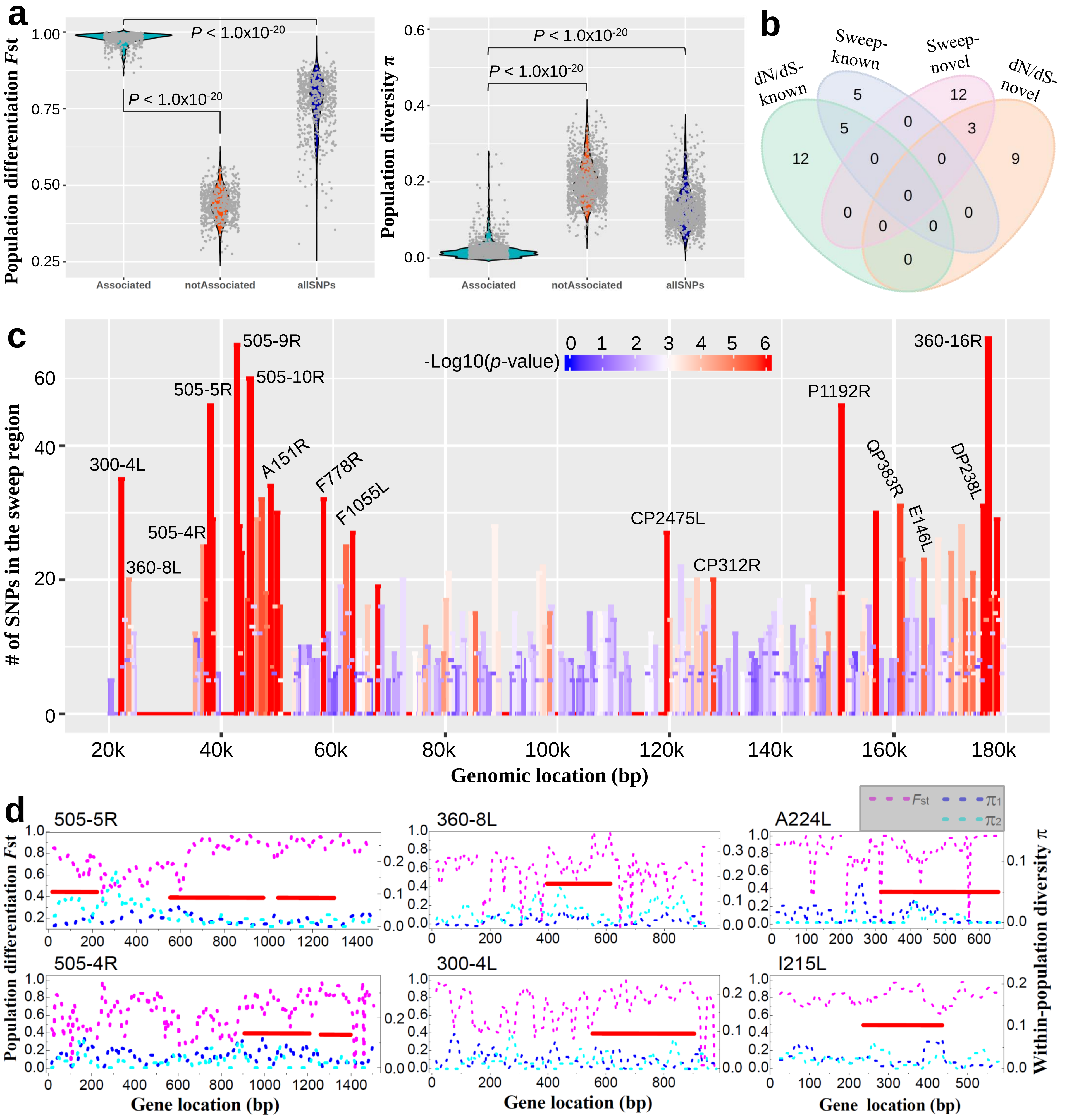
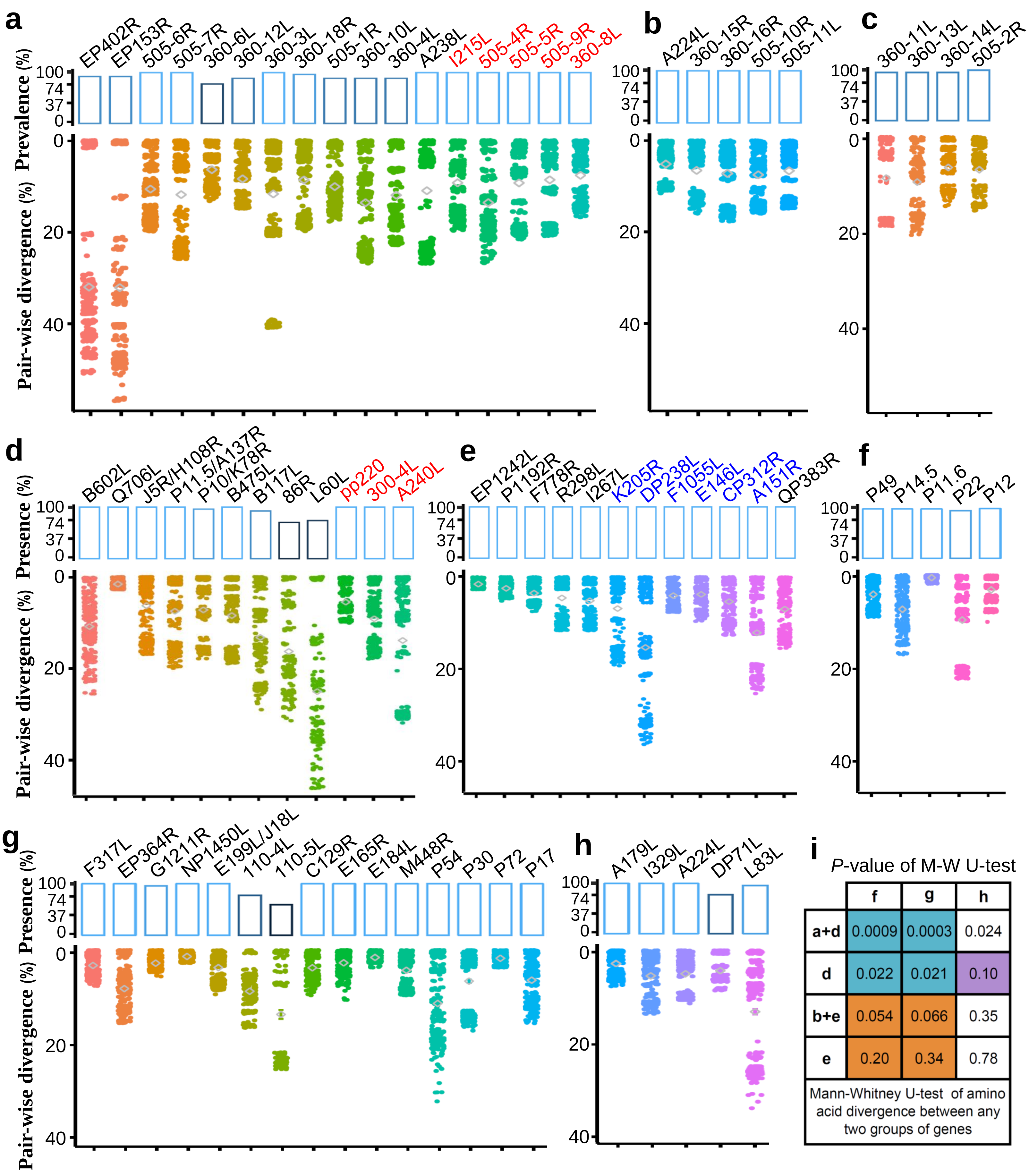

a
Population differentiation Fst
1.00
0.75
0.50
0.25
P < 1.0x10-20
P < 1.0x10-20
Associated
notAssociated
allSNPs
Population diversity π
0.6
0.4
0.2
0.0
P < 1.0x10-20
P < 1.0x10-20
Associated
notAssociated
allSNPs
b
dN/dS-known
Sweep-known
Sweep-novel
dN/dS-novel
12
5
12
9
5
0
3
0
0
0
0
0
0
0
0
0
0
c
# of SNPs in the sweep region
0
20
40
60
-Log10(p-value)
0 1 2 3 4 5 6
505-9R
505-10R
505-5R
300-4L
A151R
F778R
F1055L
505-4R
360-8L
360-16R
P1192R
QP383R
DP238L
E146L
CP2475L
CP312R
20k
40k
60k
80k
100k
120k
140k
160k
180k
Genomic location (bp)
d
Fst
π1
π2
Population differentiation Fst
Within-population diversity π
505-5R
360-8L
A224L
505-4R
300-4L
I215L
Gene location (bp)
Gene location (bp)
Gene location (bp)

Fig. 5



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

Fig. 6

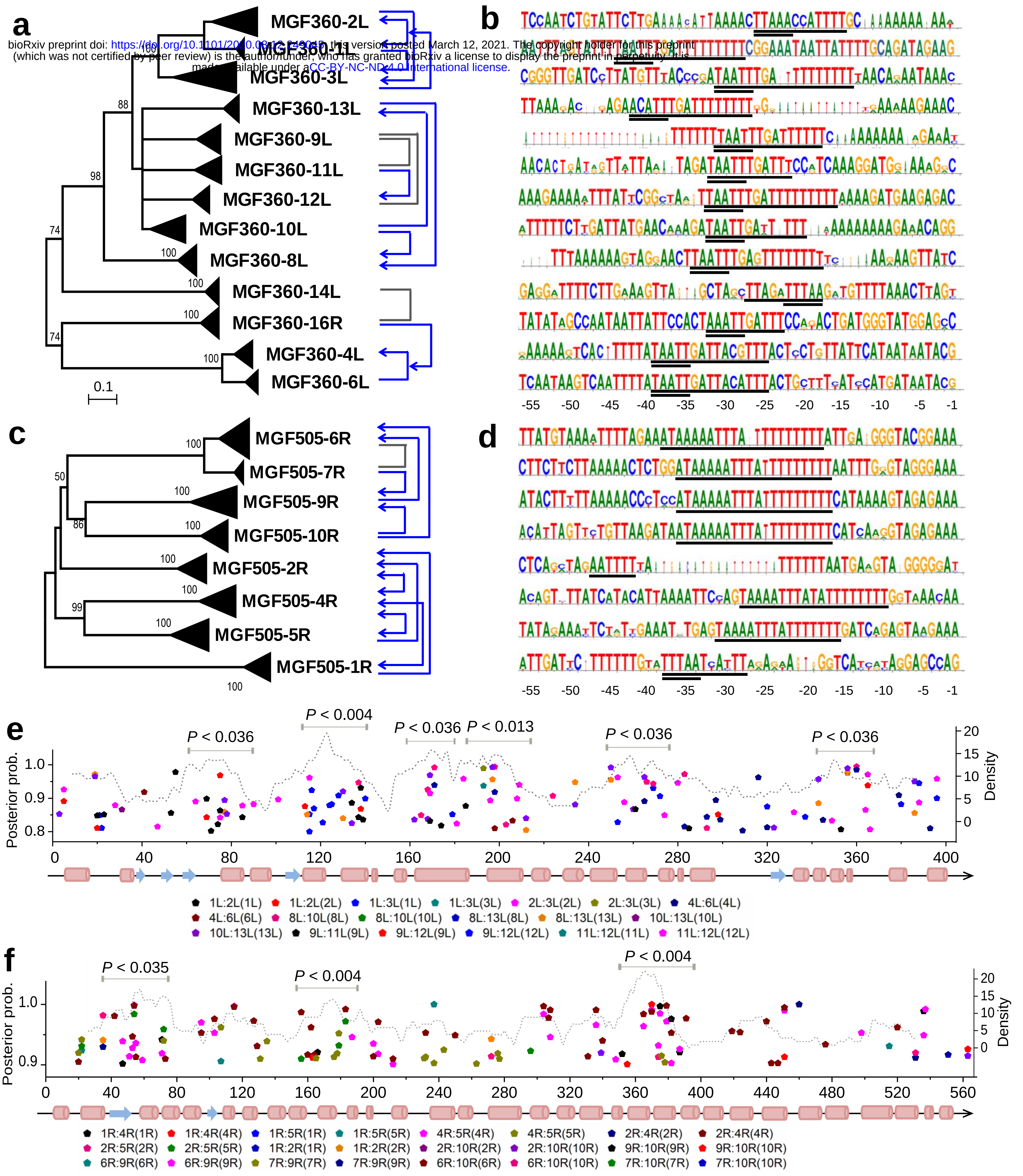

a
MGF360-2L
MGF360-1L
MGF360-3L
MGF360-13L
MGF360-9L
MGF360-11L
MGF360-12L
MGF360-10L
MGF360-8L
MGF360-14L
MGF360-16R
MGF360-4L
MGF360-6L
100
88
98
74
100
100
100
74
100
0.1
b
-55 -50 -45 -40 -35 -30 -25 -20 -15 -10 -5 -1
c
MGF505-6R
MGF505-7R
MGF505-9R
MGF505-10R
MGF505-2R
MGF505-4R
MGF505-5R
MGF505-1R
100
50
100
86
100
100
100
99
100
100
d
-55 -50 -45 -40 -35 -30 -25 -20 -15 -10 -5 -1
e
P < 0.036
P < 0.004
P < 0.036
P < 0.013
P < 0.036
P < 0.036
Posterior prob.
1.0
0.9
0.8
Density
20
15
10
5
0
0 40 80 120 160 200 240 280 320 360 400
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)
f
P < 0.035
P < 0.004
P < 0.004
Posterior prob.
1.0
0.9
Density
20
15
10
5
0
0 40 80 120 160 200 240 280 320 360 400 440 480 520 560
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)

Fig. 7

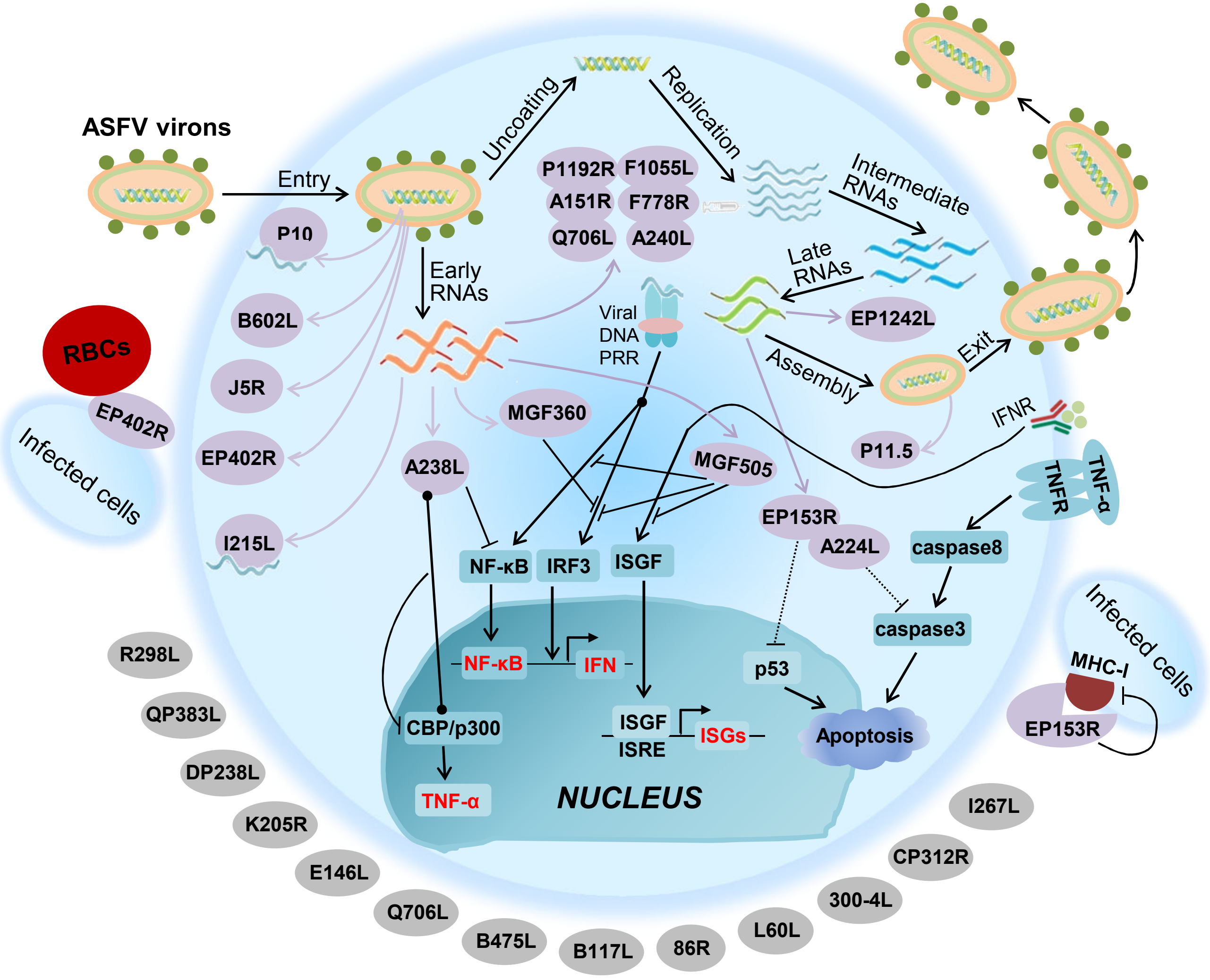

ASFV virons
Entry
Uncoating
Replication
Intermediate RNAs
Late RNAs
Early RNAs
Assembly
Exit
P1192R
F1055L
A151R
F778R
Q706L
A240L
P10
B602L
J5R
EP402R
I215L
RBCs
EP402R
Infected cells
Viral DNA PRR
MGF360
MGF505
A238L
EP1242L
P11.5
IFNR
TNFR
TNF-α
EP153R
A224L
caspase8
caspase3
NF-κB
IRF3
ISGF
NF-κB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

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

Genome accession	Virulence	Strain name	Name tag in this study	Isolation location
ASU18466	Low	BA71V	Spain-BA71V	Spain: Badajoz
KP055815	High	BA71	Spain-BA71V	Spain: Badajoz
KM262844	High	L60	Portugal-L60	Portugal
KM262845	Low	NHV	Portugal-NHV68	Portugal
AY261360	High	Kenya 1950	Kenya-1950	Kenya
KM111294	Moderate	Ken05/Tk1	Ken05-Tk1	Kenya central
KM111295	High	Ken06.Bus	Ken06-Bus	Kenya eastern
AY261366	Unknown	Warthog	Namibia-Warthog04	Namibia
AY261365	Unknown	Warmbaths	SouthAfrica-WarmBaths04	South Africa: Warmbaths
AY261364	High	Tengani 62	Malawi-Tengani62	Malawi: Tengani, Nsanje District
AY261363	High	Pretorisuskop/96/4	SouthAfrica-Pretori96	South Africa: Kruger National Park
AY261362	Unknown	Mkuzi 1979	SouthAfrica-Mkuzi1979	South Africa: Mkuzi Game Reserve
AY261361	High	Malawi Lil-20/1	Malawi-Lil83	Malawi: Chalaswa
AM712239	High	Benin 97/1	WestAfrica-Benin97	West Africa: Benin
AM712240	Low	OURT 88/3	Portugal-OURT88	Portugal
FN557520	High	E75	Spain-E75	Spain: Lerida
FR682468	High	Georgia 2007/1	Georgia-2007	Georgia
KX354450	High	47/Ss/2008	Italy-47Ss2008	Italy: Province of Sassari, Sardinia
KM102979	High	26544/OG10	Italy-26544OG10	Italy
KJ747406	Medium	Kashino 04/13	Russia-Kashino13	Russia
LS478113	Unknown	Estonia 2014	Estonia-2014	Estonia
KP843857	High	Odintsovo_02/14	Russia-Odintsovo14	Russia
MH025916	Unknown	R8	UgandaR8-2015	Uganda: Tororo district
MH025917	Unknown	R7	UgandaR7-2015	Uganda: Tororo district
MH025918	Unknown	R25	UgandaR25-2015	Uganda: Tororo district
MH025919	Unknown	N10	UgandaN10-2015	Uganda: Tororo district
MH025920	Unknown	R35	UgandaR35-2015	Uganda: Tororo district
MH681419	High	POL/2015/Podlaskie	POL2015-Podlaskie	Poland
MG939583	Unknown	Pol16_20186_o7	Pol16-o23	Poland
MG939584	Unknown	Pol16_20538_o9	Pol16-o9	Poland
MG939585	Unknown	Pol16_20540_o10	Pol16-o10	Poland
MG939586	Unknown	Pol16_29413_o23	Pol16-o23	Poland
MG939587	Unknown	Pol17_03029_C201	Pol17-C201	Poland
MG939588	Unknown	Pol17_04461_C210	Pol17-C210	Poland
MG939589	Unknown	Pol17_05838_C220	Pol17-C220	Poland
MH766894	Unknown	ASFV-SY18	China-SY18	China

Note: The strains highlighted in grey are excluded from analysis due to the high similarity with other strains of the same subtypes. Therefore, a total of 27 non-redundant strains were used for this study.

Table S2. List of unique non-synonymous mutations in the strains with low virulence.

Gene	AA change	Nucleotide change	Gene function
F1055L	Met114Ile	342C>T	Helicase superfamily II
C147L	Met11Val	33T>C	RNA polymerase subunit 6
B119L	Phe18Ser	54A>G	Component of redox pathway
CP204L/P30	Thr116Ala	348T>C	Phosphoprotein binds to ribonucleoprotein-K
CP312R	Pro190His	569C>A	Hypothetical protein
NP868R	Asp182Gly	546A>G	Guanylyl transferase (for mRNA modification)
NP868R	Gln259Arg	777A>G	
E199L/J18L	Gln61Arg	183T>C	Transmembrane domain containing protein
E120R/P14.5	His111Arg	333A>G	DNA-binding. Required for movement of virions to plasma membrane
I215L	Glu95Gly	285T>C	Ubiquitin conjugating enzyme
MGF505-11L	His251Arg	753T>C	Multigene family 505
MGF505-11L	Leu213Phe	639C>G	
MGF505-11L	Lys131Arg	393T>C	

Note: The unique mutation is defined as those occur in the two ASFV strains with low virulence (Portugal-NHV68 and Portugal-OURT88). The positions are relative to the strain Georgia-2007.

Table S3. 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 S4. Functional domain identification of the genes enriched with non-synonymous mutations (*E*-value ≤ 0.03 or score ≥ 20).

Gene	Mapped start	Mapped end	Score	E-value	PFAM accession	PFAM name	PFAM function
MGF505-4R	87	279	272.1	2.0E-81	PF03158.8	DUF249	Multigene family 505 protein
MGF505-5R	87	275	277.5	4.4E-83	PF03158.8	DUF249	Multigene family 505 protein
MGF505-6R	87	284	257	8.5E-77	PF03158.8	DUF249	Multigene family 505 protein
MGF505-9R	87	275	287.9	2.9E-86	PF03158.8	DUF249	Multigene family 505 protein
MGF505-10R	87	279	276.5	8.7E-83	PF03158.8	DUF249	Multigene family 505 protein
MGF505-11L	86	278	196.2	3.6E-58	PF03158.8	DUF249	Multigene family 505 protein
MGF360-8L	96	280	251.9	3.8E-75	PF01671.11	ASFV_360	Multigene family 360
MGF360-15R	173	262	18.2	1.3E-03	PF01671.11	ASFV_360	Multigene family 360
MGF360-16R	102	303	254.4	6.5E-76	PF01671.11	ASFV_360	Multigene family 360
I215L	7	137	154.9	7.8E-46	PF00179.21	UQ_con	Ubiquitin-conjugating enzyme
K205R	5	70	13.3	2.4E-02	PF08317.6	Spc7	Spc7 kinetochore protein
K205R	21	74	14.6	1.0E-02	PF02646.11	RmuC	RmuC family
E183L/P54	1	184	379.6	2.1E-114	PF05568.6	ASFV_J13L	African swine fever virus J13L protein
E183L/P54	29	78	16.6	2.4E-03	PF09402.5	MSC	Man1-Src1p-C-terminal domain
E183L/P54	30	73	20.9	1.7E-04	PF10717.4	ODV-E18	Occlusion-derived virus envelope protein ODV-E18
E183L/P54	31	65	16	5.5E-03	PF07423.6	DUF1510	Protein of unknown function (DUF1510)
E183L/P54	32	58	15.1	1.3E-02	PF02009.11	Rifin_STEVOR	Rifin/stevor family
E183L/P54	32	71	15.7	1.4E-02	PF14575.1	EphA2_TM	Ephrin type-A receptor 2 transmembrane domain
A240L	8	180	143.5	4.4E-42	PF02223.12	Thymidylate_kin	Thymidylate kinase
EP364R	35	147	40.2	2.6E-10	PF02732.10	ERCC4	ERCC4 domain
I177L	4	73	14.4	0.016	PF09529.5	Intg_mem_TP0381	Integral membrane domain

Table S5. Genes with the value of dN/dS lower than the average (dN/dS < 0.1) using the Nei & Gojobori method.

Gene	dN/dS	Gene function	Functional category
NP1450L	0.100	RNA polymerase subunit 1	Transcription
A859L	0.098	Helicase superfamily II	Transcription
NP419L	0.093	DNA ligase	DNA replication
E165R	0.093	dUTPase	DNA metabolism
M1249L	0.090	Ubiquitin-like domain containing protein	
D205R	0.087	RNA polymerase subunit 5	Transcription
P1192R	0.085	Topoisomerase II	DNA replication
H359L	0.082	RNA polymerase subunit 3	Transcription
NP868R	0.076	mRNA guanylyltransferase	Transcription
F1055L	0.074	Helicase superfamily II	Transcription
F778R	0.072	Ribonucleotide reductase large subunit	Nucleotide metabolism, transcription, replication and repair
C147L	0.071	RNA polymerase subunit 6	Transcription
S273R	0.069	Ulp1 protease Family	Structural protein
B263R	0.068	TATA-box binding-like protein	Nucleotide metabolism, transcription, replication and repair
E184L	0.066	Hypothetical protein	
C962R	0.064	Putative DNA primase	DNA replication
F334L	0.061	Ribonucleotide reductase small subunit	Nucleotide metabolism, transcription, replication and repair
EP424R	0.060	FTS J-like Methyltransferase domain containing protein	Nucleotide metabolism, transcription, replication and repair
B385R	0.054	A2L-like transcription factor	Transcription
B646L/P72	0.046	Structural protein P72	Structural protein
CP80R	0.040	RNA polymerase subunit 10	Transcription
CP530R	0.039	60 kDa polyprotein	Structural protein
E301R	0.035	Proliferating cell nuclear antigen-like protein	DNA replication
B125R	0.028	E2 early regulatory protein	Regulator of transcription and DNA replication
C315R	0.025	TFIIB like	Transcription
B354L	0.022	P-loop-containing nucleoside triphosphate hydrolases	Energy metabolism
EP1242L	0.019	RNA polymerase subunit 2	Transcription
A104R/P11.6	0.004	Histone-like structural protein	Structural protein

Table S6. Genes with positive selection signals at a fraction of sites with ω (dN/dS) >1 based on the likelihood ratio tests.

Gene	Parameters for M2		Parameters for M8		M2 vs. M1		M8 vs. M7	
					LRT[a]	*p*-value	LRT[a]	*p*-value
CP2475L	$p_2 = 0.013$,	$\omega_2 = 4.642$	$p_1 = 0.018$,	$\omega = 4.001$	63.492	1.63E-14	78.773	<1E-20
MGF505-4R	$p_2 = 0.040$,	$\omega_2 = 5.500$	$p_1 = 0.044$,	$\omega = 5.371$	43.337	3.89E-10	45.324	1.44E-10
EP402R	$p_2 = 0.046$,	$\omega_2 = 4.297$	$p_1 = 0.065$,	$\omega = 3.117$	34.179	3.78E-08	45.886	1.09E-10
I215L	$p_2 = 0.057$,	$\omega_2 = 6.255$	$p_1 = 0.060$,	$\omega = 6.008$	31.813	1.24E-07	32.542	8.58E-08
EP153R	$p_2 = 0.104$,	$\omega_2 = 4.427$	$p_1 = 0.113$,	$\omega = 3.639$	30.659	2.20E-07	30.156	2.83E-07
MGF505-6R	$p_2 = 0.044$,	$\omega_2 = 3.617$	$p_1 = 0.054$,	$\omega = 3.351$	20.746	3.13E-05	23.032	9.97E-06
B117L	p_2= 0.025,	$\omega_2 = 246.8$	$p_1 = 0.025$,	$\omega = 232.2$	23.076	9.75E-06	22.893	1.07E-05
MGF505-9R	$p_2 = 0.049$,	$\omega_2 = 4.534$	$p_1 = 0.054$,	$\omega = 4.334$	22.579	1.25E-05	22.881	1.08E-05
B602L	$p_2 = 0.046$,	$\omega_2 = 3.788$	$p_1 = 0.051$,	$\omega = 3.665$	19.422	6.06E-06	21.450	2.20E-05
86R	$p_2 = 0.131$,	$\omega_2 = 10.373$	$p_1 = 0.131$,	$\omega = 10.373$	19.944	4.67E-05	19.976	4.60E-05
MGF505-7R	$p_2 = 0.057$,	$\omega_2 = 3.524$	$p_1 = 0.067$,	$\omega = 3.318$	17.985	1.24E-04	18.862	8.02E-05
MGF360-6L	$p_2 = 0.016$,	$\omega_2 = 5.758$	$p_1 = 0.027$,	$\omega = 4.210$	10.329	0.006	15.378	4.58E-04
J5R /H108R	$p_2 = 0.169$,	$\omega_2 = 4.545$	$p_1 = 0.169$,	$\omega = 4.545$	12.933	0.043	8.404	0.015
A240L	$p_2 = 0.011$,	$\omega_2 = 7.581$	$p_1 = 0.0176$,	$\omega = 5.558$	6.306	0.005	12.313	0.002
MGF300-4L	$p_2 = 0.091$,	$\omega_2 = 2.927$	$p_1 = 0.099$,	$\omega = 2.872$	11.274	0.004	11.088	0.004
MGF360-12L	$p_2 = 0.015$,	$\omega_2 = 7.741$	$p_1 = 0.015$,	$\omega = 7.723$	10.806	0.005	10.810	0.004
P11.5/A137R	$p_2 = 0.177$,	$\omega_2 = 2.839$	$p_1 = 0.177$,	$\omega = 2.843$	9.740	0.008	10.283	0.006
MGF360-3L	$p_2 = 0.007$,	$\omega_2 = 9.342$	$p_1 = 0.007$,	$\omega = 8.874$	7.301	0.026	8.288	0.016
A238L	$p_2 = 0.005$,	$\omega_2 = 12.113$	$p_1 = 0.005$,	$\omega = 11.777$	6.622	0.036	7.839	0.020
B475L	$p_2 = 0.035$,	$\omega_2 = 3.895$	$p_1 = 0.118$,	$\omega = 2.358$	7.786	0.020	7.750	0.021
MGF505-5R	$p_2 = 0.035$,	$\omega_2 = 3.128$	$p_1 = 0.046$,	$\omega = 2.905$	7.085	0.029	7.676	0.022
P10/K78R	$p_2 = 0.060$,	$\omega_2 = 8.793$	$p_1 = 0.060$,	$\omega = 8.856$	7.369	0.025	7.593	0.022
L60L	$p_2 = 0.136$,	$\omega_2 = 2.763$	$p_1 = 0.144$,	$\omega = 2.742$	4.954	0.084	7.567	0.023
MGF360-8L	$p_2 = 0.217$,	$\omega_2 = 1.774$	$p_1 = 0.216$,	$\omega = 1.779$	5.819	0.055	6.288	0.043
MGF360-18R	$p_2 = 0.118$,	$\omega_2 = 2.123$	$p_1 = 0.124$,	$\omega = 2.152$	5.416	0.067	6.947	0.031
MGF505-1R	$p_2 = 0.015$,	$\omega_2 = 4.248$	$p_1 = 0.023$,	$\omega = 3.667$	4.871	0.088	6.812	0.033
MGF360-10L	$p_2 = 0.166$,	$\omega_2 = 1.352$	$p_1 = 0.163$,	$\omega = 1.365$	1.517	0.468	6.942	0.031
MGF360-4L	$p_2 = 0.041$,	$\omega_2 = 2.254$	$p_1 = 0.083$,	$\omega = 1.921$	2.809	0.245	6.267	0.044
Q706L	$p_2 = 0.002$,	$\omega_2 = 6.808$	$p_1 = 0.002$,	$\omega = 6.861$	4.258	0.119	6.737	0.034

[a] LRT is the likelihood ratio test statistic calculated as $2\Delta l$ with l the log likelihood for each model. The *p*-value was calculated using Chi-squared test.

Table S7. Three categories of proteins used for comparison of sequence variability.

Protein name	Antigenic?	Function
Structural proteins not shown to be under positive selection		
P49/B438L	-	Minor capsid protein
P14.5/E120R	-	DNA-binding protein
P11.6/A104R	-	Histone-like DNA-binding protein
P22/KP177R	-	Inner envelop protein // virus entry
P12/O61R	-	Inner envelop protein
Proteins shown to be antigenic in immunoassays		
MGF110-4L	+	Multigene family 110
MGF110-5L	+	Multigene family 110
C129R	+	Mn-dependent superoxide dismutase
E165R	+	dUTPase
E184L	+	Hypothetical protein
M448R	+	Microbody targeting signal-containing protein
F317L	+	Hypothetical protein
EP364R	+	ERCC4 nuclease domain
G1211R	+	DNA polymerase family B
NP1450L	+	RNA polymerase subunit 1
E199L/J18L	+	Inner envelop protein // virus entry
P54/E183L	+	Inner envelop protein
P30/CP204L	+	Phosphoprotein binding to ribonucleoprotein K
P72/B464L	+	Major capsid protein
P17/D117L	+	Inner envelop protein
Proteins previously shown to be involved in host-cell interactions		
A179L	-	Bcl 2 apoptosis inhibitor
I329L	+	Putative inhibitor of TLR3 signaling pathway
A224L	-	IAP apoptosis inhibitor
DP71L	+	Similar to herpes simplex virus ICP34.5 protein
L83L	-	Putative IL-1b binding protein

Table S8. 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.

Tested pairs	Model A					Model A null			Model A vs. null	
	p_1	p_{2a}	p_{2b}	ω_{2a}	ω_{2b}	p_1	p_{2a}	p_{2b}	LRT[a]	*p*-value
MGF360-1L vs. 2L (1L)	0.4714	0.0183	0.0175	4.4375	4.4375	0.4715	0.0171	0.0163	10.0124	0.0067
MGF360-1L vs. 2L (2L)	0.4788	0.0167	0.0164	4.8263	4.8263	0.4340	0.0411	0.0365	15.2116	0.0005
MGF360-1L vs. 3L (1L)	0.1823	0.2683	0.1117	1.2538	1.2538	0.1747	0.2918	0.1248	1.3494	0.5093
MGF360-1L vs. 3L (3L)	0.4672	0.0157	0.0146	4.3621	4.3621	0.4341	0.0419	0.0374	8.8637	0.0119
MGF360-2L vs. 3L (2L)	0.4691	0.0243	0.0236	4.3006	4.3006	0.2005	0.2655	0.1325	-9.7854	Nil
MGF360-2L vs. 3L (3L)	0.5050	0.0051	0.0053	7.1295	7.1295	0.4157	0.0591	0.0519	7.7830	0.0204
MGF360-4L vs. 6L (4L)	0.2244	0.1116	0.0401	1.9819	1.9819	0.2126	0.1485	0.0540	14.5861	0.0007
MGF360-4L vs. 6L (6L)	0.3530	0.0305	0.0180	3.1154	3.1154	0.2746	0.1300	0.0676	-1.8117	Nil
MGF360-8L vs. 10L (8L)	0.1853	0.1043	0.0283	2.3876	2.3876	0.1543	0.2058	0.0542	10.0030	0.0067
MGF360-8L vs. 10L (10L)	0.2272	0.0443	0.0141	1.9266	1.9266	0.1904	0.1208	0.0352	-1.2883	Nil
MGF360-8L vs. 13L (8L)	0.3448	0.0477	0.0284	3.3790	3.3790	0.2372	0.1981	0.1014	7.1805	0.0276
MGF360-8L vs. 13L (13L)	0.2814	0.0789	0.0368	3.0148	3.0148	0.2001	0.2358	0.1021	5.0706	0.0792
MGF360-10L vs. 13L (10L)	0.2358	0.1560	0.0681	1.0000	1.0000	0.2358	0.1560	0.0681	0.0000	1.0000
MGF360-10L vs. 13L (13L)	0.2696	0.0980	0.0450	3.0578	3.0578	0.2018	0.2196	0.0909	18.6248	9.0E-05
MGF360-9L vs. 11L (9L)	0.1734	0.2448	0.0855	1.0000	1.0000	0.1734	0.2448	0.0855	0.0000	1
MGF360-9L vs. 11L (11L)	0.2993	0.0834	0.0435	1.0000	1.0000	0.2993	0.0834	0.0435	0.0000	1.0000
MGF360-9L vs. 12L (9L)	0.1722	0.2625	0.0964	1.0000	1.0000	0.1722	0.2625	0.0964	0.0000	1.0000
MGF360-9L vs. 12L (12L)	0.2552	0.0889	0.0366	1.9514	1.9514	0.2362	0.1363	0.0563	4.5839	0.1011
MGF360-11L vs. 12L (11L)	0.2372	0.0659	0.0232	1.0000	1.0000	0.2372	0.0659	0.0232	0.0000	1.0000
MGF360-11L vs. 12L (12L)	0.1543	0.1224	0.0271	2.1345	2.1345	0.1414	0.1922	0.0436	7.5229	0.0232
MGF360-14L vs. 16R (14L)	0.1711	0.2513	0.0878	1.0000	1.0000	0.1711	0.2513	0.0878	0.0000	1
MGF360-14L vs. 16R (16R)	0.1514	0.2951	0.0980	1.0000	1.0000	0.1514	0.2951	0.0980	0.0000	1.0000
MGF360-1L2L vs. 3L (1L2L)	0.2348	0.1683	0.0758	1.6554	1.6554	0.1796	0.2532	0.0967	12.6640	0.0018
MGF360-1L2L vs. 3L (3L)	0.3951	0.0883	0.0799	1.0000	1.0000	0.3951	0.0883	0.0799	0.0000	1.0000
MGF360-4L6L vs. 16R (4L6L)	0.1239	0.2061	0.0406	1.4131	1.4131	0.1116	0.2546	0.0485	7.1307	0.0283
MGF360-4L6L vs. 16R (16R)	0.31775	0.0108	0.00515	3.45243	3.45243	0.2142	0.1115	0.0375	5.5447	0.0625
MGF505-1R vs. 4R (1R)	0.45069	0.01305	0.0112	3.95315	3.95315	0.3583	0.0878	0.0643	6.0140	0.0494
MGF505-1R vs. 4R (4R)	0.31299	0.10294	0.06167	2.98983	2.98983	0.1983	0.2583	0.1213	26.9201	1.4E-06
MGF505-1R vs. 5R (1R)	0.21734	0.19668	0.0854	1	1	0.2174	0.1966	0.0854	-5.4E-05	Nil

MGF505-1R vs. 5R (5R)	0.21678	0.16056	0.06209	1.37617	1.37617	0.2056	0.1988	0.0792	1.8239	0.4017
MGF505-4R vs. 5R (4R)	0.27519	0.14555	0.08027	2.66908	2.66908	0.2012	0.2658	0.134	25.6902	2.6E-06
MGF505-4R vs. 5R (5R)	0.45594	0.01751	0.01562	4.03558	4.03558	0.3846	0.0717	0.0566	12.0711	0.0024
MGF505-2R vs. 4R (2R)	0.40467	0.01935	0.01393	4.00212	4.00212	0.371	0.0514	0.0352	10.7772	0.0046
MGF505-2R vs. 4R (4R)	0.17107	0.21592	0.06774	2.20738	2.20738	0.1356	0.3269	0.1017	26.5228	1.7E-06
MGF505-2R vs. 5R (2R)	0.24959	0.03936	0.0141	3.16633	3.16633	0.2189	0.1058	0.0363	15.2186	0.0005
MGF505-2R vs. 5R (5R)	0.14866	0.22888	0.06055	1	1	0.1487	0.2289	0.0606	0	1
MGF505-2R vs. 1R (1R)	0.15462	0.21421	0.05776	1	1	0.1546	0.2142	0.0578	1.8E-05	1.0000
MGF505-2R vs. 1R (2R)	0.2557	0.04886	0.01845	2.87911	2.87911	0.2313	0.1061	0.0394	12.9997	0.0015
MGF505-2R vs. 10R (2R)	0.21476	0.12844	0.04509	1	1	0.2147	0.1285	0.0451	-4.4E-05	Nil
MGF505-2R vs. 10R (10R)	0.14742	0.20415	0.05032	1.20072	1.20072	0.1457	0.228	0.0585	0.6030	0.7397
MGF505-9R vs. 10R (9R)	0.27408	0.04728	0.01967	4.3231	4.3231	0.1791	0.2205	0.0752	32.5048	8.7E-08
MGF505-9R vs. 10R (10R)	0.16557	0.26257	0.09028	1	1	0.1656	0.2626	0.0903	0	1
MGF505-6R vs. 7R (6R)	0.36338	0.00366	0.00211	4.20616	4.20616	0.3643	0	0	0.8609	0.6502
MGF505-6R vs. 7R (7R)	0.36427	0	0	3.26102	3.26102	0.3643	0	0	1.8E-05	1.0000
MGF505-6R vs. 9R (6R)	0.2458	0.07828	0.02978	2.7473	2.7473	0.1764	0.2046	0.0652	11.0664	0.0040
MGF505-6R vs. 9R (9R)	0.16781	0.29106	0.11448	1	1	0.1678	0.291	0.1145	-2.6E-05	Nil
MGF505-7R vs. 9R (7R)	0.18129	0.19575	0.06342	1.49047	1.49047	0.1726	0.2436	0.0841	5.8248	0.0543
MGF505-7R vs. 9R (9R)	0.41359	0.01256	0.0092	8.49966	8.49966	0.2328	0.2126	0.1117	17.6709	0.0001
MGF505-6R vs. 10R (6R)	0.27971	0.0519	0.02247	3.15776	3.15776	0.2149	0.1437	0.0525	24.4340	4.9E-06
MGF505-6R vs. 10R (10R)	0.19109	0.23978	0.09706	1	1	0.1911	0.2398	0.0971	4E-06	1.0000
MGF505-7R vs. 10R (7R)	0.354	0.0558	0.03562	2.92338	2.92338	0.199	0.225	0.0926	5.2533	0.0723
MGF505-7R vs. 10R (10R)	0.24485	0.20909	0.12022	1	1	0.2449	0.2091	0.1202	-6.6E-05	Nil
MGF505-6R7R vs. 9R10R (6R7R)	0.21414	0.11373	0.03843	2.71635	2.71635	0.2115	0.1335	0.0464	103.7484	0
MGF505-6R7R vs. 9R10R (9R10R)	0.15043	0.36485	0.18029	1	1	0.1504	0.3649	0.1803	0	1

Note: The foreground in the tests is indicated in the parentheses.

[a] LRT is the likelihood ratio test statistic calculated as $2\Delta l$ with l the log likelihood for each model. The *p*-value was calculated using Chi-squared test. The *p*-value was shown as Nil if LRT is negative.

Fig. S1

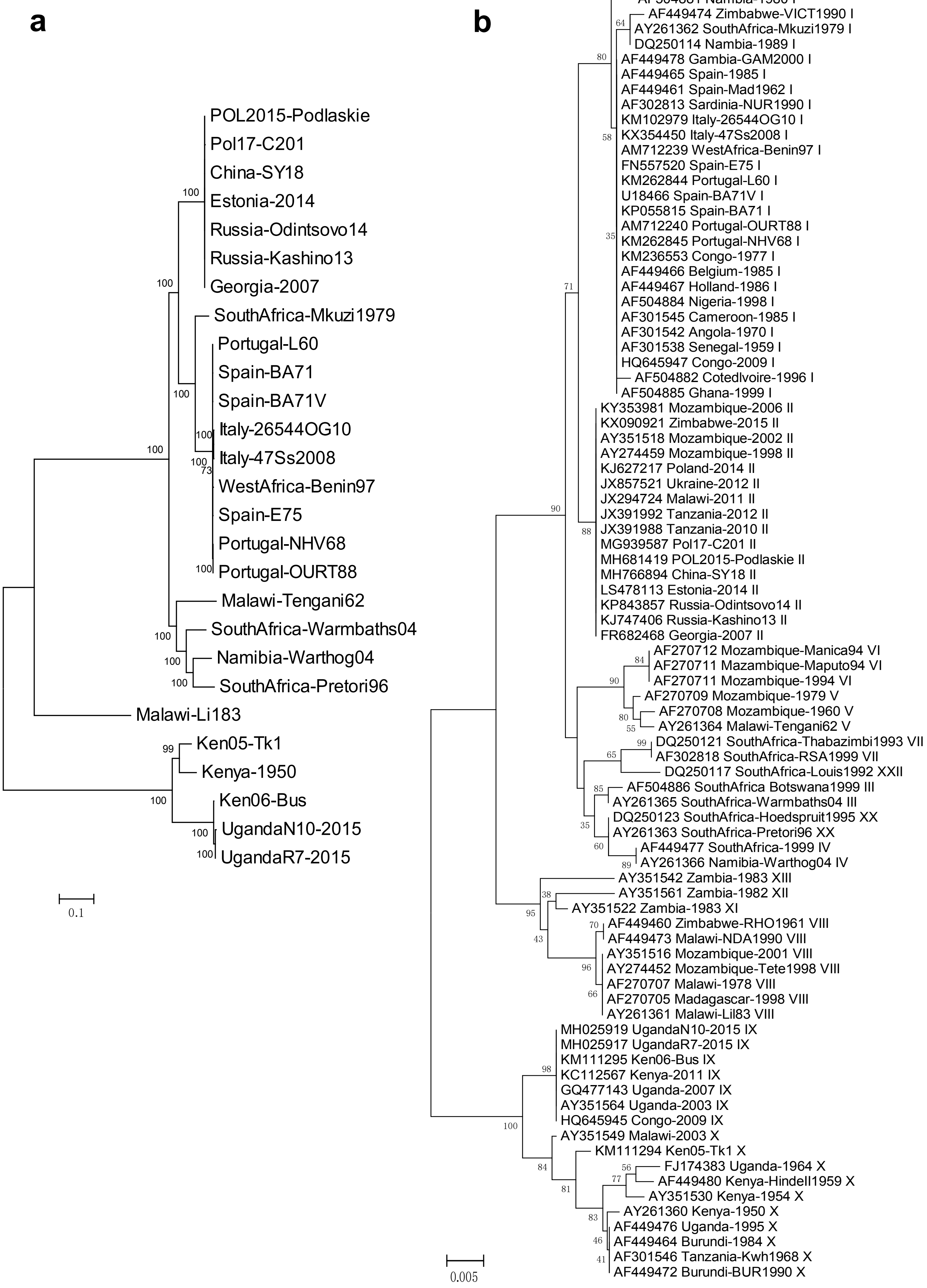

a
POL2015-Podlaskie
Pol17-C201
China-SY18
Estonia-2014
Russia-Odintsovo14
Russia-Kashino13
Georgia-2007
SouthAfrica-Mkuzi1979
Portugal-L60
Spain-BA71
Spain-BA71V
Italy-26544OG10
Italy-47Ss2008
WestAfrica-Benin97
Spain-E75
Portugal-NHV68
Portugal-OURT88
Malawi-Tengani62
SouthAfrica-Warmbaths04
Namibia-Warthog04
SouthAfrica-Pretori96
Malawi-Li183
Ken05-Tk1
Kenya-1950
Ken06-Bus
UgandaN10-2015
UgandaR7-2015
100
100
100
100
100
100
73
100
100
100
100
99
100
100
100
0.1
b
AF504881 Nambia-1980 I
AF449474 Zimbabwe-VICT1990 I
AY261362 SouthAfrica-Mkuzi1979 I
DQ250114 Nambia-1989 I
AF449478 Gambia-GAM2000 I
AF449465 Spain-1985 I
AF449461 Spain-Mad1962 I
AF302813 Sardinia-NUR1990 I
KM102979 Italy-26544OG10 I
KX354450 Italy-47Ss2008 I
AM712239 WestAfrica-Benin97 I
FN557520 Spain-E75 I
KM262844 Portugal-L60 I
U18466 Spain-BA71V I
KP055815 Spain-BA71 I
AM712240 Portugal-OURT88 I
KM262845 Portugal-NHV68 I
KM236553 Congo-1977 I
AF449466 Belgium-1985 I
AF449467 Holland-1986 I
AF504884 Nigeria-1998 I
AF301545 Cameroon-1985 I
AF301542 Angola-1970 I
AF301538 Senegal-1959 I
HQ645947 Congo-2009 I
AF504882 Cotedlvoire-1996 I
AF504885 Ghana-1999 I
KY353981 Mozambique-2006 II
KX090921 Zimbabwe-2015 II
AY351518 Mozambique-2002 II
AY274459 Mozambique-1998 II
KJ627217 Poland-2014 II
JX857521 Ukraine-2012 II
JX294724 Malawi-2011 II
JX391992 Tanzania-2012 II
JX391988 Tanzania-2010 II
MG939587 Pol17-C201 II
MH681419 POL2015-Podlaskie II
MH766894 China-SY18 II
LS478113 Estonia-2014 II
KP843857 Russia-Odintsovo14 II
KJ747406 Russia-Kashino13 II
FR682468 Georgia-2007 II
AF270712 Mozambique-Manica94 VI
AF270711 Mazambique-Maputo94 VI
AF270711 Mozambique-1994 VI
AF270709 Mozambique-1979 V
AF270708 Mozambique-1960 V
AY261364 Malawi-Tengani62 V
DQ250121 SouthAfrica-Thabazimbi1993 VII
AF302818 SouthAfrica-RSA1999 VII
DQ250117 SouthAfrica-Louis1992 XXII
AF504886 SouthAfrica Botswana1999 III
AY261365 SouthAfrica-Warmbaths04 III
DQ250123 SouthAfrica-Hoedspruit1995 XX
AY261363 SouthAfrica-Pretori96 XX
AF449477 SouthAfrica-1999 IV
AY261366 Namibia-Warthog04 IV
AY351542 Zambia-1983 XIII
AY351561 Zambia-1982 XII
AY351522 Zambia-1983 XI
AF449460 Zimbabwe-RHO1961 VIII
AF449473 Malawi-NDA1990 VIII
AY351516 Mozambique-2001 VIII
AY274452 Mozambique-Tete1998 VIII
AF270707 Malawi-1978 VIII
AF270705 Madagascar-1998 VIII
AY261361 Malawi-Lil83 VIII
MH025919 UgandaN10-2015 IX
MH025917 UgandaR7-2015 IX
KM111295 Ken06-Bus IX
KC112567 Kenya-2011 IX
GQ477143 Uganda-2007 IX
AY351564 Uganda-2003 IX
HQ645945 Congo-2009 IX
AY351549 Malawi-2003 X
KM111294 Ken05-Tk1 X
FJ174383 Uganda-1964 X
AF449480 Kenya-Hindell1959 X
AY351530 Kenya-1954 X
AY261360 Kenya-1950 X
AF449476 Uganda-1995 X
AF449464 Burundi-1984 X
AF301546 Tanzania-Kwh1968 X
AF449472 Burundi-BUR1990 X
64
80
58
35
71
90
88
84
90
80
55
99
65
85
35
60
89
38
95
43
70
96
66
98
100
84
81
56
77
83
46
41
0.005

Fig. S2

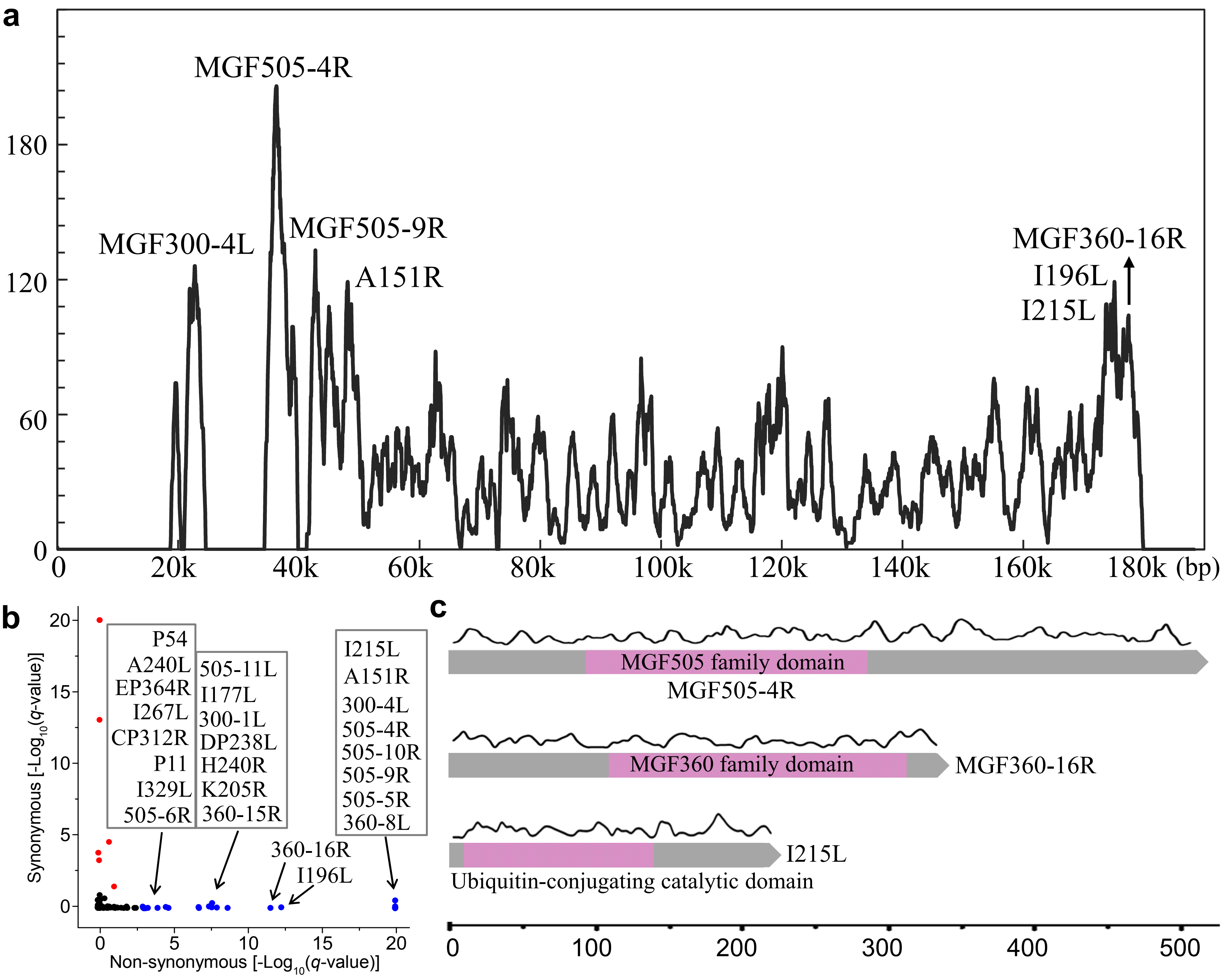
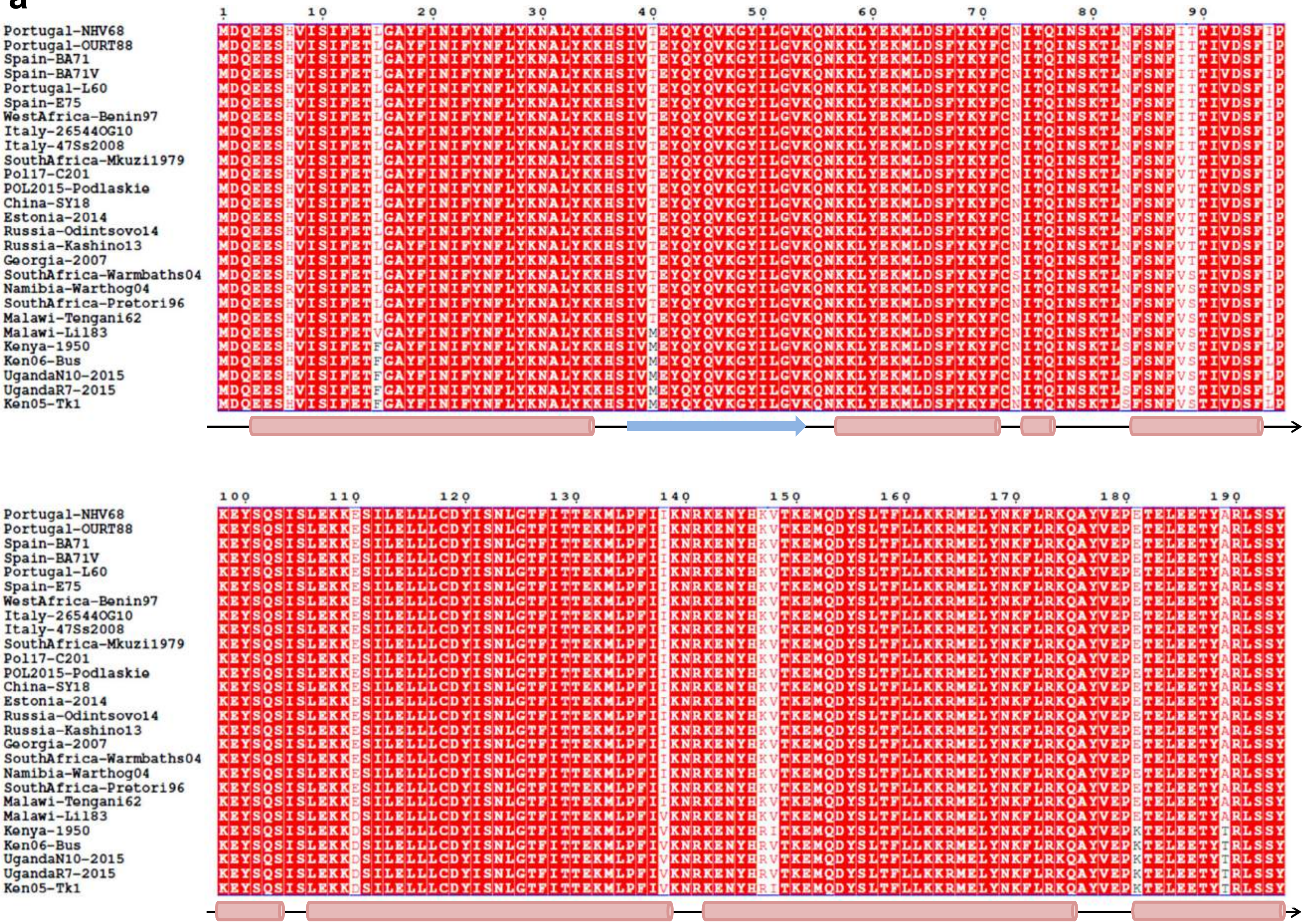
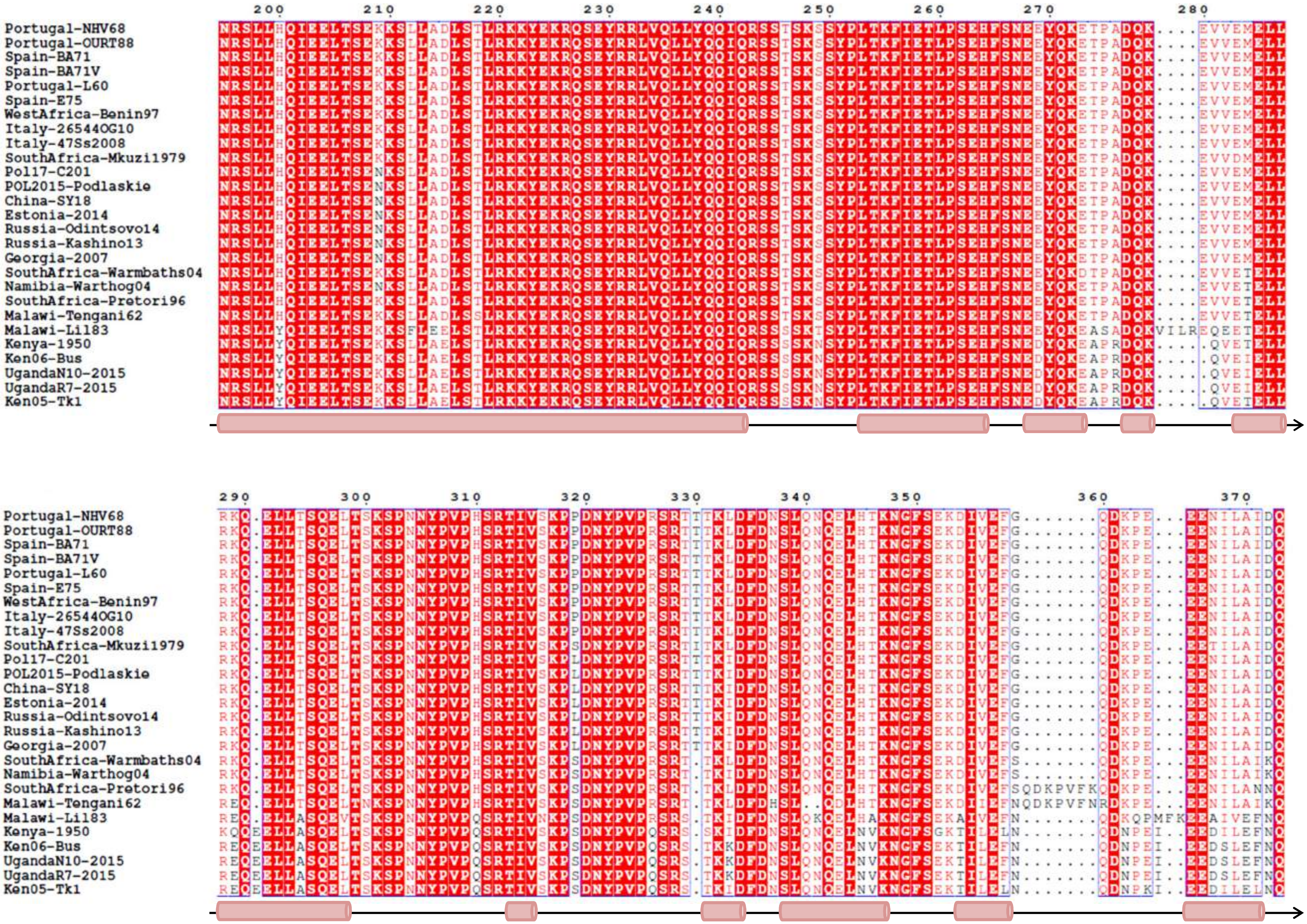
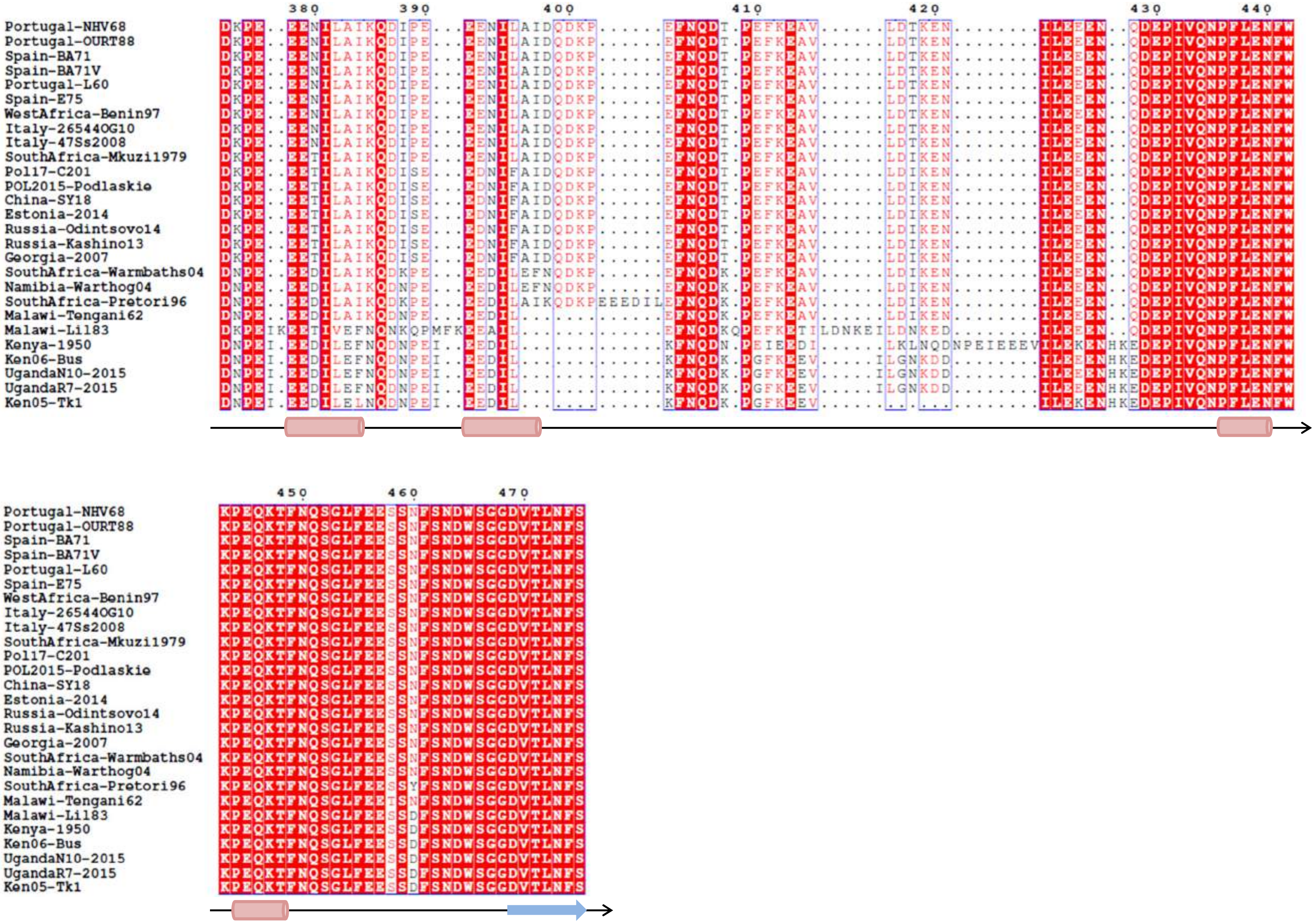

a
MGF505-4R
MGF300-4L
MGF505-9R
A151R
MGF360-16R
I196L
I215L
180
120
60
0
0
20k
40k
60k
80k
100k
120k
140k
160k
180k (bp)
b
20
15
10
5
0
Synonymous [-Log$_{10}$(q-value)]
0
5
10
15
20
Non-synonymous [-Log$_{10}$(q-value)]
P54
A240L
EP364R
I267L
CP312R
P11
I329L
505-6R
505-11L
I177L
300-1L
DP238L
H240R
K205R
360-15R
I215L
A151R
300-4L
505-4R
505-10R
505-9R
505-5R
360-8L
360-16R
I196L
c
MGF505 family domain
MGF505-4R
MGF360 family domain
MGF360-16R
I215L
Ubiquitin-conjugating catalytic domain
0
100
200
300
400
500

Fig. S3

a

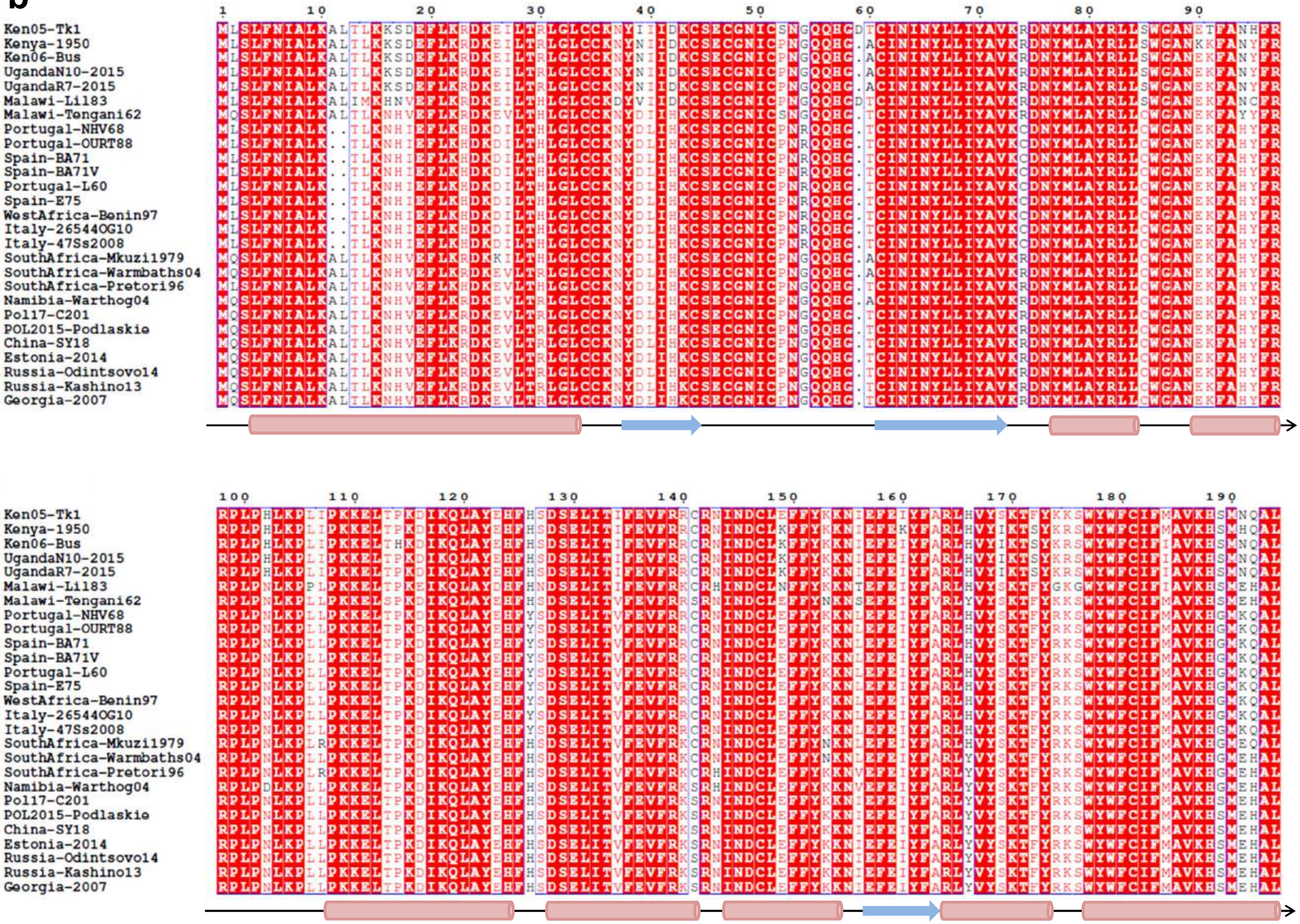

1 10 20 30 40 50 60 70 80 90
Portugal-NHV68 MDQEESHVISIFETLGAYFINIFYNFLYKNALYKKHSIVTEYQYQVKGYILGVKQNKKLYEKMLDSFYKYFCNITQINSKTLNFSNFITTIVDSFIP
Portugal-OURT88 MDQEESHVISIFETLGAYFINIFYNFLYKNALYKKHSIVTEYQYQVKGYILGVKQNKKLYEKMLDSFYKYFCNITQINSKTLNFSNFITTIVDSFIP
Spain-BA71 MDQEESHVISIFETLGAYFINIFYNFLYKNALYKKHSIVTEYQYQVKGYILGVKQNKKLYEKMLDSFYKYFCNITQINSKTLNFSNFITTIVDSFIP
Spain-BA71V MDQEESHVISIFETLGAYFINIFYNFLYKNALYKKHSIVTEYQYQVKGYILGVKQNKKLYEKMLDSFYKYFCNITQINSKTLNFSNFITTIVDSFIP
Portugal-L60 MDQEESHVISIFETLGAYFINIFYNFLYKNALYKKHSIVTEYQYQVKGYILGVKQNKKLYEKMLDSFYKYFCNITQINSKTLNFSNFITTIVDSFIP
Spain-E75 MDQEESHVISIFETLGAYFINIFYNFLYKNALYKKHSIVTEYQYQVKGYILGVKQNKKLYEKMLDSFYKYFCNITQINSKTLNFSNFITTIVDSFIP
WestAfrica-Benin97 MDQEESHVISIFETLGAYFINIFYNFLYKNALYKKHSIVTEYQYQVKGYILGVKQNKKLYEKMLDSFYKYFCNITQINSKTLNFSNFITTIVDSFIP
Italy-265440G10 MDQEESHVISIFETLGAYFINIFYNFLYKNALYKKHSIVTEYQYQVKGYILGVKQNKKLYEKMLDSFYKYFCNITQINSKTLNFSNFITTIVDSFIP
Italy-47Ss2008 MDQEESHVISIFETLGAYFINIFYNFLYKNALYKKHSIVTEYQYQVKGYILGVKQNKKLYEKMLDSFYKYFCNITQINSKTLNFSNFITTIVDSFIP
SouthAfrica-Mkuzi1979 MDQEESHVISIFETLGAYFINIFYNFLYKNALYKKHSIVTEYQYQVKGYILGVKQNKKLYEKMLDSFYKYFCNITQINSKTLNFSNFVTTIVDSFIP
Pol17-C201 MDQEESHVISIFETLGAYFINIFYNFLYKNALYKKHSIVTEYQYQVKGYILGVKQNKKLYEKMLDSFYKYFCNITQINSKTLNFSNFVTTIVDSFIP
POL2015-Podlaskie MDQEESHVISIFETLGAYFINIFYNFLYKNALYKKHSIVTEYQYQVKGYILGVKQNKKLYEKMLDSFYKYFCNITQINSKTLNFSNFVTTIVDSFIP
China-SY18 MDQEESHVISIFETLGAYFINIFYNFLYKNALYKKHSIVTEYQYQVKGYILGVKQNKKLYEKMLDSFYKYFCNITQINSKTLNFSNFVTTIVDSFIP
Estonia-2014 MDQEESHVISIFETLGAYFINIFYNFLYKNALYKKHSIVTEYQYQVKGYILGVKQNKKLYEKMLDSFYKYFCNITQINSKTLNFSNFVTTIVDSFIP
Russia-Odintsovo14 MDQEESHVISIFETLGAYFINIFYNFLYKNALYKKHSIVTEYQYQVKGYILGVKQNKKLYEKMLDSFYKYFCNITQINSKTLNFSNFVTTIVDSFIP
Russia-Kashino13 MDQEESHVISIFETLGAYFINIFYNFLYKNALYKKHSIVTEYQYQVKGYILGVKQNKKLYEKMLDSFYKYFCNITQINSKTLNFSNFVTTIVDSFIP
Georgia-2007 MDQEESHVISIFETLGAYFINIFYNFLYKNALYKKHSIVTEYQYQVKGYILGVKQNKKLYEKMLDSFYKYFCNITQINSKTLNFSNFVTTIVDSFIP
SouthAfrica-Warmbaths04 MDQEESHVISIFETLGAYFINIFYNFLYKNALYKKHSIVTEYQYQVKGYILGVKQNKKLYEKMLDSFYKYFCSITQINSKTLNFSNFVSTIVDSFIP
Namibia-Warthog04 MDQEESRVISIFETLGAYFINIFYNFLYKNALYKKHSIVTEYQYQVKGYILGVKQNKKLYEKMLDSFYKYFCNITQINSKTLNFSNFVSTIVDSFIP
SouthAfrica-Pretori96 MDQEESHVISIFETLGAYFINIFYNFLYKNALYKKHSIVTEYQYQVKGYILGVKQNKKLYEKMLDSFYKYFCNITQINSKTLNFSNFVSTIVDSFIP
Malawi-Tengani62 MDQEESHVISIFETLGAYFINIFYNFLYKNALYKKHSIVTEYQYQVKGYILGVKQNKKLYEKMLDSFYKYFCNITQINSKTLNFSNFVSTIVDSFIP
Malawi-Lil83 MDQEESHVISIFETVGAYFINIFYNFLYKNALYKKHSIVMEYQYQVKGYILGVKQNKKLYEKMLDSFYKYFCNITQINSKTLNFSNFVSTIVDSFLP
Kenya-1950 MDQEESHVISIFETFGAYFINIFYNFLYKNALYKKHSIVMEYQYQVKGYILGVKQNKKLYEKMLDSFYKYFCNITQINSKTLSFSNFVSTIVDSFLP
Ken06-Bus MDQEESHVISIFETFGAYFINIFYNFLYKNALYKKHSIVMEYQYQVKGYILGVKQNKKLYEKMLDSFYKYFCNITQINSKTLSFSNFVSTIVDSFLP
UgandaN10-2015 MDQEESHVISIFETFGAYFINIFYNFLYKNALYKKHSIVMEYQYQVKGYILGVKQNKKLYEKMLDSFYKYFCNITQINSKTLSFSNFVSTIVDSFLP
UgandaR7-2015 MDQEESHVISIFETFGAYFINIFYNFLYKNALYKKHSIVMEYQYQVKGYILGVKQNKKLYEKMLDSFYKYFCNITQINSKTLSFSNFVSTIVDSFLP
Ken05-Tk1 MDQEESHVISIFETFGAYFINIFYNFLYKNALYKKHSIVMEYQYQVKGYILGVKQNKKLYEKMLDSFYKYFCNITQINSKTLSFSNFVSTIVDSFLP
100 110 120 130 140 150 160 170 180 190
Portugal-NHV68 KEYSQSISLEKKESILELLLCDYISNLGTFITTEKMLPFIIKNRKENYHKVTKEMQDYSLTFLLKKRMELYNKFLRKQAYVEPETELEETYARLSSY
Portugal-OURT88 KEYSQSISLEKKESILELLLCDYISNLGTFITTEKMLPFIIKNRKENYHKVTKEMQDYSLTFLLKKRMELYNKFLRKQAYVEPETELEETYARLSSY
Spain-BA71 KEYSQSISLEKKESILELLLCDYISNLGTFITTEKMLPFIIKNRKENYHKVTKEMQDYSLTFLLKKRMELYNKFLRKQAYVEPETELEETYARLSSY
Spain-BA71V KEYSQSISLEKKESILELLLCDYISNLGTFITTEKMLPFIIKNRKENYHKVTKEMQDYSLTFLLKKRMELYNKFLRKQAYVEPETELEETYARLSSY
Portugal-L60 KEYSQSISLEKKESILELLLCDYISNLGTFITTEKMLPFIIKNRKENYHKVTKEMQDYSLTFLLKKRMELYNKFLRKQAYVEPETELEETYARLSSY
Spain-E75 KEYSQSISLEKKESILELLLCDYISNLGTFITTEKMLPFIIKNRKENYHKVTKEMQDYSLTFLLKKRMELYNKFLRKQAYVEPETELEETYARLSSY
WestAfrica-Benin97 KEYSQSISLEKKESILELLLCDYISNLGTFITTEKMLPFIIKNRKENYHKVTKEMQDYSLTFLLKKRMELYNKFLRKQAYVEPETELEETYARLSSY
Italy-265440G10 KEYSQSISLEKKESILELLLCDYISNLGTFITTEKMLPFIIKNRKENYHKVTKEMQDYSLTFLLKKRMELYNKFLRKQAYVEPETELEETYARLSSY
Italy-47Ss2008 KEYSQSISLEKKESILELLLCDYISNLGTFITTEKMLPFIIKNRKENYHKVTKEMQDYSLTFLLKKRMELYNKFLRKQAYVEPETELEETYARLSSY
SouthAfrica-Mkuzi1979 KEYSQSISLEKKESILELLLCDYISNLGTFITTEKMLPFIIKNRKENYHKVTKEMQDYSLTFLLKKRMELYNKFLRKQAYVEPETELEETYARLSSY
Pol17-C201 KEYSQSISLEKKESILELLLCDYISNLGTFITTEKMLPFIIKNRKENYHKVTKEMQDYSLTFLLKKRMELYNKFLRKQAYVEPETELEETYARLSSY
POL2015-Podlaskie KEYSQSISLEKKESILELLLCDYISNLGTFITTEKMLPFIIKNRKENYHKVTKEMQDYSLTFLLKKRMELYNKFLRKQAYVEPETELEETYARLSSY
China-SY18 KEYSQSISLEKKESILELLLCDYISNLGTFITTEKMLPFIIKNRKENYHKVTKEMQDYSLTFLLKKRMELYNKFLRKQAYVEPETELEETYARLSSY
Estonia-2014 KEYSQSISLEKKESILELLLCDYISNLGTFITTEKMLPFIIKNRKENYHKVTKEMQDYSLTFLLKKRMELYNKFLRKQAYVEPETELEETYARLSSY
Russia-Odintsovo14 KEYSQSISLEKKESILELLLCDYISNLGTFITTEKMLPFIIKNRKENYHKVTKEMQDYSLTFLLKKRMELYNKFLRKQAYVEPETELEETYARLSSY
Russia-Kashino13 KEYSQSISLEKKESILELLLCDYISNLGTFITTEKMLPFIIKNRKENYHKVTKEMQDYSLTFLLKKRMELYNKFLRKQAYVEPETELEETYARLSSY
Georgia-2007 KEYSQSISLEKKESILELLLCDYISNLGTFITTEKMLPFIIKNRKENYHKVTKEMQDYSLTFLLKKRMELYNKFLRKQAYVEPETELEETYARLSSY
SouthAfrica-Warmbaths04 KEYSQSISLEKKESILELLLCDYISNLGTFITTEKMLPFIIKNRKENYHKVTKEMQDYSLTFLLKKRMELYNKFLRKQAYVEPETELEETYARLSSY
Namibia-Warthog04 KEYSQSISLEKKESILELLLCDYISNLGTFITTEKMLPFIIKNRKENYHKVTKEMQDYSLTFLLKKRMELYNKFLRKQAYVEPETELEETYARLSSY
SouthAfrica-Pretori96 KEYSQSISLEKKESILELLLCDYISNLGTFITTEKMLPFIIKNRKENYHKVTKEMQDYSLTFLLKKRMELYNKFLRKQAYVEPETELEETYARLSSY
Malawi-Tengani62 KEYSQSISLEKKESILELLLCDYISNLGTFITTEKMLPFIIKNRKENYHKVTKEMQDYSLTFLLKKRMELYNKFLRKQAYVEPETELEETYARLSSY
Malawi-Lil83 KEYSQSISLEKKDSILELLLCDYISNLGTFITTEKMLPFIVKNRKENYHKVTKEMQDYSLTFLLKKRMELYNKFLRKQAYVEPETELEETYARLSSY
Kenya-1950 KEYSQSISLEKKDSILELLLCDYISNLGTFITTEKMLPFIVKNRKENYHRITKEMQDYSLTFLLKKRMELYNKFLRKQAYVEPKTELEETYTRLSSY
Ken06-Bus KEYSQSISLEKKDSILELLLCDYISNLGTFITTEKMLPFIVKNRKENYHRVTKEMQDYSLTFLLKKRMELYNKFLRKQAYVEPKTELEETYTRLSSY
UgandaN10-2015 KEYSQSISLEKKDSILELLLCDYISNLGTFITTEKMLPFIVKNRKENYHRVTKEMQDYSLTFLLKKRMELYNKFLRKQAYVEPKTELEETYTRLSSY
UgandaR7-2015 KEYSQSISLEKKDSILELLLCDYISNLGTFITTEKMLPFIVKNRKENYHRVTKEMQDYSLTFLLKKRMELYNKFLRKQAYVEPKTELEETYTRLSSY
Ken05-Tk1 KEYSQSISLEKKDSILELLLCDYISNLGTFITTEKMLPFIVKNRKENYHRITKEMQDYSLTFLLKKRMELYNKFLRKQAYVEPKTELEETYTRLSSY

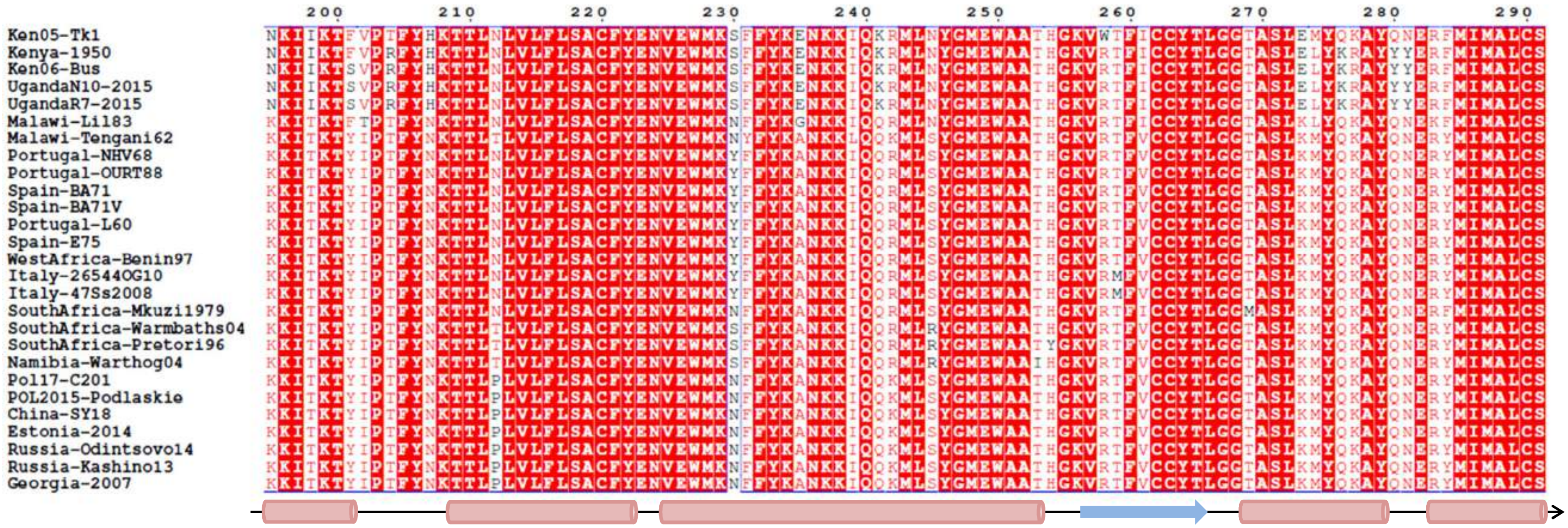

200
210
220
230
240
250
260
270
280
290
300
310
320
330
340
350
360
370
Portugal-NHV68
Portugal-OURT88
Spain-BA71
Spain-BA71V
Portugal-L60
Spain-E75
WestAfrica-Benin97
Italy-265440G10
Italy-47Ss2008
SouthAfrica-Mkuzi1979
Pol17-C201
POL2015-Podlaskie
China-SY18
Estonia-2014
Russia-Odintsovo14
Russia-Kashino13
Georgia-2007
SouthAfrica-Warmbaths04
Namibia-Warthog04
SouthAfrica-Pretori96
Malawi-Tengani62
Malawi-Lil83
Kenya-1950
Ken06-Bus
UgandaN10-2015
UgandaR7-2015
Ken05-Tk1



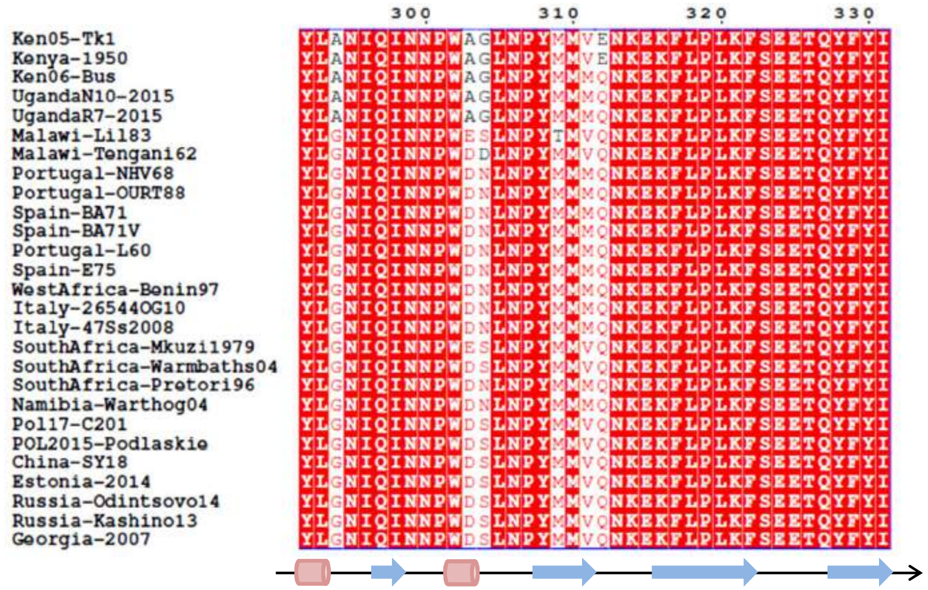
```
Portugal-NHV68              DKPE..EENILAIKQDIPE...EENILAIDQDKP......EFNQDT.PEFKEAV......LDTKEN........ILEEEN..QDEPIVQNPFLENFW
Portugal-OURT88             DKPE..EENILAIKQDIPE...EENILAIDQDKP......EFNQDT.PEFKEAV......LDTKEN........ILEEEN..QDEPIVQNPFLENFW
Spain-BA71                  DKPE..EENILAIKQDIPE...EENILAIDQDKP......EFNQDT.PEFKEAV......LDTKEN........ILEEEN..QDEPIVQNPFLENFW
Spain-BA71V                 DKPE..EENILAIKQDIPE...EENILAIDQDKP......EFNQDT.PEFKEAV......LDTKEN........ILEEEN..QDEPIVQNPFLENFW
Portugal-L60                DKPE..EENILAIKQDIPE...EENILAIDQDKP......EFNQDT.PEFKEAV......LDTKEN........ILEEEN..QDEPIVQNPFLENFW
Spain-E75                   DKPE..EENILAIKQDIPE...EENILAIDQDKP......EFNQDT.PEFKEAV......LDTKEN........ILEEEN..QDEPIVQNPFLENFW
WestAfrica-Benin97          DKPE..EENILAIKQDIPE...EENILAIDQDKP......EFNQDT.PEFKEAV......LDTKEN........ILEEEN..QDEPIVQNPFLENFW
Italy-265440G10             DKPE..EENILAIKQDIPE...EENILAIDQDKP......EFNQDT.PEFKEAV......LDTKEN........ILEEEN..QDEPIVQNPFLENFW
Italy-47Ss2008              DKPE..EENILAIKQDIPE...EENILAIDQDKP......EFNQDT.PEFKEAV......LDTKEN........ILEEEN..QDEPIVQNPFLENFW
SouthAfrica-Mkuzi1979       DKPE..EETILAIKQDIPE...EENILAIDQDKP......EFNQDT.PEFKEAV......LDIKEN........ILEEEN..QDEPIVQNPFLENFW
Pol17-C201                  DKPE..EETILAIKQDISE...EDNIFAIDQDKP......EFNQDT.PEFKEAV......LDIKEN........ILEEEN..QDEPIVQNPFLENFW
POL2015-Podlaskie           DKPE..EETILAIKQDISE...EDNIFAIDQDKP......EFNQDT.PEFKEAV......LDIKEN........ILEEEN..QDEPIVQNPFLENFW
China-SY18                  DKPE..EETILAIKQDISE...EDNIFAIDQDKP......EFNQDT.PEFKEAV......LDIKEN........ILEEEN..QDEPIVQNPFLENFW
Estonia-2014                DKPE..EETILAIKQDISE...EDNIFAIDQDKP......EFNQDT.PEFKEAV......LDIKEN........ILEEEN..QDEPIVQNPFLENFW
Russia-Odintsovo14          DKPE..EETILAIKQDISE...EDNIFAIDQDKP......EFNQDT.PEFKEAV......LDIKEN........ILEEEN..QDEPIVQNPFLENFW
Russia-Kashino13            DKPE..EETILAIKQDISE...EDNIFAIDQDKP......EFNQDT.PEFKEAV......LDIKEN........ILEEEN..QDEPIVQNPFLENFW
Georgia-2007                DKPE..EETILAIKQDISE...EDNIFAIDQDKP......EFNQDT.PEFKEAV......LDIKEN........ILEEEN..QDEPIVQNPFLENFW
SouthAfrica-Warmbaths04     DNPE..EEDILAIKQDKPE...EEDILEFNQDKP......EFNQDK.PEFKEAV......LDIKEN........ILEEEN..QDEPIVQNPFLENFW
Namibia-Warthog04           DNPE..EEDILAIKQDNPE...EEDILEFNQDKP......EFNQDK.PEFKEAV......LDIKEN........ILEEEN..QDEPIVQNPFLENFW
SouthAfrica-Pretori96       DNPE..EEDILAIKQDKPE...EEDILAIKQDKPEEEDILEFNQDK.PEFKEAV......LDIKEN........ILEEEN..QDEPIVQNPFLENFW
Malawi-Tengani62            DNPE..EEDILAIKQDNPE...EEDIL.............EFNQDK.PEFKEAV......LDIKEN........ILEEEN..QDEPIVQNPFLENFW
Malawi-Lil83                DKPEIKEETIVEFNQNKQPMFKEEAIL.............EFNQDKQPEFKETILDNKEILDNKED........ILEEEN..QDEPIVQNPFLENFW
Kenya-1950                  DNPEI.EEDILEFNQDNPEI..EEDIL.............KFNQDN.PEIEEDI......LKLNQDNPEIEEEVILEKENHKEDEPIVQNPFLENFW
Ken06-Bus                   DNPEI.EEDILEFNQDNPEI..EEDIL.............KFNQDK.PGFKEEV.....ILGNKDD........ILEEENHKEDEPIVQNPFLENFW
UgandaN10-2015              DNPEI.EEDILEFNQDNPEI..EEDIL.............KFNQDK.PGFKEEV.....ILGNKDD........ILEEENHKEDEPIVQNPFLENFW
UgandaR7-2015               DNPEI.EEDILEFNQDNPEI..EEDIL.............KFNQDK.PGFKEEV.....ILGNKDD........ILEEENHKEDEPIVQNPFLENFW
Ken05-Tk1                   DNPEI.EEDILELNQDNPEI..EEDIL.............KFNQDK.PGFKEEV....................ILEKENHKEDEPIVQNPFLENFW

Portugal-NHV68              KPEQKTFNQSGLFEESSNFSNDWSGGDVTLNFS
Portugal-OURT88             KPEQKTFNQSGLFEESSNFSNDWSGGDVTLNFS
Spain-BA71                  KPEQKTFNQSGLFEESSNFSNDWSGGDVTLNFS
Spain-BA71V                 KPEQKTFNQSGLFEESSNFSNDWSGGDVTLNFS
Portugal-L60                KPEQKTFNQSGLFEESSNFSNDWSGGDVTLNFS
Spain-E75                   KPEQKTFNQSGLFEESSNFSNDWSGGDVTLNFS
WestAfrica-Benin97          KPEQKTFNQSGLFEESSNFSNDWSGGDVTLNFS
Italy-265440G10             KPEQKTFNQSGLFEESSNFSNDWSGGDVTLNFS
Italy-47Ss2008              KPEQKTFNQSGLFEESSNFSNDWSGGDVTLNFS
SouthAfrica-Mkuzi1979       KPEQKTFNQSGLFEESSNFSNDWSGGDVTLNFS
Pol17-C201                  KPEQKTFNQSGLFEESSNFSNDWSGGDVTLNFS
POL2015-Podlaskie           KPEQKTFNQSGLFEESSNFSNDWSGGDVTLNFS
China-SY18                  KPEQKTFNQSGLFEESSNFSNDWSGGDVTLNFS
Estonia-2014                KPEQKTFNQSGLFEESSNFSNDWSGGDVTLNFS
Russia-Odintsovo14          KPEQKTFNQSGLFEESSNFSNDWSGGDVTLNFS
Russia-Kashino13            KPEQKTFNQSGLFEESSNFSNDWSGGDVTLNFS
Georgia-2007                KPEQKTFNQSGLFEESSNFSNDWSGGDVTLNFS
SouthAfrica-Warmbaths04     KPEQKTFNQSGLFEESSNFSNDWSGGDVTLNFS
Namibia-Warthog04           KPEQKTFNQSGLFEESSNFSNDWSGGDVTLNFS
SouthAfrica-Pretori96       KPEQKTFNQSGLFEESSYFSNDWSGGDVTLNFS
Malawi-Tengani62            KPEQKTFNQSGLFEETSNFSNDWSGGDVTLNFS
Malawi-Lil83                KPEQKTFNQSGLFEESSDFSNDWSGGDVTLNFS
Kenya-1950                  KPEQKTFNQSGLFEESSDFSNDWSGGDVTLNFS
Ken06-Bus                   KPEQKTFNQSGLFEESSDFSNDWSGGDVTLNFS
UgandaN10-2015              KPEQKTFNQSGLFEESSDFSNDWSGGDVTLNFS
UgandaR7-2015               KPEQKTFNQSGLFEESSDFSNDWSGGDVTLNFS
Ken05-Tk1                   KPEQKTFNQSGLFEESSDFSNDWSGGDVTLNFS
```

b


1
10
20
30
40
50
60
70
80
90
100
110
120
130
140
150
160
170
180
190
Ken05-Tk1
Kenya-1950
Ken06-Bus
UgandaN10-2015
UgandaR7-2015
Malawi-Lil83
Malawi-Tengani62
Portugal-NHV68
Portugal-OURT88
Spain-BA71
Spain-BA71V
Portugal-L60
Spain-E75
WestAfrica-Benin97
Italy-26544OG10
Italy-47Ss2008
SouthAfrica-Mkuzi1979
SouthAfrica-Warmbaths04
SouthAfrica-Pretori96
Namibia-Warthog04
Pol17-C201
POL2015-Podlaskie
China-SY18
Estonia-2014
Russia-Odintsovo14
Russia-Kashino13
Georgia-2007



```
                                200       210       220       230       240       250       260       270       280       290
Ken05-Tk1                     NKIIKTFVPTFYHKTTLNLVLFLSACFYENVEWMKSFFYKENKKIQKRMLNYGMEWAATHGKVWTFICCYTLGGTASLEMYQKAYQNERFMIMALCS
Kenya-1950                    NKIIKTFVPRFYHKTTLNLVLFLSACFYENVEWMKSFFYKENKKIQKRMLNYGMEWAATHGKVRTFICCYTLGGTASLELYKRAYYYERFMIMALCS
Ken06-Bus                     NKIIKTSVPRFYHKTTLNLVLFLSACFYENVEWMKSFFYKENKKIQKRMLNYGMEWAATHGKVRTFICCYTLGGTASLELYKRAYYYERFMIMALCS
UgandaN10-2015                NKIIKTSVPRFYHKTTLNLVLFLSACFYENVEWMKSFFYKENKKIQKRMLNYGMEWAATHGKVRTFICCYTLGGTASLELYKRAYYYERFMIMALCS
UgandaR7-2015                 NKIIKTSVPRFYHKTTLNLVLFLSACFYENVEWMKSFFYKENKKIQKRMLNYGMEWAATHGKVRTFICCYTLGGTASLELYKRAYYYERFMIMALCS
Malawi-Lil83                  KKITKTFTPTFYNKTTLNLVLFLSACFYENVEWMKNFFYKGNKKIQQRMLNYGMEWAATHGKVRTFICCYTLGGTASLKLYQKAYQNEKFMIMALCS
Malawi-Tengani62              KKITKTYIPTFYNKTTLTLVLFLSACFYENVEWMKNYFYKANKKLQQKMLSYGMEWAATHGKVRTFVCCYTLGGTASLKMYQKAYQNERYMIMALCS
Portugal-NHV68                KKITKTYIPTFYNKTTLNLVLFLSACFYENVEWMKYFFYKANKKIQQRMLSYGMEWAATHGKVRTFVCCYTLGGTASLKMYQKAYQNERYMIMALCS
Portugal-OURT88               KKITKTYIPTFYNKTTLNLVLFLSACFYENVEWMKYFFYKANKKIQQRMLSYGMEWAATHGKVRTFVCCYTLGGTASLKMYQKAYQNERYMIMALCS
Spain-BA71                    KKITKTYIPTFYNKTTLNLVLFLSACFYENVEWMKYFFYKANKKIQQRMLSYGMEWAATHGKVRTFVCCYTLGGTASLKMYQKAYQNERYMIMALCS
Spain-BA71V                   KKITKTYIPTFYNKTTLNLVLFLSACFYENVEWMKYFFYKANKKIQQRMLSYGMEWAATHGKVRTFVCCYTLGGTASLKMYQKAYQNERYMIMALCS
Portugal-L60                  KKITKTYIPTFYNKTTLNLVLFLSACFYENVEWMKYFFYKANKKIQQRMLSYGMEWAATHGKVRTFVCCYTLGGTASLKMYQKAYQNERYMIMALCS
Spain-E75                     KKITKTYIPTFYNKTTLNLVLFLSACFYENVEWMKYFFYKANKKIQQRMLSYGMEWAATHGKVRTFVCCYTLGGTASLKMYQKAYQNERYMIMALCS
WestAfrica-Benin97            KKITKTYIPTFYNKTTLNLVLFLSACFYENVEWMKYFFYKANKKIQQRMLSYGMEWAATHGKVRTFVCCYTLGGTASLKMYQKAYQNERYMIMALCS
Italy-26544OG10               KKITKTYIPTFYNKTTLNLVLFLSACFYENVEWMKYFFYKANKKIQQRMLSYGMEWAATHGKVRMFVCCYTLGGTASLKMYQKAYQNERYMIMALCS
Italy-47Ss2008                KKITKTYIPTFYNKTTLNLVLFLSACFYENVEWMKYFFYKANKKIQQRMLSYGMEWAATHGKVRMFVCCYTLGGTASLKMYQKAYQNERYMIMALCS
SouthAfrica-Mkuzi1979         KKITKTYIPTFYNKTTLNLVLFLSACFYENVEWMKNFFYKANKKIQQRMLSYGMEWAATHGKVRTFICCYTLGGMASLKMYQKAYQNERFMIMALCS
SouthAfrica-Warmbaths04       KKITKTYIPTFYNKTTLTLVLFLSACFYENVEWMKSFFYKANKKIQQRMLRYGMEWAATHGKVRTFVCCYTLGGTASLKMYQKAYQNERYMIMALCS
SouthAfrica-Pretori96         KKITKTYIPTFYNKTTLTLVLFLSACFYENVEWMKSFFYKANKKIQQRMLRYGMEWAATYGKVRTFVCCYTLGGTASLKMYQKAYQNERYMIMALCS
Namibia-Warthog04             KKITKTYIPTFYNKTTLTLVLFLSACFYENVEWMKSFFYKANKKIQQRMLRYGMEWAAIHGKVRTFVCCYTLGGTASLKMYQKAYQNERYMIMALCS
Pol17-C201                    KKITKTYIPTFYNKTTLPLVLFLSACFYENVEWMKNFFYKANKKIQQKMLSYGMEWAATHGKVRTFVCCYTLGGTASLKMYQKAYQNERYMIMALCS
POL2015-Podlaskie             KKITKTYIPTFYNKTTLPLVLFLSACFYENVEWMKNFFYKANKKIQQKMLSYGMEWAATHGKVRTFVCCYTLGGTASLKMYQKAYQNERYMIMALCS
China-SY18                    KKITKTYIPTFYNKTTLPLVLFLSACFYENVEWMKNFFYKANKKIQQKMLSYGMEWAATHGKVRTFVCCYTLGGTASLKMYQKAYQNERYMIMALCS
Estonia-2014                  KKITKTYIPTFYNKTTLPLVLFLSACFYENVEWMKNFFYKANKKIQQKMLSYGMEWAATHGKVRTFVCCYTLGGTASLKMYQKAYQNERYMIMALCS
Russia-Odintsovo14            KKITKTYIPTFYNKTTLPLVLFLSACFYENVEWMKNFFYKANKKIQQKMLSYGMEWAATHGKVRTFVCCYTLGGTASLKMYQKAYQNERYMIMALCS
Russia-Kashino13              KKITKTYIPTFYNKTTLPLVLFLSACFYENVEWMKNFFYKANKKIQQKMLSYGMEWAATHGKVRTFVCCYTLGGTASLKMYQKAYQNERYMIMALCS
Georgia-2007                  KKITKTYIPTFYNKTTLPLVLFLSACFYENVEWMKNFFYKANKKIQQKMLSYGMEWAATHGKVRTFVCCYTLGGTASLKMYQKAYQNERYMIMALCS
```



```
                                   300       310       320       330
Ken05-Tk1                     YLANIQINNPWAGLNPYMMVENKEKFLPLKFSEETQYFYI
Kenya-1950                    YLANIQINNPWAGLNPYMMVENKEKFLPLKFSEETQYFYI
Ken06-Bus                     YLANIQINNPWAGLNPYMMMQNKEKFLPLKFSEETQYFYI
UgandaN10-2015                YLANIQINNPWAGLNPYMMMQNKEKFLPLKFSEETQYFYI
UgandaR7-2015                 YLANIQINNPWAGLNPYMMMQNKEKFLPLKFSEETQYFYI
Malawi-Lil83                  YLGNIQINNPWESLNPYTMVQNKEKFLPLKFSEETQYFYI
Malawi-Tengani62              YLGNIQINNPWDDLNPYMMVQNKEKFLPLKFSEETQYFYI
Portugal-NHV68                YLGNIQINNPWDNLNPYMMMQNKEKFLPLKFSEETQYFYI
Portugal-OURT88               YLGNIQINNPWDNLNPYMMMQNKEKFLPLKFSEETQYFYI
Spain-BA71                    YLGNIQINNPWDNLNPYMMMQNKEKFLPLKFSEETQYFYI
Spain-BA71V                   YLGNIQINNPWDNLNPYMMMQNKEKFLPLKFSEETQYFYI
Portugal-L60                  YLGNIQINNPWDNLNPYMMMQNKEKFLPLKFSEETQYFYI
Spain-E75                     YLGNIQINNPWDNLNPYMMMQNKEKFLPLKFSEETQYFYI
WestAfrica-Benin97            YLGNIQINNPWDNLNPYMMMQNKEKFLPLKFSEETQYFYI
Italy-26544OG10               YLGNIQINNPWDNLNPYMMMQNKEKFLPLKFSEETQYFYI
Italy-47Ss2008                YLGNIQINNPWDNLNPYMMMQNKEKFLPLKFSEETQYFYI
SouthAfrica-Mkuzi1979         YLGNIQINNPWESLNPYMMVQNKEKFLPLKFSEETQYFYI
SouthAfrica-Warmbaths04       YLGNIQINNPWDSLNPYMMVQNKEKFLPLKFSEETQYFYI
SouthAfrica-Pretori96         YLGNIQINNPWDNLNPYMMMQNKEKFLPLKFSEETQYFYI
Namibia-Warthog04             YLGNIQINNPWDNLNPYMMMQNKEKFLPLKFSEETQYFYI
Pol17-C201                    YLGNIQINNPWDSLNPYMMVQNKEKFLPLKFSEETQYFYI
POL2015-Podlaskie             YLGNIQINNPWDSLNPYMMVQNKEKFLPLKFSEETQYFYI
China-SY18                    YLGNIQINNPWDSLNPYMMVQNKEKFLPLKFSEETQYFYI
Estonia-2014                  YLGNIQINNPWDSLNPYMMVQNKEKFLPLKFSEETQYFYI
Russia-Odintsovo14            YLGNIQINNPWDSLNPYMMVQNKEKFLPLKFSEETQYFYI
Russia-Kashino13              YLGNIQINNPWDSLNPYMMVQNKEKFLPLKFSEETQYFYI
Georgia-2007                  YLGNIQINNPWDSLNPYMMVQNKEKFLPLKFSEETQYFYI
```

Fig. S4

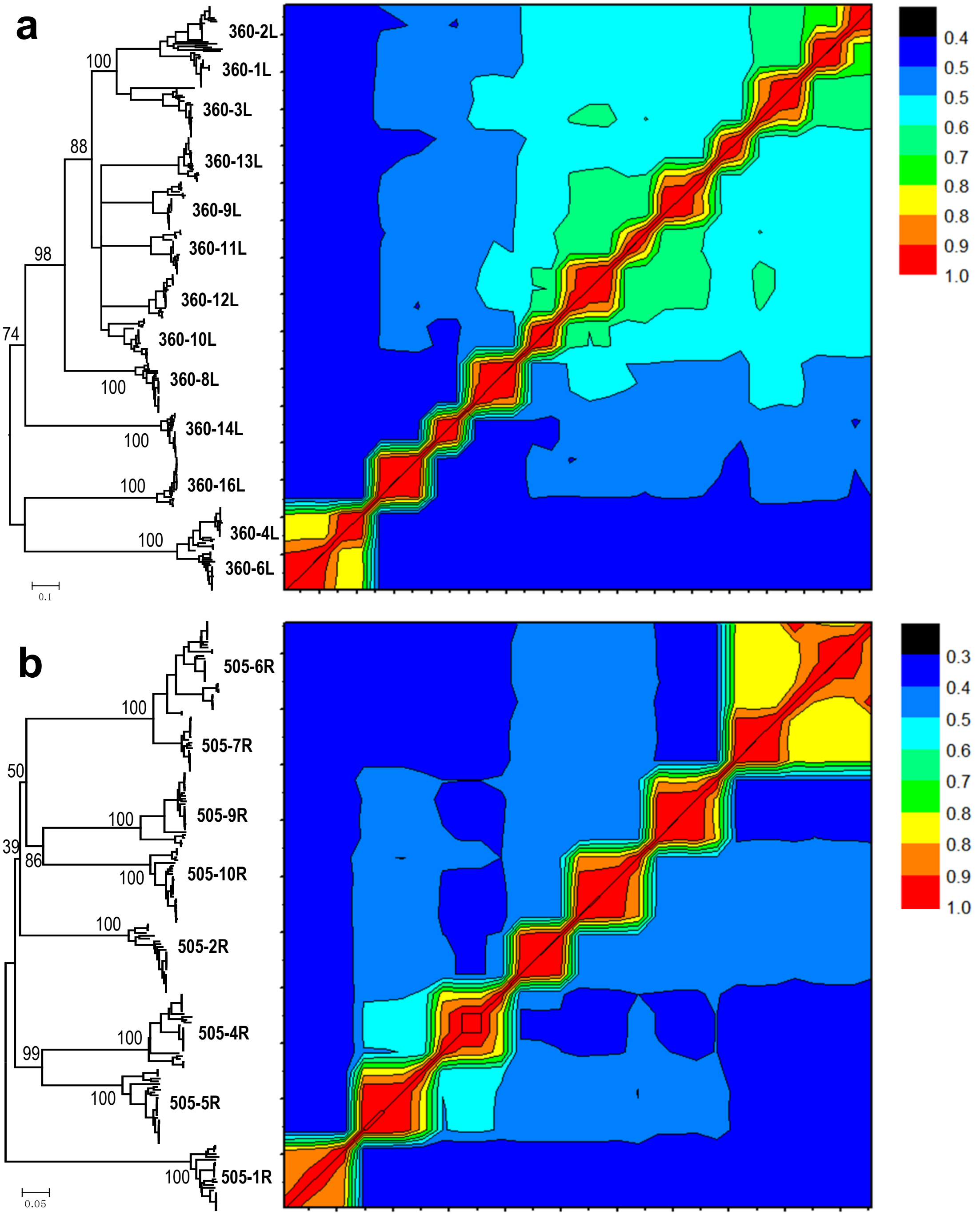
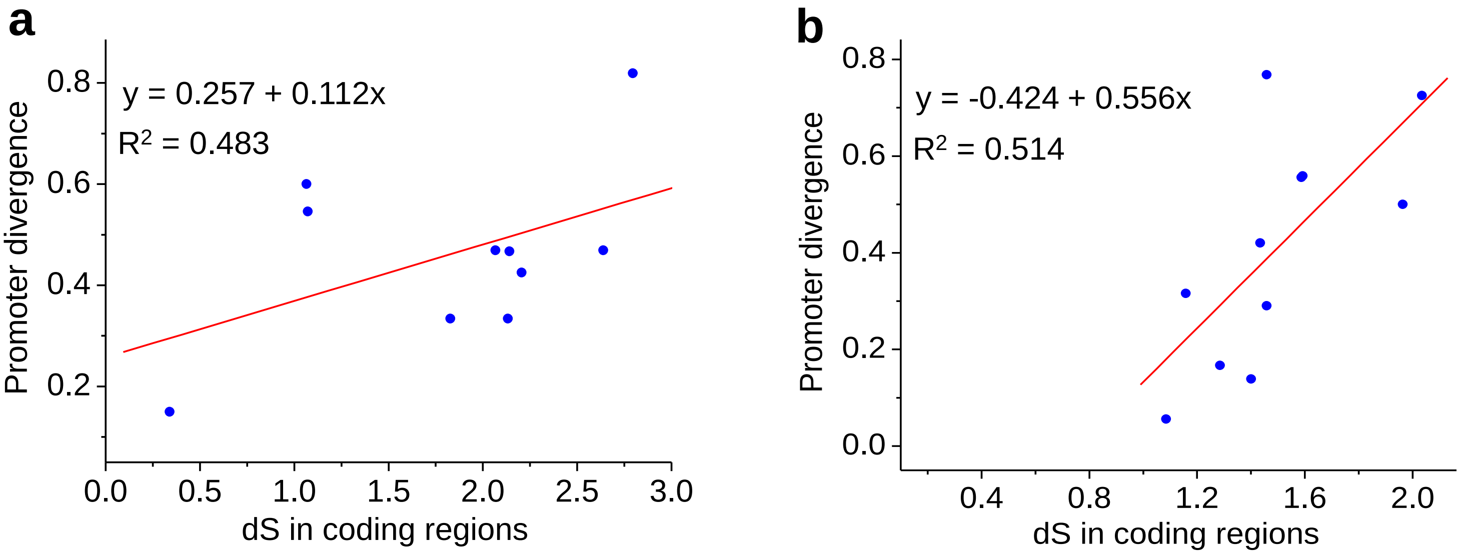

a
360-2L
100
360-1L
360-3L
88
360-13L
360-9L
98
360-11L
360-12L
74
360-10L
100
360-8L
360-14L
100
100
360-16L
100
360-4L
360-6L
0.1
0.4
0.5
0.5
0.6
0.7
0.8
0.8
0.9
1.0
b
505-6R
100
505-7R
50
100
505-9R
39
86
100
505-10R
100
505-2R
100
505-4R
99
100
505-5R
100
505-1R
0.05
0.3
0.4
0.5
0.6
0.7
0.8
0.8
0.9
1.0

Fig. S5


a
y = 0.257 + 0.112x
R2 = 0.483
Promoter divergence
0.2
0.4
0.6
0.8
0.0
0.5
1.0
1.5
2.0
2.5
3.0
dS in coding regions
b
y = -0.424 + 0.556x
R2 = 0.514
Promoter divergence
0.0
0.2
0.4
0.6
0.8
0.4
0.8
1.2
1.6
2.0
dS in coding regions