MHC Class I-like Gene in Squirrel Monkey Cytomegalovirus and the Evolution of the Virus

Eiichi Hondo[1], Tetsuya Mizutani[2], Hiroshi Shimoda[3], Atsuo Iida[1]

[1] Laboratory of Animal Morphology, Department of Animal Science, Graduate School of Bioagricultural Sciences, Nagoya University, 464-8603 Nagoya, Japan

[2]Center for Infectious Disease Epidemiology and Prevention Research, Tokyo University of Agriculture and Technology, Tokyo 183-8538, Japan

[3] Laboratory of Microbiology, Joint Faculty of Veterinary Medicine, Yamaguchi University, Yamaguchi 1677-1, Japan

Corresponce: Eiichi Hondo, DVM, PhD
e-mail: ehondo@agr.nagoya-u.ac.jp
Tel&Fax: +81-52-789-4185

Abstract

Homology searches between viral and human peptides revealed a squirrel monkey cytomegalovirus (Saimiriine herpesvirus 4 strain [SBHV4]) S9 protein sharing 60.6% amino acid identity across 86.3% of full-length human HLA-A. This represents the most extensive and highest homology observed to date between a viral protein and an MHC class I molecule.

Phylogenetic analysis of MHC class I-like sequences from primate HLA-A orthologs and SBHV4 suggests that SBHV4-S9 was likely acquired via LINE-1-mediated horizontal gene transfer from the African Old World monkey Papio anubis or a close ancestor 43.7-74.3 million years ago. SBHV4-S9 shows evidence of strong purifying selection following divergence from *Piliocolobus tephrosceles*. Moreover, phylogenetic analysis of the SBHV4-S1 gene and primate SLAMF6-like genes indicates this gene was acquired after divergence from a close ancestor of the New World monkey *Saimiri oerstedii* and *Saimiri sciureus* .

Comprehensive evaluation of the similarity between eight SBHV4 genes, including S1 and S9, and their homologous sequences in individual primate species,assessed using k-mer based Wasserstein distances, revealed that the overall similarity between SBHV4 genes and the homologous gene set of *Saimiri oerstedii* was greater than that between the SBHV4 genome and the homologous genes of *Piliocolobus tephrosceles*. This overall similarity to SBHV4 in *Saimiri oerstedii* was comparable to that observed in certain lemur species inhabiting Madagascar. Notably, S1 and S9 genes were more distantly related to these lemurs than to *Saimiri oerstedii*. Given that comparisons between *Saimiri oerstedii* and these lemurs were based on homologous S1, S9, and S28 genes, the S28 gene appears to be especially closely related to those of the lemurs. Although the route by which SBHV4, originating from the African Old World monkey *Piliocolobus tephrosceles*, was transmitted to the genus *Saimiri* remains unclear, our results suggest that SBHV4 has spread across diverse primate lineages from Africa to South America, initiating the acquisition of novel genomic features in its new hosts.

Introduction

Vertebrate genomes harbor a diverse array of virus-like sequences [1]. These represent remnants of ancient viral integrations resulting from host-virus interactions, existing in a spectrum from static accumulations to structurally and functionally dynamic elements. The distinction between static and dynamic is likely a continuum influenced by the accumulation of mutations over time, rather than a clear dichotomy.

Endogenous retroviruses (ERVs), in particular, exhibit dynamic behavior. Including those that have become static due to mutations, ERVs are estimated to constitute approximately 8% of the human genome [2]. While implicated in human host defense and cellular functions [3], ERVs have also been associated with the pathogenesis of diseases such as cancer and Alzheimer's disease [4, 5]. Furthermore, considering retroelements such as SINEs, their suspected involvement in the formation of mouse genome boundaries [6] suggests that these elements have played a crucial role in the acquisition and maintenance of vertebrate morphology and function from ancient times to the present.

The significant proportion of retroelements in vertebrate genomes, the recent endogenization of a retrovirus [7], and the existence of a range of sequences from static to dynamic make retroelements a fascinating subject. While retroviruses possess enzymes that facilitate the insertion of their own genomes into the host genome, The integration of other viruses is currently understood to be largely contingent upon seemingly random and stochastic events. Consequently, non-mutated, dynamic non-retroviral sequences are seldom found within vertebrate genomes, leading to the inference that ancient ancestral viruses integrated and now exist in a static state or as quasi-dynamic sequences with limited transcription in extant vertebrate genomes. Nevertheless, even static sequences, when present in substantial quantities, possess the potential to influence the three-dimensional architecture of vertebrate genomes, and any transcriptional activity originating from these non-retroviral elements could potentially modulate immune responses during infection with related viruses.

While initial compilations of non-retroviral integrations date back to the early 2000s [1], the recent accumulation of vast viral data and advancements in bioinformatics have enabled comprehensive analyses [8]. This prior study identified 2,040 virus-like sequences within 295 vertebrate genomes by screening 24,478 viral

proteins from the *Shotokuvirae* (16 families) and *Orthornavirae* (112 families) kingdoms (tblastn E-value 1e-5), and discussed the host phylogenetic relationships for major viral groups.

In this study, we performed a search using the FASTA program [9] against the entire human peptides set, specifically peptides derived from the human genome build GRCh38, with all viral proteins registered in the Virus-Host DB (https://www.genome.jp/virushostdb/) as of 2022 (RefSeq 1,860,785 entries). This analysis revealed a high degree of homology between the S9 protein of Simian beta-herpesvirus 4 (simian cytomegalovirus; SBHV4) and the human Major Histocompatibility Complex class I molecule, HLA-A, covering 86.3% of the HLA-A full length. Notably, this S9 protein is unique and was identified in a genomic region distinct from the herpesviral regions found in the aforementioned comprehensive search for virus-like proteins [8]. In this study, we discuss the evolution of this virus, focusing on the S9 protein, from a perspective slightly different from the hypothesis of ancient endogenization of a *Saimiriine beta-herpesvirus* ancestor in the human lineage.

Results＆Discussion

Homology searches using FASTA36 software between the entire viral peptide set and the human peptide set yielded 372 hits with ≥60% identity and ≥200 amino acid residues. Among these, we focused on a 315-residue overlap (86.3% of full-length HLA-A) between the membrane protein S9 (SBHV4-S9; 344 residues in length) of Saimiriine herpesvirus 4 strain (NC_016448.1) and human major histocompatibility complex class IA (HLA-A; 365 residues in length), showing 191 identical amino acids (60.6%) (Supplemental data 1). This represented the most extensive and highest homology observed thus far between a viral protein and an MHC class I molecule.

Further homology analysis of the SBHV4-S9 peptide—restricted to the HLA-A homologous region—against all other known viral peptides revealed no substantial similarity, with the highest identity being 43.2% (137/317) to a sequence from *Murmansk poxvirus* (Supplemental data 2). Notably, SBHV4-S9 exhibited higher similarity to mammalian peptides than to any viral sequence, underscoring its unique status among viral proteins. Given the low probability that the corresponding proteins in the viral species listed in Supplemental data 2 shared a recent common viral ancestor, phylogenetic analysis was not performed.

Structural analysis was conducted on the full-length peptide sequences of SBHV4-S9 and HLA-A. Based on subsequent phylogenetic analysis, we included the MHC class I-like gene from *Piliocolobus tephrosceles*—the closest relative to SBHV4—and that from *Scleropages formosus*—the most distantly related yet confidently identified homolog—for structural comparison (Fig 1). Visual inspection revealed a high degree of structural similarity among the four (Fig 1, upper panel; SBHV4-S9 magenta, HLA-A cyan, *Piliocolobus tephrosceles* gold, *Scleropages Formosus* green). Template Modeling (TM)-scores and Root Mean Square Deviation (RMSD) values between SBHV4-S9 and HLA-A, SBHV4-S9 and "MHC class I-like" (*Piliocolobus tephrosceles*), and SBHV4-S9 and "MHC class I-like"(*Scleropages formosus*) were (0.83, 1.98 Å), (0.89, 1.98 Å), and (0.78, 2.86 Å), respectively, further demonstrating significant structural resemblance based on intermolecular distances. The SBHV4-S9 peptide exhibited greater similarity to HLA-A and the *Piliocolobus tephrosceles* MHC class I-like peptide than to the *Scleropages formosus* MHC class I-like peptide. Superimposition of the SBHV4-S9 and HLA-A structures showed particularly high similarity in the α1, α2, and α3 domains of HLA-A, with SBHV4-S9 also possessing the peptide-binding pocket between α1 and α2. However, differences in the angles at the base of the branching towards the N- and C-termini and variations in individual molecular features were observed (Fig 1 A and B). Considering the potential for cross-species infection, the immunogenicity of this structure itself warrants consideration. The human cytomegalovirus UL18 protein has been previously reported as the viral protein exhibiting the highest peptide homology to HLA, displaying approximately 25% sequence identity [10]. UL18 has been shown to bind inhibitory NK cell receptors with surprisingly high affinity (over 1000-fold greater than HLA), suggesting a potential mechanism for NK cell function suppression [11].

What are the implications of the uniqueness of the SBHV4-S9 protein among previously isolated or sequenced viruses, and its striking structural similarity to HLA-A? Beyond the integration of viral genomes into the host genome, as observed with endogenous retroviruses, there are several documented instances suggesting the acquisition of host genomic fragments by viruses. For example, in the *Pestivirus* genus, the genomes of three bovine viral diarrhea virus (BVDV) strains exhibit distinct insertion sites of host-derived ubiquitin genes. Notably, the pathogenic CP1 strain and the non-pathogenic NCP1 strain, isolated from the same bovine host, showed the presence of a host ubiquitin gene insertion specifically in CP1, which was absent in

NCP1. The potential role of this ubiquitin gene insertion in the differential pathogenicity has been discussed [12].

Furthermore, studies on poxviruses, which are DNA viruses, have revealed the presence of target site duplications (TSDs) generated by the host Long Interspersed Nuclear Element-1 (LINE-1) retrotransposon. This observation suggests a LINE-1-mediated molecular mechanism for the incorporation of host genomic fragments (DNA reverse-transcribed from mRNA) into the viral genome [13].

Regarding the squirrel monkey cytomegalovirus focused on in this study, its S1 protein (285 amino acids; NC_016448.1) has been shown to possess high homology to the signaling lymphocytic activation molecule (SLAM) family receptors of the Bolivian squirrel monkey (77% identity to sbSLAMF6) [14]. It has been proposed that this S1 protein was acquired through retrotransposition during the co-evolution of the virus and its host. Similarly, the absence of intron sequences characteristic of HLA-A within the homologous region of SBHV4-S9 may indicate a potential acquisition via reverse transcription. This specific S1 homologous sequence was also identified in our current investigation.

Initially, we conducted a search for TSDs flanking the SBHV4-S9 gene. Multiple sequences, suggestive of TSDs, were identified at both ends of the HLA-A homologous region within the S9 gene, as well as at both ends of the entire S9 gene (Fig.2 The blue gradient boxes in the upper panel). While TSDs mediated by LINE-1 are typically directly adjacent to the inserted DNA, the presence of pseudo-empty sites containing polyadenylation signals can serve as supplementary evidence for LINE-1 involvement [13]. In this study, some clear polyadenylation signals were observed within the low GC content regions; however, due to accumulated mutations, the authentic polyadenylation signal utilized remains undetermined. Consequently, while the precise length and integration site of the SBHV4 MHC class I-like sequence cannot be definitively established, these findings suggest a potential transfer of genetic material from the host genome to this cytomegalovirus via a LINE-1-mediated mechanism, analogous to that reported in poxviruses.

Phylogenetic analysis (maximum likelihood method) was performed on a total of 133 MHC class I-like gene sequences, including the finally obtained SBHV4-S9 and seven human variants. To initially ascertain the branching position of SBHV4-S9, a cladogram devoid of genetic distance information was generated to assess its placement

and statistical support (Fig 3). Both SHaLRT and Ultra-fast bootstrap (UFbootstrap) values were calculated, with values ≥80% and ≥95%, respectively, considered reliable. As no integrated index for assessing both measures was available, a novel index, the S-value, based on their geometric mean, was calculated, with a threshold of 87 established for reliable support ($\sqrt{80 \ X \ 95}$ = 87.18). Phylogenetic analysis (maximum likelihood) using *Anas zonorhyncha* as an outgroup confirmed robust branching leading to SBHV4-S9, as depicted in Fig 3. Given that the SBHV4-S9 gene had diverged from a primate clade, we expanded our dataset by incorporating additional primate sequences to enable a more detailed phylogenetic analysis within primates using the maximum likelihood method. Analysis of 186 primate species together with the SBHV4-S9 gene revealed that SBHV4 diverged from *Piliocolobus tephrosceles* after the split between the hominoid clade and a subset of Old World monkeys (Fig. 4). The topology of the clade comprising 11 Old World monkey species containing the SBHV4-S9 gene was consistent with that inferred for the same clade—excluding SBHV4—by Bayesian inference (Fig. 5, right panel). For this clade, both the maximum likelihood S-values and the Bayesian posterior probabilities (BPP) were high for all nodes. The genetic distance between SBHV4-S9 and *Piliocolobus tephrosceles* estimated by maximum likelihood was substantial (Fig 4, right panel). Bayesian inference was performed excluding SBHV4-S9 for the following reasons: initial analysis including SBHV4-S9 placed its divergence basal to Asian arowana, a discrepancy likely arising from the substantial post-divergence distance of SBHV4-S9 as indicated by maximum likelihood analysis. Viruses can undergo rapid diversification during specific short periods, potentially driven by interspecies transmission and spread [15], a phenomenon likely experienced by SBHV4, distinct from vertebrate evolution. The high sequence similarity of SBHV4-S9 to primate peptides revealed by maximum likelihood analysis (Fig 4, right panel) and homology searches (Asian arowana was ranked outside the top hits in Supplemental data 3.), coupled with the greater structural similarity of SBHV4-S9 to HLA-A and *Piliocolobus tephrosceles* MHC class I-like molecules compared to that of Asian arowana (Fig 1), supported the exclusion of a pre-arowana divergence for SBHV4-S9. Consequently, Bayesian inference was conducted without the SBHV4-S9 sequence. The divergence time among the other 10 Old World monkey species belonging to the same clade as *Piliocolobus tephrosceles* was considered the oldest likely divergence of SBHV4-S9 from *Piliocolobus tephrosceles* (Fig 5, right panel, the violet arrow), estimated to have occurred between 43.7 and 74.3 million years ago (Height: 58.5 million years). This suggests that SBHV4-S9 likely diverged from *Piliocolobus tephrosceles* or a closely related ancestor within this timeframe.

Collectively, one hypothesis posits that SBHV4-S9 was transferred from the *Piliocolobus tephrosceles* MHC class I-like gene to SBHV4 via LINE-1 approximately 43.7 to 74.3 million years ago. The current host range of SBHV4 includes the common squirrel monkey *(Saimiri sciureus)*, a New World primate, while *Piliocolobus tephrosceles* is an Old World primate, implying cross-species transmission across continents. To investigate the evolutionary dynamics of this process, we first calculated dN/dS (ω) values (Fig 5, left panel), which were generally neutral across vertebrates. Although strong purifying selection or mild positive selection was observed in some primates, the values were generally consistent with neutrality. In SBHV4-S9, however, purifying selection was inferred. To assess this, a likelihood ratio test (LRT) was conducted using SBHV4-S9, yielding an LRT statistic of 240.33 (df = 1), $p < 3.33 \times 10^{-}$. In 10 million simulation trials, no instance exceeded the observed LRT statistic ($p < 1.0 \times 10^{-}$). These results reject the null hypothesis and demonstrate that SBHV4-S9 is under purifying selection. While purifying selection is common in viruses [16], our calculation was performed within the context of human HLA-A and primate MHC class I-like sequences, not solely among viral strains. Considering the generally high host specificity of cytomegaloviruses, it is plausible that SBHV4-S9 has survived under strong selective constraints imposed by its host during co-existence.

Piliocolobus tephrosceles is a primate species native to Africa, while SBHV4 was isolated from squirrel monkeys inhabiting the rainforests in Northern South America. Direct transmission of a virus with the highest sequence similarity to MHC class I like gene (*Piliocolobus tephrosceles*) across such geographical and species barriers is improbable. Parallels could be drawn to the human introduction of yellow fever virus from Africa to South America [17]; however, historical records of human-mediated transport of *Piliocolobus tephrosceles* to South America are lacking. It has been suggested that the *Chlorocebus sabaeus* was introduced from Africa to the West Indies in the 16th century, where it remains extant, while the habitat of *Piliocolobus tephrosceles* was different.[18]. Given the difficulty of resolving the close relationship between *Chlorocebus sabaeus* and *Piliocolobus tephrosceles* using phylogenetic inference based on a single MHC class I like gene, a comprehensive analysis incorporating multiple SBHV4 genes and their homologous primate genes would likely be required. Our hypothesis posits the transfer of an orthologous sequence with >60% identity to MHC class I like gene from *Piliocolobus tephrosceles* to SBHV4. As previously noted, such instances of host gene acquisition by viruses are not unprecedented; as seen with the S1 protein of squirrel monkey cytomegalovirus is proposed to have originated from a SLAMF receptor gene of the Bolivian squirrel

monkey [14]. Thus, supporting both our current hypothesis and previous findings suggests the potential for squirrel monkey cytomegalovirus to have acquired cDNA sequences from other primate genomes as well.

First, we performed maximum likelihood (ML) phylogenetic inference between the SBHV4 S1 gene—previously suggested to have originated from the squirrel monkey SLAMF6-like gene—and primate (121 species) SLAMF6-like genes (Fig. 6). The analysis confirmed that the S1 gene diverged from Saimiri boliviensis and subsequently from the common ancestor of *Saimiri oerstedii* and *Saimiri sciureus*.

To conduct a comprehensive analysis of multiple SBHV4 genes and their primate homologs, we retrieved, using Spaln3, all primate genes homologous to SBHV4 proteins from the genomes of 241 primate species (Supplemental data 4). This search identified eight SBHV4 genes, including S1 and S9, with high sequence similarity to primate homologs. Using all these homologous sequences, we examined the overall relatedness between SBHV4 and 241 primate species, aiming to bridge the “missing link” between *Piliocolobus tephrosceles* and the common squirrel monkey. For this purpose, we applied a k-mer–based Wasserstein distance approach to investigate the phylogenetic relationships between SBHV4 and these primates. Prior to the main analysis, the validity of this method was assessed through a benchmark comparison with Iqtree2. Specifically, we used 49 species spanning tarsiers, New World monkeys, Old World monkeys, and apes—whose species phylogenetic relationships are well established—and analyzed 13 mitochondrial DNA genes plus two rRNA sequences with Iqtree2. For Iqtree2, sequences were aligned for each gene across the 49 species using MAFFT, concatenated, and analyzed under a partitioned model to account for differences in substitution rates. Because Wasserstein distances cannot be directly compared with Iqtree2 outputs, we applied neighbor-joining (NJ) clustering to the Wasserstein distance matrix, generated a Newick file, and visualized the resulting trees for comparison (ML estimation is not possible directly from distance matrices). Correlation between pairwise path lengths from the NJ trees and the Wasserstein distances was assessed using a Mantel test, revealing a highly significant correlation ($r = 0.9064$, $P = 0.0001$). This indicates that pairwise path lengths from NJ clustering can serve as a valid proxy for Wasserstein distances. While the Iqtree2 analysis did not fully preserve known phylogenetic relationships among the tested species, the present method retained the major clade separations (Fig. 8): divergence of tarsiers from the primate ancestor, followed by the split between New World monkeys and the lineage leading to Old World monkeys and apes. Within the apes, the method correctly placed gibbons as diverging from the common ancestor of the Hominoidea, but subsequently grouped

humans with gorillas, deviating from the established species-level topology. Nonetheless, in the broader context, our multi-gene phylogenetic analysis exhibited performance on par with, or superior to, that of Iqtree2 (ML), supporting the robustness of our approach.

Using 100 unprocessed Wasserstein distance matrices derived from homologous sequences of SBHV4 genes and various primate species, we compared the genetic proximity of SBHV4 to each primate taxon (Fig. 9). Median pairwise distances between SBHV4 and *Saimiri oerstedii* were significantly shorter than those between SBHV4 and *Piliocolobus tephrosceles* (Dunn's Post-Hoc Test: p=8.36e-10), indicating closer affinity to *Saimiri oerstedii*. A simple ranking of median Wasserstein distances for each primate species placed *Saimiri oerstedii* second overall, with the top five taxa all harboring homologs of SBHV4 S1, S9, and S28 genes. Kruskal–Wallis H-tests detected no significant differences among the medians of these top five species, reinforcing the conclusion that *Saimiri oerstedii* is among the closest living relatives of SBHV4. Conversely, three lemurs, *Propithecus_deckenii_coronatus, Eulemur rufifrons Lepilemur ruficaudatus* living in native to Madagascar, are as closely related to SBHV4 as *Saimiri oerstedii*. Notably, the S1 and S9 genes of these lemurs were more distantly related to SBHV4 than those of *Saimiri oerstedii* (S1: Lemurs lack SLAMF6 gene homolog to that of humans, S9: Fig. 4), implying that the S28 gene exhibits a particularly close relationship to lemur homologs. Based solely on the present findings, a plausible scenario is that SBHV4 acquired the S9 gene from *Piliocolobus tephrosceles* approximately 43.7–74.3 million years ago, subsequently acquired the S28 gene via lemurs in Madagascar, and finally reached the *Saimiri* lineage, where it obtained the S1 gene. Such a trajectory cannot be inferred from host phylogenies alone.

In the primate species closely related to SBHV4, as shown in Figure 9, the expected phylogenetic relationships were clearly disrupted. It is important to note that the Wasserstein distance calculation inherently accommodates unequal gene numbers among taxa (Supplemental data 4). While this study assumed that SBHV4 evolved through successive acquisition of homologous genes, “gene snatching", from its primate hosts, the hosts themselves may have independently evolved their own versions of these genes, suggesting a highly complex co-evolutionary history between SBHV4 and its primate hosts.

In conclusion, although the “missing link” between *Piliocolobus tephrosceles* and the common squirrel monkey remains unresolved, our results support the hypothesis that SBHV4 acquired the S9 gene from the MHC class I–like gene of *Piliocolobus tephrosceles* and later obtained additional genes from *Saimiri oerstedii*. This, in turn,

suggests that anthropogenic translocation of primates from Africa to South America—possibly associated with transoceanic navigation beginning around 8000 BCE [19]—underlies the current genetic composition of SBHV4. Accordingly, some SBHV4 genes may be relatively recent acquisitions, and the virus may have undergone novel evolutionary changes after its arrival in South America.

Materials&Methods

Entire Viral Peptide vs. Entire Human Peptide Set

The complete viral peptide set (release 225), updated August 6, 2024, was downloaded from GenomeNet (Virus-Host Database). The entire human peptide set, corresponding to the GRCh38 genome build, was obtained from NCBI. Homology searches were performed using FASTA36 (-E 1e-20) [9], and results were screened for hits with ≥60% identity and ≥200 amino acid residues (Supplemental data 1).

SBHV4-S9 vs. All Viral Peptides

To investigate the relationship of SBHV4-S9 (YP_004940188.1; 344 residues, full length) to other viral peptides, a FASTA36 homology search was performed using the SBHV4-S9 peptide sequence (HLA-A homologous region only) and the aforementioned entire viral peptide dataset (Supplemental data 2).

SBHV4-S9 TSD Identification

To investigate the presence of TSDs flanking the HLA-A homologous region of SBHV4-S9, TSD searches were performed using the SBHV4 genome (NC_016448.1) and the high HLA-A homologous region with SBHV4-S9. The SBHV4-S9 locus within the SBHV4 genome was determined, and 200 bp flanking sequences were retrieved. The 5' 200 bp region was split into overlapping fragments (4-20 bp sliding window), and each fragment was used to query the 3' 200 bp region using the Needleman-Wunsch alignment algorithm [20] (Code 1). TSD sequences exhibiting ≥60% identity were output in FASTA format, converted to FASTQ format (with maximum quality scores), and subsequently mapped using HISAT2 [21] to a composite sequence (“the SBHV4 concatemer”) comprising the 200 bp flanking the 5' end of the SBHV4-S9 gene, 50 'N's

mimicking its HLA-A homologous region, and the 200 bp flanking its 3' end. This procedure was applied to both the HLA-A homologous region and the full-length DNA of SBHV4-S9. Lastly, the similarity between the paired TSDs flanking the region of 50 consecutive "N"s was ranked by calculating their Hamming distance.

SBHV4-S9 vs. Peptides from Other Animal Species

To identify peptide sequences homologous to SBHV4-S9 in diverse species, the SBHV4-S9 peptide (HLA-A homologous region only) was queried against the comprehensive peptide datasets of all species available at https://ftp.ensembl.org/pub/release-112 using a homology search (-b 1 -d 1 -E 1e-10 -m BB). Results were filtered for sequences exhibiting ≥300 amino acid residues, ≥50% identity, and a query start position <3 amino acid residues in NCBI BLAST output format. A tabular summary of the homology observed for each species was compiled (Supplemental data 3).

Structural Analysis of Full-Length SBHV4-S9 and MHC Class I Peptides

Structural analysis of the full-length sequences of SBHV4-S9 (YP_004940188.1), HLA-A (AZU89775.1), the *Piliocolobus tephrosceles* MHC class I-like peptide (African origin; closest homolog identified by Spaln3[33]), and the *Scleropages formosus* MHC class I-like peptide (XP_029104321.1; most distant homolog identified by FASTA36) was performed using AlphaFold 2.3.2 [22]. Full-length sequences were employed to explore potential functional similarities beyond high-homology subregions. TM-scores for pairwise comparisons were normalized by the length of the shorter sequence. The resulting four PDB files were used to assess pairwise structural similarity via the TM-align web server [23], and visualized using ChimeraX software [24].

Molecular Phylogenetic Analysis of MHC Class I-like and SBHV4-S9 Genes

To elucidate the molecular evolutionary history of SBHV4-S9 in relation to MHC class I-like genes, we performed DNA phylogenetic analysis.

The sequence corresponding to the high-homology region of SBHV4-S9 within HLA-A peptide was manually extracted from the Ensembl entry ENST00000706898.1

(HLA-A-216). For SBHV4-S9, the homologous region to HLA-A was extracted from the SBHV4 genome (NC_016448.1).

Subsequently, CDS data for all species available on the Ensembl FTP site (https://ftp.ensembl.org/pub/release-112) were downloaded and subjected to homology searches (-b 1 -d 1 -E 1e-10 -m BB). Resulting BLAST-formatted .txt files within this directory were filtered to retain sequences ≥900 bp in length, initiating with a query start position ≤15 (to ensure data integrity), yielding FASTA36 results for 126 animal species + 1 virus species. All MHC class I-like sequences for these 126 animal species + 1 virus species (SBHV4-S9) were compiled into a single FASTA file

To investigate associations with human variants, HLA-A haplotype sequences were collected from human genome CRAM files available from the International Genome Sample Resources (IGSR; https://www.internationalgenome.org/data-portal/data-collection/30x-grch38). A comprehensive list of available CRAM files (igsr_30x GRCh38.tsv.tsv) was downloaded from the IGSR website (Available data tab, as of July 2024). A list of up to four individuals per cohort with non-redundant Population designations was generated (igsr_30x GRCh38_rev.tsv; 104 individuals). CRAM files for these 104 individuals were retrieved. FASTQ files extracted from these CRAM files were mapped to HLA-A_gene_GRCh38.fa (EU445472.1). De novo assembly was performed on the resulting .sam files using SPAdes software (-k 31) [25]. Following careful examination of open reading frames and homologies, and excluding redundant HLA-A haplotypes, six human individuals exhibiting ≥99% homology to HLA-A CDS datum (GRCh38) were selected as candidates for phylogenetic analysis (IDs specified in Fig 3 and Fig 4). The FASTA data for these human individuals, previously recovered other species (126 animal species including humans + 1 virus species), and SBHV4-S9 were concatenated into a final file (containing 133 fasta entries; Supplemental data 5). The presence of the motif characteristic of MHC class I molecules was confirmed across all FASTA data using MEME suite (https://meme-suite.org/meme/index.html) [26]. Finally, this dataset underwent global alignment with MAFFT [27] and trimming with ClipKIT [28], followed by maximum likelihood phylogenetic tree inference using Iqtree2 [29] (substitution model selection: -m MFP+LM; UFbootstrap with 7 million replicates including SHaLRT calculation). Branch support reliability was assessed using a composite S-value, calculated as the geometric mean of SHaLRT and UFbootstrap values output by Iqtree2.

$$S\text{-}value=\sqrt{SHaLRT\ X\ UFbootstrap}$$

Nodes with S-values ≥87% were considered reliably supported and visualized using FigTree (https://github.com/rambaut/figtree). Initially, a cladogram was generated to illustrate branching patterns. Given the close phylogenetic affinity of SBHV4-S9 to primates among vertebrates, as of June 2025 we downloaded all primate genomes available in NCBI. Using the HLA-A peptide sequence of the SBHV4-S9 homologous region, we retrieved the corresponding coding sequences via Spaln3 (-H 1000). Species with excessive ambiguous nucleotides (N), extremely short nucleotide sequences, or in-frame stop codons were excluded, yielding a final dataset of 186 primate sequences (Supplemental data 6). The presence of an MHC class I motif in these sequences was verified. Following global alignment with MAFFT, phylogenetic inference was performed using Iqtree2.

Divergence Time Estimation of SBHV4-S9

From the final FASTA file used for maximum likelihood analysis (Supplemental data 6, excluding SBHV4-S9), FASTA data for primates and outgroups (*Monodelphis domestica*, *Cavia porcellus*, *Ornithorhynchus anatinus*, *Oryctolagus cuniculus*) were constructed. Following global alignment with MAFFT and parameter acquisition with Iqtree2, an XML file was generated using BEAUti [30] and BEAST 2 was executed [30] (1 billion MCMC iterations; ESS values ≥200 for all parameters). From the resulting treefile, a Maximum Clade Credibility tree was generated after discarding the initial 30% burn-in using TreeAnnotator [30]. Numerical values pertaining to the divergence time estimation of *Piliocolobus tephrosceles* and 10 Old World monkeys were extracted from this Nexus format file.

dN/dS Calculation

dN/dS (ω) values were calculated using CODEML in PAML [31]. For codon alignment, the FASTA file from Supplemental data 6 was aligned using PRANK software [32], employing the phylogenetic tree from Fig 5. CODEML was configured (model = 1, NSsites = 0) to estimate a single ω value for all branches, assuming a uniform ω across sites, to infer the ω value of SBHV4-S9 within the broader vertebrate context (details in Supplemental data 7). To facilitate visualization given the range of ω values, log□ □(ω) was calculated.. To statistically assess negative selection on

SBHV4-S9, ω value for SBHV4-S9 was estimated separately under the same site-homogeneous ω assumption (model = 2; all other settings identical to Supplemental data 7). A likelihood ratio test (LRT) was subsequently performed using the results from a null model where all ω values were fixed (model = 0).

Molecular Phylogenetic Analysis of SLAMF6-like and Primate SBHV4-S1 Genes

To elucidate the co-evolutionary history of SBHV4-S1 with primate SLAMF6-like genes, the corresponding DNA sequences for the high-homology region between SBHV4-S1 and human SLAMF6 (AB590474.1) were obtained using human SLAMF6 peptides (Spaln3; -H 800). Species with excessive ambiguous nucleotides (N), extremely short nucleotide sequences, or in-frame stop codons were excluded, yielding SLAMF6-like sequences from 112 primate sequences. To assess preliminary orthology, motif searches were performed using MEME suite [26], as previously described. All sequences commonly exhibited Ig_3 motif characteristic of SLAMF6. Following global alignment with MAFFT, maximum likelihood phylogenetic inference was conducted using Iqtree2, employing -m MFP for substitution model selection, with Ultra-fast bootstrap (10 thousands replicates) including SHaLRT. The results were converted to the aforementioned S-values and visualized.

Phylogenetic Relationship Between SBHV4 and Primates

We retrieved the complete set of 141 peptide sequences encoded by SBHV4 (NC_016448). Using these sequences, we queried the genomes of 244 primate species with Spaln3 (-H 100). Regions with at least one match were detected in 243 species. Among the SBHV4 genes, eight (S1, S8, S9, S11, S28, UL54, UL116, and A23) exhibited homology to primate genomes (Supplemental data 4). None of the primate species contained all eight genes. Homologous sequences for these eight SBHV4 genes were extracted in multi-FASTA format using the GFF files generated by Spaln3 for each primate species. After excluding one species lacking any homologous genes (Tarsius wallacei) and two species whose retrieved sequences contained excessive ambiguous nucleotides (N), the dataset comprised 241 primate species plus the eight SBHV4 genes (Supplemental data 8). These were used for the subsequent analyses. The relationship between SBHV4 and primate species was assessed by inferring phylogenetic relationships among the above DNA sequences using a method based on

k-mer frequency distributions and the Wasserstein distance (Fig.8). The output of the Wasserstein distance matrix was generated using Code 2. Supplemental data 8 contains the 241 files used, with the entire fasta data for each file listed sequentially after its filename.

1. *Construction of the Nucleotide Substitution Cost Matrix*

   To quantify the "cost" of nucleotide substitutions, we constructed a substitution cost matrix from the Q-matrix obtained by analyzing all fasta files with Iqtree2. The off-diagonal elements of the Q-matrix represent the relative probability or rate of substitution between two nucleotides. The reciprocal of each off-diagonal element was adopted as the raw substitution cost between those nucleotides, while diagonal elements (identical nucleotides) were assigned a cost of zero. The raw cost matrix was linearly scaled to the range 0.5–1.5 to normalize values and prevent the contribution of substitution costs in subsequent k-mer distance calculations from depending excessively on the scale of the original Q-matrix. The resulting normalized substitution cost matrix, C_{nuc}, is a 4 × 4 square matrix where $C_{nuc}(i,j)$ denotes the cost of substituting nucleotide i with j.

2. *Selection of Representative k-mers*

   To reduce computational burden, we selected 2000 representative k-mers (K_{rep}) from the entire dataset. First, minimizers were extracted from each fasta file (one per species) using k-mer length (k=180) and window (size w=90), capturing local genomic information efficiently. Next, from the combined set of unique minimizers across all fasta files, we randomly sampled 2000 k-mers with approximately equal representation from each file, avoiding duplication.

3. *Construction of the Base Distance Matrix Between k-mers*

For every pair of k-mers in K_{rep}, we computed the weighted Hamming distance to produce the base distance matrix DG. Given two k-mers

$s_1=s_{1,1}s_{1,2}\ldots s_{1,k}$ and $s_2=s_{2,1}s_{2,2}\ldots s_{2,k}$, the weighted Hamming distance $d\mathrm{H}(s_1,s_2)$ was defined as:

$$d\mathrm{H}(s_1,s_2)=\sum_{i=1}^{k} C_{nuc}(s_{1,i},s_{2,i})$$

where C_{nuc} is as defined in Section 1. DG is symmetric, with DG(i,j) representing the weighted Hamming distance between k-mer k_i and k_j. This matrix served as the ground metric in the Wasserstein distance calculation described in Section 5.

4. *Calculation of k-mer Frequency Distributions for Each Species*

For each primate species, we computed the frequency distribution of k-mers in its DNA sequences, using K_{rep} as the reference set. Each sequence-derived k-mer was mapped to the nearest element K_{rep} based on weighted Hamming distance, subject to a mismatch cost threshold $T_{mismatch}$, defined as:

$$T_{mismatch}=k\times r\times C_{min_nuc}$$

where C_{min_nuc} = 0.5 (the normalized minimum substitution cost), k=180, and r=0.175. This threshold corresponds to allowing substitutions at the minimum cost across 17.5% of the k-mer length.

If $d\mathrm{H}(s_{seq},k_{min}) \leq T_{mismatch}$, the sequence k-mer was "fuzzily" assigned to k_{min} (i.e., in a non-exact, proximity-based manner), and its frequency count was incremented for that species. Raw frequencies were normalized by the total number of k-mers to yield probability distribution vectors P_f, where $P_f(j)$ denotes the normalized frequency of k_j in species f.

5. *Computation of the Wasserstein Distance Matrix*

Pairwise similarities between species were quantified by computing Wasserstein distances between their k-mer frequency distributions P_f using the POT library's ot.emd2 function. Inputs to ot.emd2 included two probability vectors and the base

distance matrix D_G; the square root of the resulting squared Wasserstein distance was taken as the final distance. All pairwise distances were compiled into the Wasserstein distance matrix D_W.

6. *Phylogenetic Tree Construction*

The D_W matrix was used to construct phylogenetic trees via the Neighbor-Joining (NJ) method (Code 3). To assess tree robustness, 100 bootstrap replicates were generated by resampling k-mers with replacement, followed by recalculation of k-mer distributions, D_W, and NJ trees as in Sections 4–5. The resulting 100 NJ trees were output in Newick format. Robinson–Foulds (RF) distances among these trees were computed, and a consensus tree was generated, with branch frequencies across the bootstrap trees recorded as bootstrap support values (Code 4).

Benchmarking Against Iqtree2

Benchmarking was conducted using 49 primate species selected from the SBHV4 homology list:

- Apes: Family *Hominidae* and Family *Hylobatidae* (19 species)
- Tarsiers: Family *Tarsiidae* (3 species + *Tarsius wallacei*)
- New World monkeys: Family *Aotidae* (5 species) and Genus *Saimiri* (6 species)
- Old World monkeys: Genus *Macaca* (15 species)

From the mitochondrial genome of *Carlito syrichta* (NC_012774), all 13 protein-coding genes and two rRNAs were retrieved. Using fasta36 (-E 1e-10), 15 homologous DNA sequences were obtained for each of the 49 species. Multiple sequence alignments were performed gene-by-gene with mafft, then concatenated per species. Iqtree2 was run with 15 partitions (-m MFP -alrt 10000 -bb 10000 -ninit 10000 -ntop 1000 -o Carlito_syrichta), producing a Q-matrix that was then applied in the Wasserstein distance analysis described above.

Comparison of SBHV4–Primate Relationships

Following benchmarking, 100 Wasserstein distance matrices were used to compute the

median Wasserstein distance between SBHV4 and each primate species. Statistical testing was performed with the Kruskal–Wallis H-test; when significant ($p < 0.05$), Dunn's post hoc test was applied.

Figure Legends

Fig1 **Molecular Models of MHC Class I-like Peptides (AlphaFold).** Magenta: SBHV4-S9; cyan: human HLA-A; gold: *Piliocolobus tephrosceles* MHC class I-like peptide; green: *Scleropages formosus* MHC class I-like peptide. **Middle Left:** Superimposed structures of SBHV4-S9 and HLA-A. **Middle Right:** Superimposed structures of SBHV4-S9 and *Scleropages formosus* MHC class I-like peptide. TM-score and RMSD values quantifying structural similarity between the respective molecular models are indicated. α1, α2, α3 denote the HLA-A subunits. **Bottom:** Structural divergence between SBHV4-S9 (magenta) and HLA-A (cyan) at the C-terminus (A) and N-terminus (B).

Fig2 **Predicted Target Site Duplications (TSDs) Flanking the SBHV4 S9 Gene.** The middle color-coded section illustrates the MHC class I like gene in S9 gene (grey) and surrounding nucleotides. For visualization, the 945 bp inside S9 gene is compressed to 50 bp (referred to as MHC-I like). Putative flanking short sequences, suggestive of TSDs, are observed at both ends of the compressed HLA-A homologous region in and around the S9 gene (the upper panel). The top three pairs with the shortest Hamming distances were numbered within the box, with the complete sequence for each pair listed. The bottom panel displays a histogram (window size 25) of the GC content..

Fig3 **Phylogenetic Analysis of SBHV4-S9 Gene and Primate MHC Class I-like Sequences.** Cladogram (maximum likelihood) illustrating branching patterns of 133 FASTA sequences, including SBHV4-S9 and seven human individuals, with the magenta arrow indicating the SBHV4-S9 gene's position. Node colors represent S-values (geometric mean of SHaLRT and UFbootstrap values), with ≥87 indicating reliable support.

Fig4 **Phylogenetic relationship between the MHC class I-like genes of 186 primate species and the SBHV4-S9 gene**
Left panel: A cladogram omitting genetic distances, inferred using the maximum likelihood method. The position of SBHV4 is indicated by a magenta arrow. Primate species belonging to the genus *Saimiri* are highlighted in light blue. Node colors reflect the magnitude of S-values. *Right panel:* A maximum likelihood phylogenetic tree incorporating genetic distances. The branching pattern is identical to that in the left panel. The analysis reveals a bifurcation into two major clades. The SBHV4-S9 gene

diverges from *Piliocolobus tephrosceles* with an extended branch length. MHC class I genes of lemurs are located within the major clade to which SBHV4-S9 does not belong, and are positioned as the most distantly related to SBHV4-S9.

Fig5 **dN/dS (ω) Values and Bayesian Phylogenetic Tree. Left:** dN/dS (ω) values for each branch calculated from 186 primate sequences. The yellow arrow indicates the SBHV4-S9 gene. Values ≥1.5 (log□□(ω)) are red, 0 (log□□(ω) = 1) is green, and ≤-1.5 (log□□(ω)) are blue. Purifying selection is primarily inferred for SBHV4-S9. **Right:** Bayesian phylogenetic tree excluding SBHV4-S9. The purple arrow indicates the divergence point of *Piliocolobus tephrosceles*. Node colors represent Bayesian Posterior Probabilities (BPP; threshold ≥80%): red (<80%), green (90%), blue (100%). Bars on nodes indicate 95% Highest Posterior Density intervals (HPD).

Fig6 **Maximum Likelihood Phylogenetic Analysis of 121 Primate SLAMF6-like and SBHV4-S1 Genes.** Node colors represent S-values (geometric mean of SHaLRT and UFbootstrap values). Nodes with missing values during tree construction (rooted with *Carlito syrichta*) are black. SBHV4-S1 **arose** from a lineage that had already separated from *Saimiri boliviensis*, and it later branched from the **predecessor** of *Saimiri oerstedii* and *Saimiri sciureus*.

Fig7 **Schematic overview of the Wasserstein distance calculation based on k-mer frequency distributions.**

The computed Wasserstein distances (right) were used to construct phylogenetic trees via the Neighbor-Joining (NJ) method, which is well-suited to numerical distance matrices. These trees were then compared with those inferred using the maximum likelihood method implemented in IQ-TREE2. Prior to benchmarking (Fig. 8), the correlation between the raw Wasserstein distances and the pairwise path lengths derived from the NJ trees was assessed using the Mantel test. The actual phylogenetic proximity between SBHV4 and primates was evaluated by applying violin plot analysis to the raw Wasserstein distances (Fig. 9).

Fig8 **Benchmark comparison between phylogenetic trees inferred using the Neighbor-Joining (NJ) method based on Wasserstein distances and the maximum likelihood method implemented in IQ-TREE2.**

Phylogenetic inference with Iqtree2 was performed using 15 mitochondrial and rRNA genes (13 mitochondrial protein-coding genes and 2 rRNAs) from 49 primate species

spanning Tarsiers, New World monkeys (NWM), Old World monkeys (OWM), and Apes (left panel). Genes were aligned individually with MAFFT, concatenated by species, and partitioned for analysis by Iqtree2. Node support values represent S-values.
The NJ tree (right panel) was constructed from the Wasserstein distance matrix. A representative NJ tree was selected based on Robinson–Foulds (RF) distances from 100 random NJ trees. Node values indicate the frequency of each split across replicates.
Although the inferred trees do not fully recapitulate the true species phylogeny based on fossil evidence, the overall topology—separating Tarsiers, followed by NWM, and then OWM and Apes—is retained.

Fig9 **Ranking of primate species based on Wasserstein distances to SBHV4.**
Wasserstein distances were calculated between eight SBHV4 genes and their homologs in 241 primate species. The ten species with the smallest distances to SBHV4 are shown in light blue (upper panel), ranked from left to right. Additionally, *Piliocolobus tephrosceles*, *Saimiri oerstedii*, *Saimiri sciureus*, and *Saimiri boliviensis* are included in brown. Notably, *Saimiri oerstedii* ranked second. Below each species name, the corresponding major taxonomic group and homologous SBHV4 genes are indicated. All of these genes were used in the calculation of Wasserstein distances. Conversely, any of the eight SBHV4 genes not listed here indicates the absence of a detectable homologue in the corresponding species. Among the four brown-highlighted species, significant differences were confirmed by Kruskal–Wallis H-test ($p < 0.05$), followed by Dunn's post hoc test (lower panel), which revealed significant divergence involving *Saimiri oerstedii*. In contrast, among the top five ranked species including *Saimiri oerstedii*, no significant pairwise differences were detected at the Kruskal–Wallis stage.

References

[1] Endogenous viruses: insights into viral evolution and impact on host biology
Cédric Feschotte, Clément Gilbert. Nat Rev Genet 2012 Mar 16;13(4):283-96. doi: 10.1038/nrg3199.

[2] Initial sequencing and analysis of the human genome
Lander E.S. et al. Nature 2001 Feb 15;409(6822):860-921. doi: 10.1038/35057062.

[3] Human Endogenous Retroviruses Are Ancient Acquired Elements Still Shaping Innate Immune Response
Nicole Grandi, Enzo Tramontano. Front Immunol 2018 Sep 10:9:2039. doi: 10.3389/fimmu.2018.02039. eCollection 2018.

[4] Human Endogenous Retrovirus-K (HML-2)-Related Genetic Variation: Human Genome Diversity and Disease
Wonseok Shin, Seyoung Mun, Kyudong Han. Genes (Basel) 2023 Nov 28;14(12):2150. doi: 10.3390/genes14122150.

[5] Awakening the dark side: Retrotransposon activation in neurodegenerative disorders
Elizabeth Ochoa Thomas, Gabbe Zuniga, Wenyan Sun, Bess Frost. Curr Opin Neurobiol. 2020 Feb 21;61:65–72. doi: 10.1016/j.conb.2020.01.012

[6] B2 SINE Copies Serve as a Transposable Boundary of DNA Methylation and Histone Modifications in the Mouse
Tomoko Ichiyanagi, Hirokazu Katoh, Yoshinobu Mori, Keigo Hirafuku, Beverly Ann Boyboy, Masaki Kawase, Kenji Ichiyanagi. Mol Biol Evol 2021 May 19;38(6):2380-2395. doi: 10.1093/molbev/msab033.

[7] Retroviral invasion of the koala genome
Rachael E. Tarlinton, Joanne Meers, Paul R. Young. Nature 2006 Jul 6;442(7098):79-81. doi: 10.1038/nature04841.

[8] Deep mining reveals the diversity of endogenous viral elements in vertebrate genomes
Jose Gabriel Nino Barreat, Aris Katzourakis. Nat Microbiol 2024 Nov;9(11):3013-3024.

doi: 10.1038/s41564-024-01825-4. Epub 2024 Oct 22.

[9] Finding protein and nucleotide similarities with FASTA
William Pearson. Curr Protoc Bioinformatics 2004 Feb:Chapter 3:Unit3.9. doi: 10.1002/0471250953.bi0309s04.

[10] Human cytomegalovirus encodes a glycoprotein homologous to MHC class-I antigens
Beck S, Barrell B. G. Nature 1988 Jan 21;331(6153):269-72. doi: 10.1038/331269a0.

[11] Structure of UL18, a peptide-binding viral MHC mimic, bound to a host inhibitory receptor
Zhiru Yang, Pamela J Bjorkman. Proc Natl Acad Sci USA 2008 Jul 22;105(29):10095-100. doi: 10.1073/pnas.0804551105. Epub 2008 Jul 15.

[12] Viral cytopathogenicity correlated with integration of ubiquitin-coding sequences
Gregor Meyers, Norbert Tautz, Edward J Dubovi, Heinz-Jürgen Thiel. Virology 2004 Feb 9;180(2):602–616. doi: 10.1016/0042-6822(91)90074-L

[13] Poxviruses capture host genes by LINE-1 retrotransposition
Sarah M Fixsen, Kelsey R Cone, Stephen A Goldstein, Thomas A Sasani, Aaron R Quinlan, Stefan Rothenburg, Nels C Elde. eLife 2022 Sep 7;11:e63332. doi: 10.7554/eLife.63332

[14] Signaling Lymphocytic Activation Molecule Family Receptor Homologs in New World Monkey Cytomegaloviruses
Natàlia Pérez-Carmona, Domènec Farré, Pablo Martínez-Vicente, Cox Terhorst, Pablo Engel, Ana Angulo. J Virol 2015 Nov;89(22):11323-36. doi: 10.1128/JVI.01296-15. Epub 2015 Sep 2.

[15] Dispersal history of Miniopterus fuliginosus bats and their associated viruses in east Asia
Kimprasit T et al. PLoS One 2021 Jan 14;16(1):e0244006. doi: 10.1371/journal.pone.0244006. eCollection 2021.

[16] Evolutionary insights into host–pathogen interactions from mammalian sequence

data
Manuela Sironi, Rachele Cagliani, Diego Forni, Mario Clerici. Nat Rev Genet 2015 Apr;16(4):224-36. doi: 10.1038/nrg3905.

[17] Out of Africa: a molecular perspective on the introduction of yellow fever virus into the Americas
Juliet E Bryant, Edward C. Holmes, Alan D. T Barrett. PLoS Pathog 2007 May 18;3(5):e75. doi: 10.1371/journal.ppat.0030075.

[18] Insights into the Geographical Origins of the Cabo Verde Green Monkey
Lara Almeida, Ivo Colmonero-Costeira, Maria J Ferreira da Silva, Cecilia Veracini, Raquel Vasconcelos. Genes (Basel) 2024 Apr 17;15(4):504. doi: 10.3390/genes15040504.

[19] The earliest evidence of ocean navigation
Robert G. Bednarik. Int J Naut Archaeol 1997; 26(3):183 doi: 10.1006/ijna.1997.0092

[20] A general method applicable to the search for similarities in the amino acid sequence of two proteins
Needleman S. B., Wunsch C. D. J. Mol Biol 1970 Mar;48(3):443-53. doi: 10.1016/0022-2836(70)90057-4.

[21] HISAT: a fast spliced aligner with low memory requirements
Daehwan Kim, Ben Langmead, Steven L Salzberg. Nat Methods 2015 Apr;12(4):357-60. doi: 10.1038/nmeth.3317. Epub 2015 Mar 9.

[22] Highly accurate protein structure prediction with AlphaFold
John Jumper et al. Nature 2021 Aug;596(7873):583-589. doi: 10.1038/s41586-021-03819-2. Epub 2021 Jul 15.

[23] TM-align: a protein structure alignment algorithm based on the TM-score
Yang Zhang, Jeffrey Skolnick. Nucleic Acids Res. 2005 Apr 22;33(7):2302–2309. doi: 10.1093/nar/gki524

[24] UCSF ChimeraX: Tools for structure building and analysis

Elaine C. Meng, Thomas D. Goddard, Eric F. Pettersen, Greg S. Couch, Zach J. Pearson, John H. Morris, Thomas E. Ferrin. Protein Sci. 2023 Nov 1;32(11):e4792. doi: 10.1002/pro.4792

[25] SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing
Anton Bankevich et al. J Comput Biol 2012 May;19(5):455-77. doi: 10.1089/cmb.2012.0021. Epub 2012 Apr 16.

[26] The MEME Suite
Timothy L. Bailey, James Johnson, Charles E. Grant, William S. Noble Nucleic Acids Res 2015 Jul 1;43(W1):W39-49. doi: 10.1093/nar/gkv416. Epub 2015 May 7.

[27] MAFFT version 5: improvement in accuracy of multiple sequence alignment.
Kazutaka Katoh, Kei-ichi Kuma, Hiroyuki Toh, Takashi Miyata. Nucleic Acids Res 2005 Jan 20;33(2):511-8. doi: 10.1093/nar/gki198. Print 2005.

[28] ClipKIT: A multiple sequence alignment trimming software for accurate phylogenomic inference
Jacob L. Steenwyk, Thomas J. Buida 3rd, Yuanning Li , Xing-Xing Shen, Antonis Rokas. PLoS Biol 2020 Dec 2;18(12):e3001007. doi: 10.1371/journal.pbio.3001007. eCollection 2020 Dec.

[29] IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era
Bui Quang Minh, Heiko A. Schmidt, Olga Chernomor, Dominik Schrempf, Michael D. Woodhams, Arndt von Haeseler, Robert Lanfear. Mol Biol Evol. 2020 May 1;37(5):1530-1534. doi: 10.1093/molbev/msaa015.

[30] BEAST 2: a software platform for Bayesian evolutionary analysis
Remco Bouckaert, Joseph Heled, Denise Kühnert, Tim Vaughan, Chieh-Hsi Wu, Dong Xie, Marc A. Suchard, Andrew Rambaut, Alexei J. Drummond. PLoS Comput Biol 2014 Apr 10;10(4):e1003537. doi: 10.1371/journal.pcbi.1003537. eCollection 2014 Apr.

[31] PAML 4: phylogenetic analysis by maximum likelihood
Ziheng Yang. Mol Biol Evol 2007 Aug;24(8):1586-91. doi: 10.1093/molbev/msm088.

Epub 2007 May 4.

[32] Phylogeny-aware alignment with PRANK
Ari Löytynoja. Methods Mol Biol 2014:1079:155-70. doi: 10.1007/978-1-62703-646-7_10.

[33] Spaln3: improvement in speed and accuracy of genome mapping and spliced alignment of protein query sequences Open Access
Osamu Gotoh. Bioinformatics 2024:40(8)btae517. doi: 10.1093/bioinformatics/btae517

Data availability

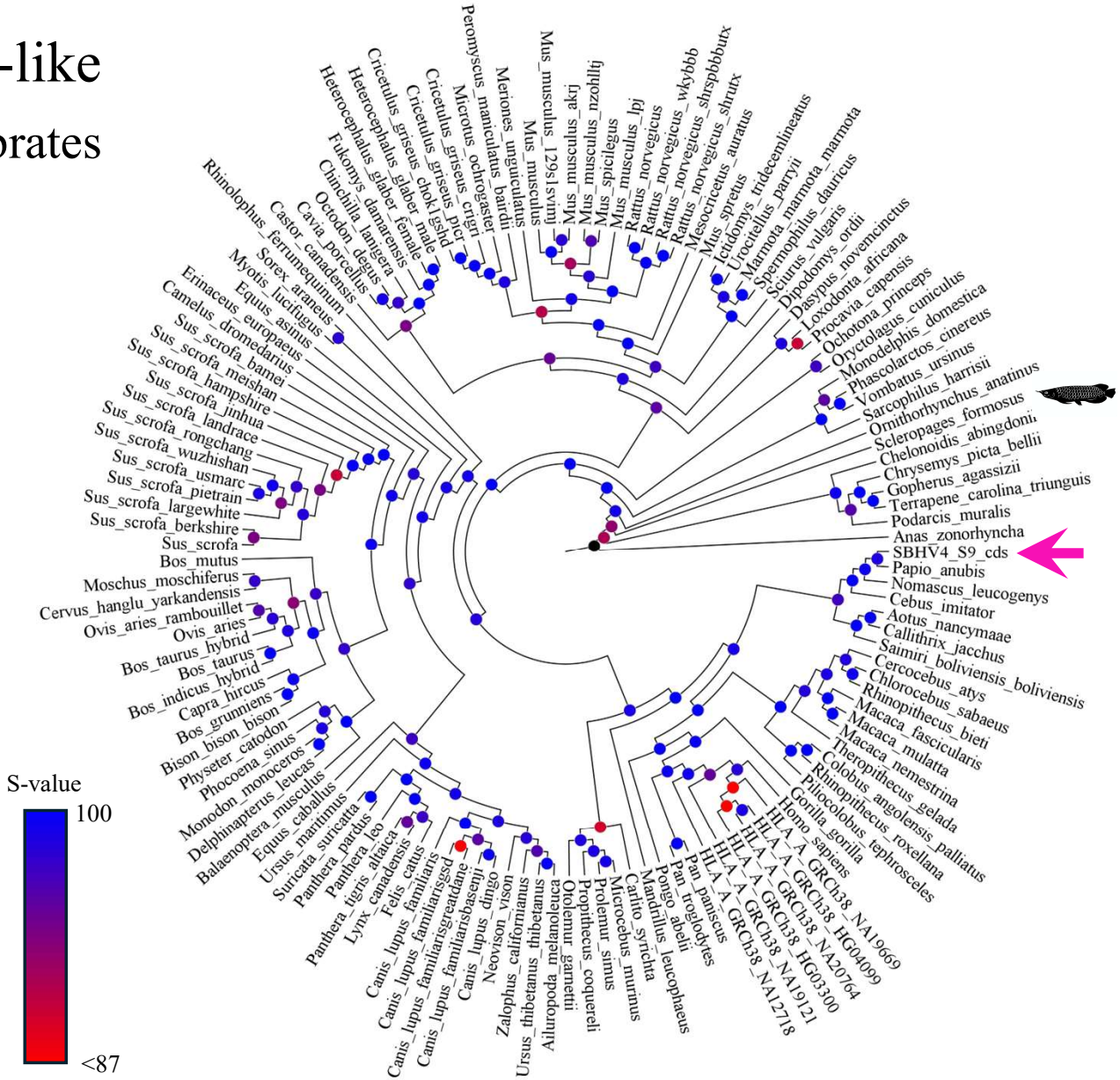
All data utilized in this study are presented within the Figures or as Supplemental Data.

Acknowledgements

This work was partially supported by Grants-in-Aid for Scientific Research from the JSPS (24K21913 to E.H.). Computations were performed on the NIG supercomputer at ROIS National Institute of Genetics, Japan.

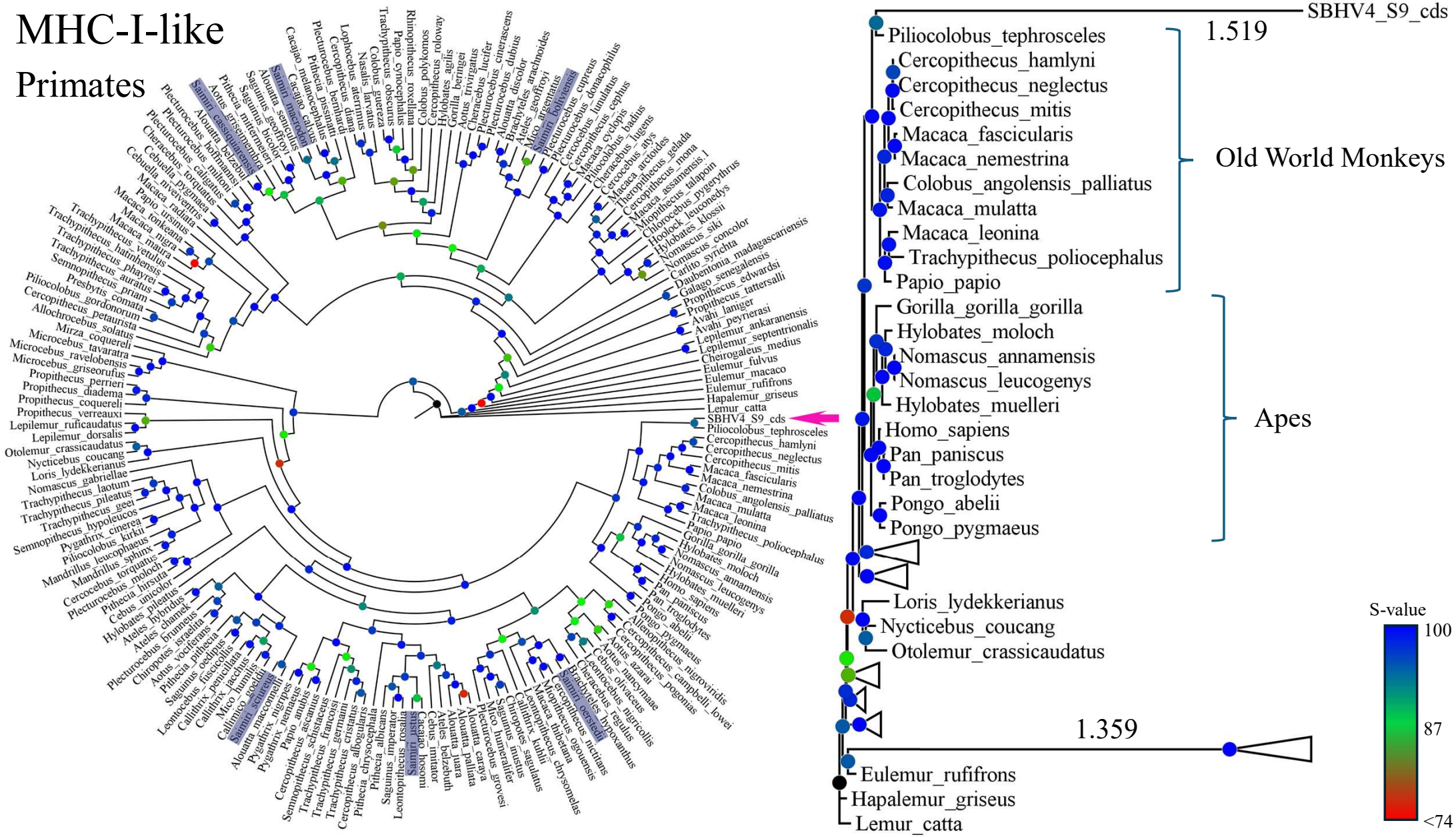
MHC-I-like

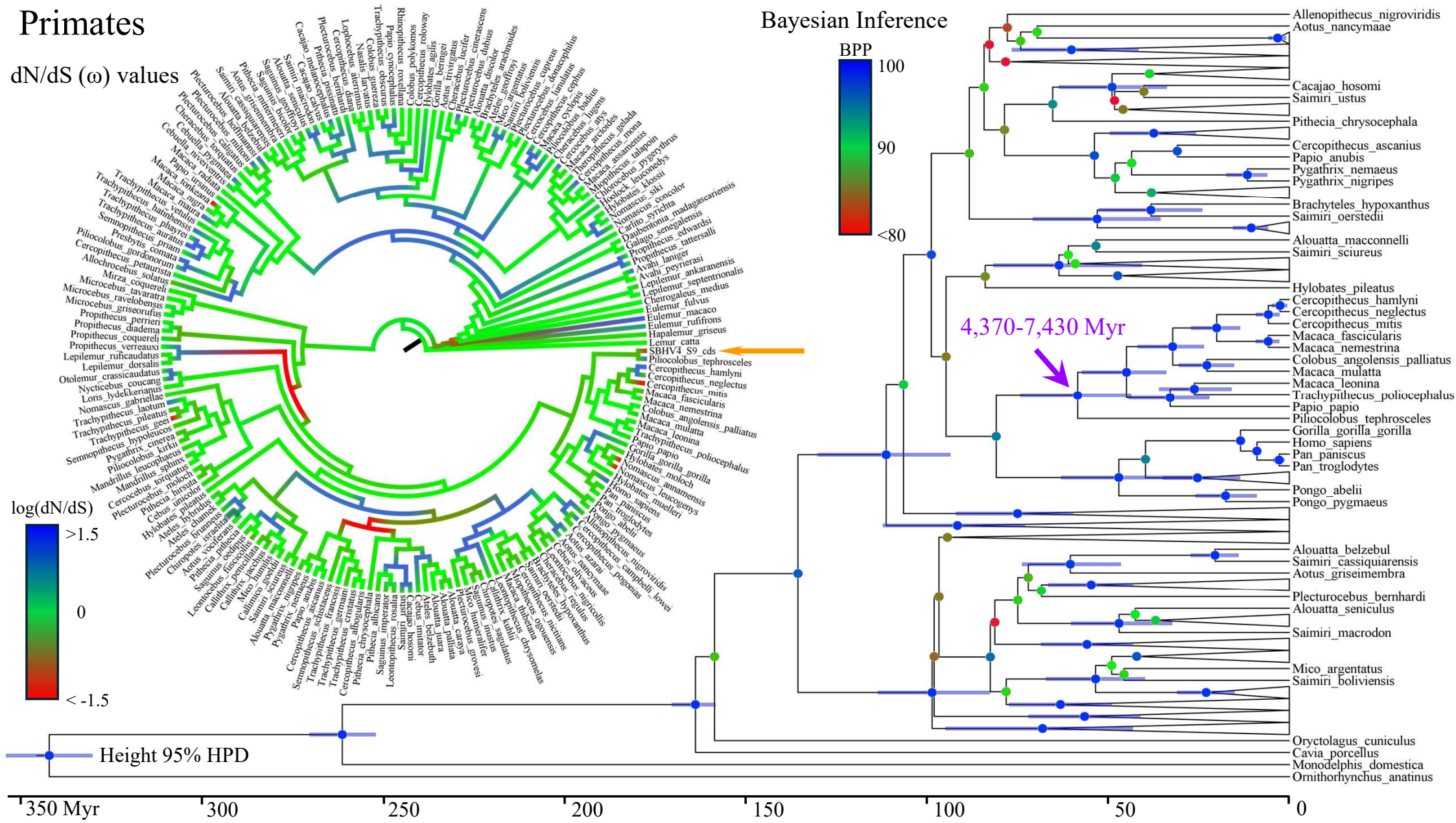
Vertebrates



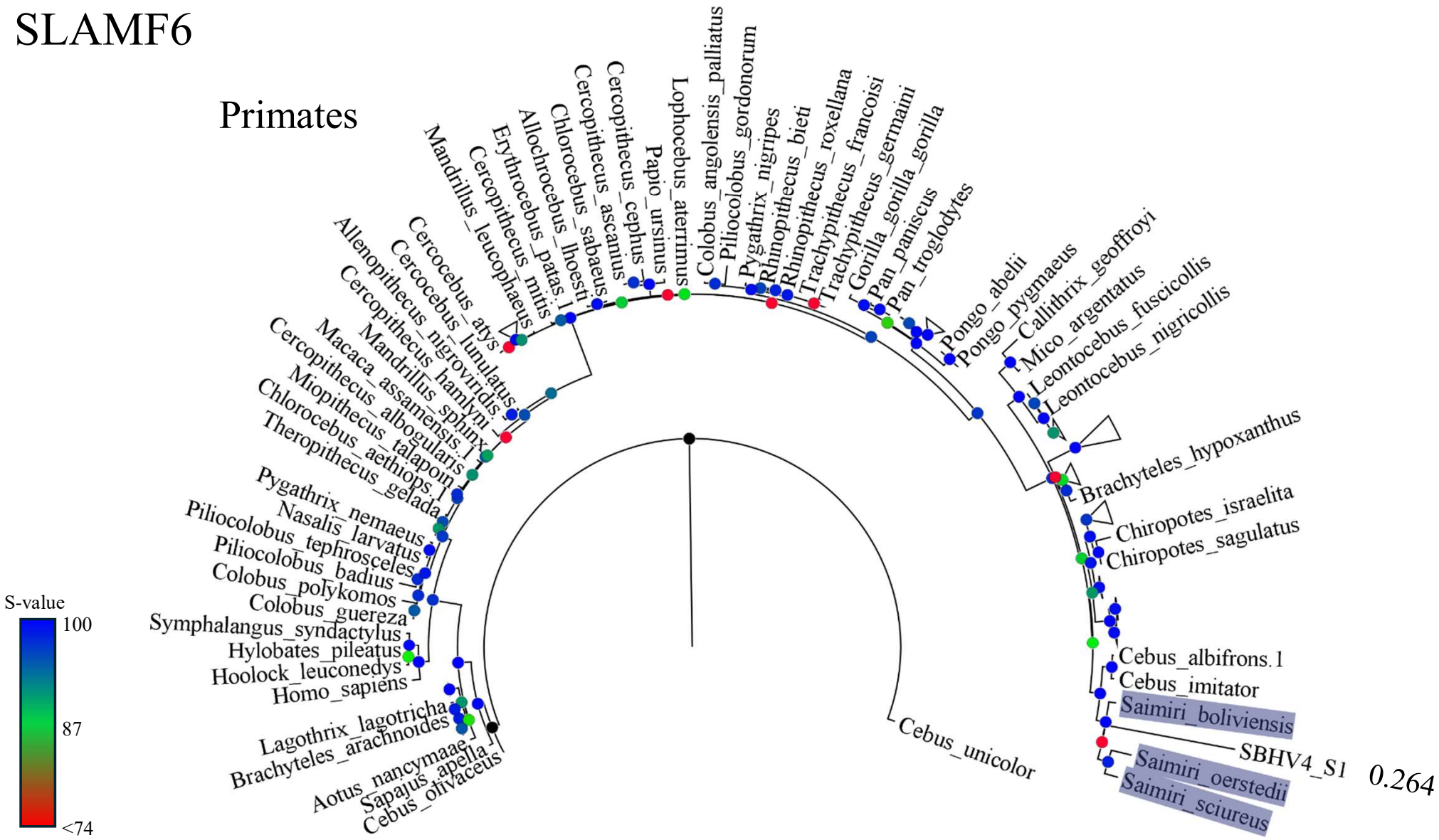
MHC-I-like

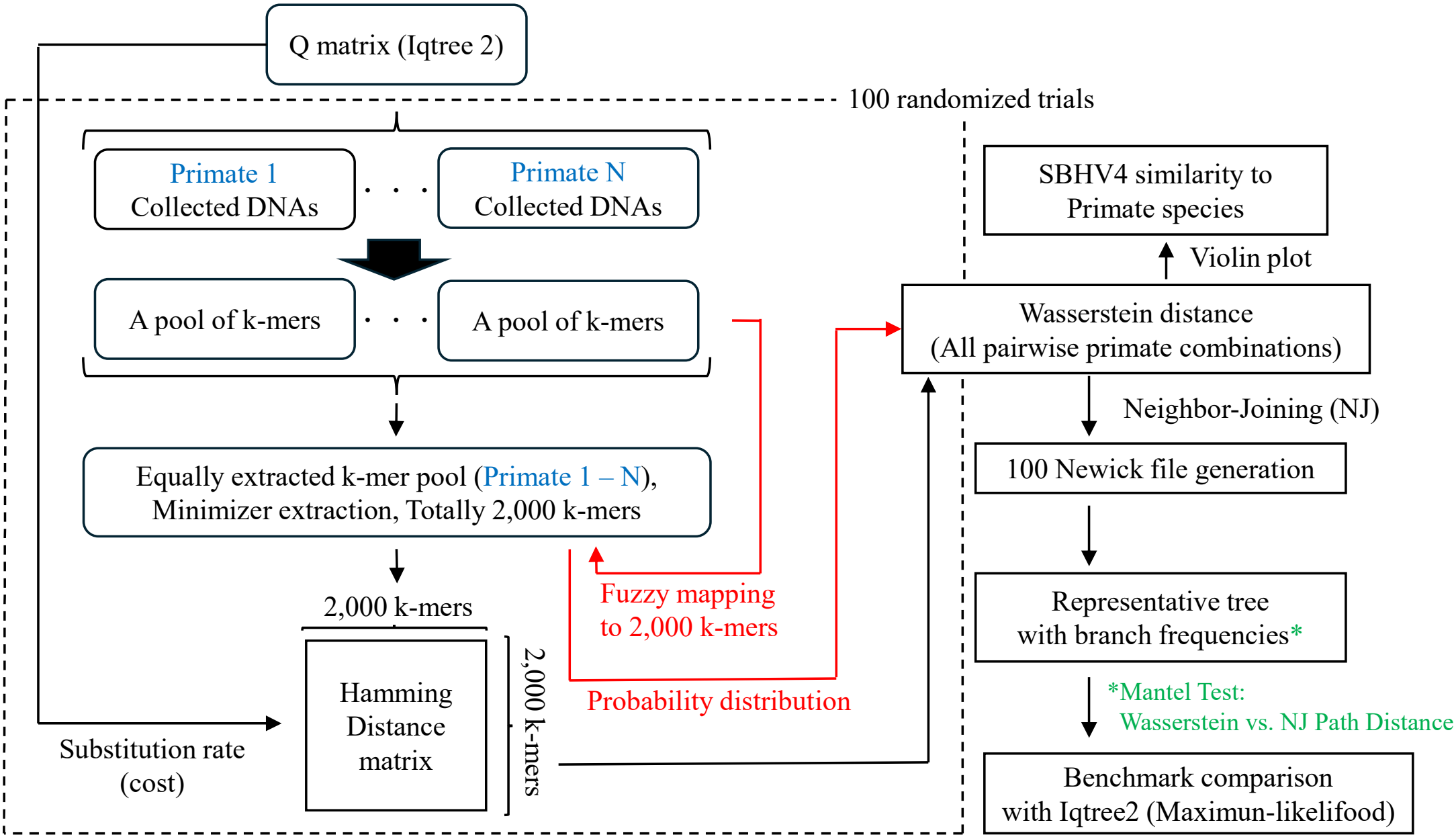
Primates

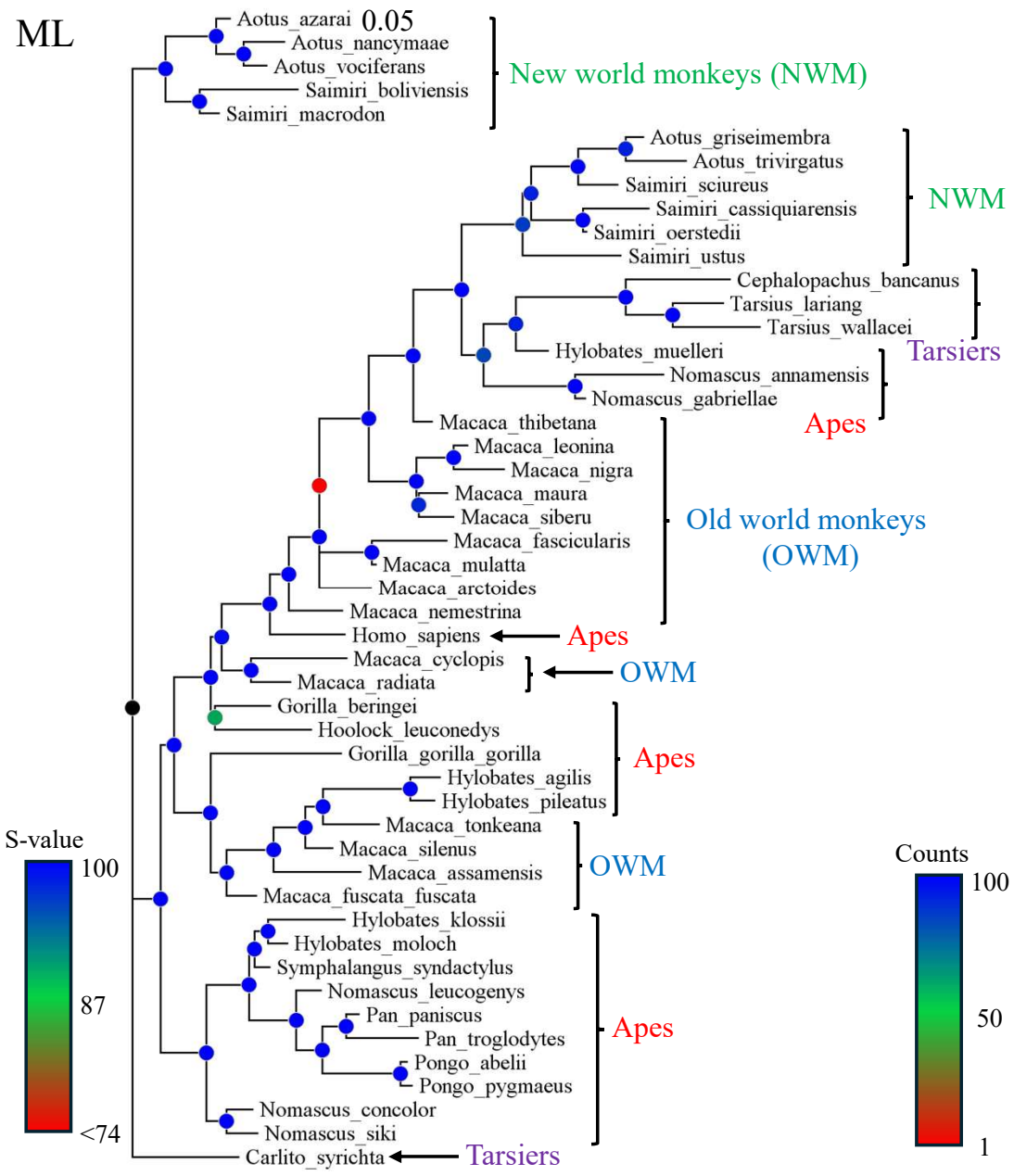


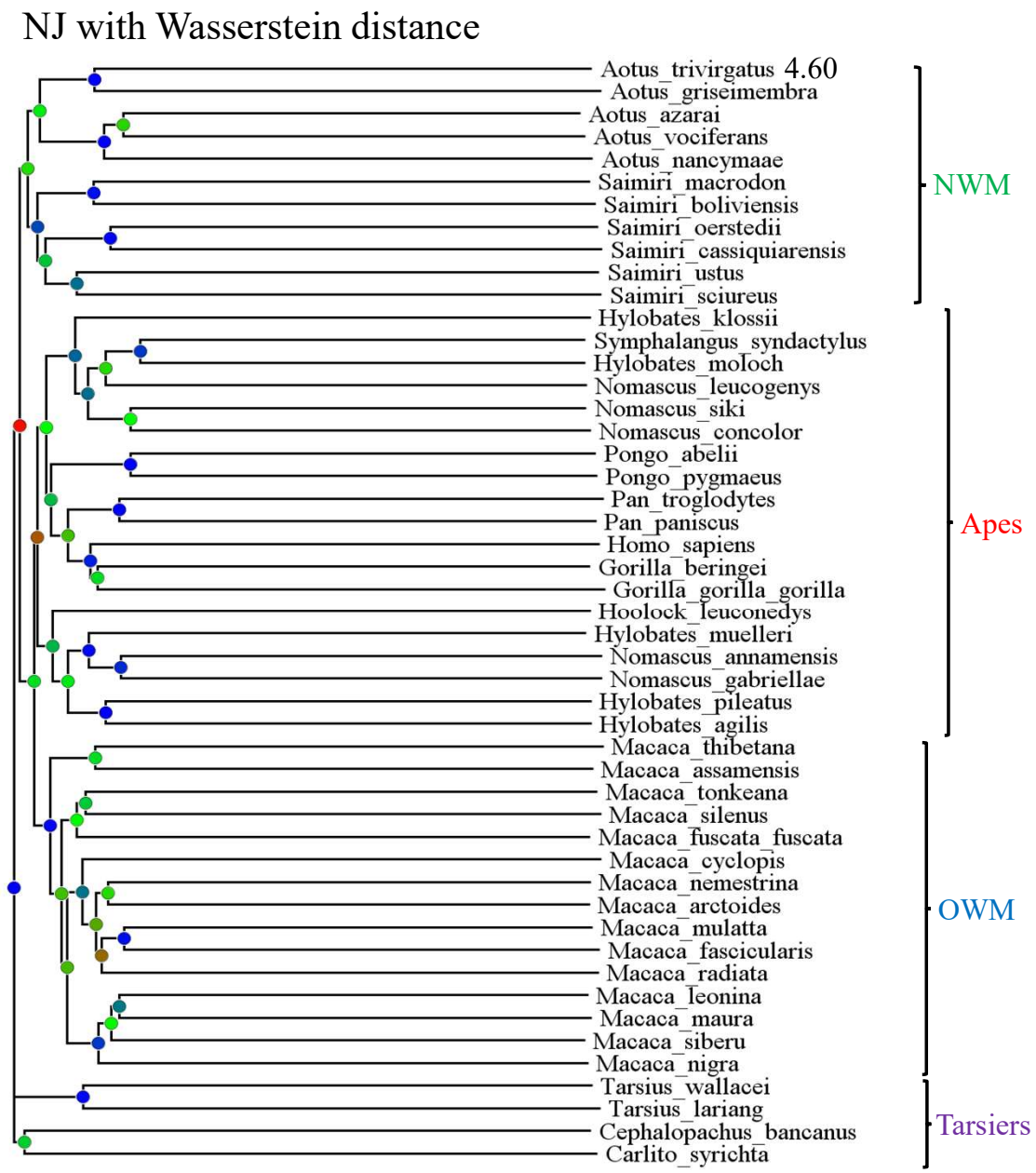

Primates
dN/dS (ω) values
log(dN/dS)
>1.5
0
< -1.5
Bayesian Inference
BPP
100
90
<80
4,370-7,430 Myr
Height 95% HPD
350 Myr
300
250
200
150
100
50
0
Allenopithecus_nigroviridis
Aotus_nancymaae
Cacajao_hosomi
Saimiri_ustus
Pithecia_chrysocephala
Cercopithecus_ascanius
Papio_anubis
Pygathrix_nemaeus
Pygathrix_nigripes
Brachyteles_hypoxanthus
Saimiri_oerstedii
Alouatta_macconnelli
Saimiri_sciureus
Hylobates_pileatus
Cercopithecus_hamlyni
Cercopithecus_neglectus
Cercopithecus_mitis
Macaca_fascicularis
Macaca_nemestrina
Colobus_angolensis_palliatus
Macaca_mulatta
Macaca_leonina
Trachypithecus_poliocephalus
Papio_papio
Piliocolobus_tephrosceles
Gorilla_gorilla_gorilla
Homo_sapiens
Pan_paniscus
Pan_troglodytes
Pongo_abelii
Pongo_pygmaeus
Alouatta_belzebul
Saimiri_cassiquiarensis
Aotus_griseimembra
Plecturocebus_bernhardi
Alouatta_seniculus
Saimiri_macrodon
Mico_argentatus
Saimiri_boliviensis
Oryctolagus_cuniculus
Cavia_porcellus
Monodelphis_domestica
Ornithorhynchus_anatinus

SLAMF6

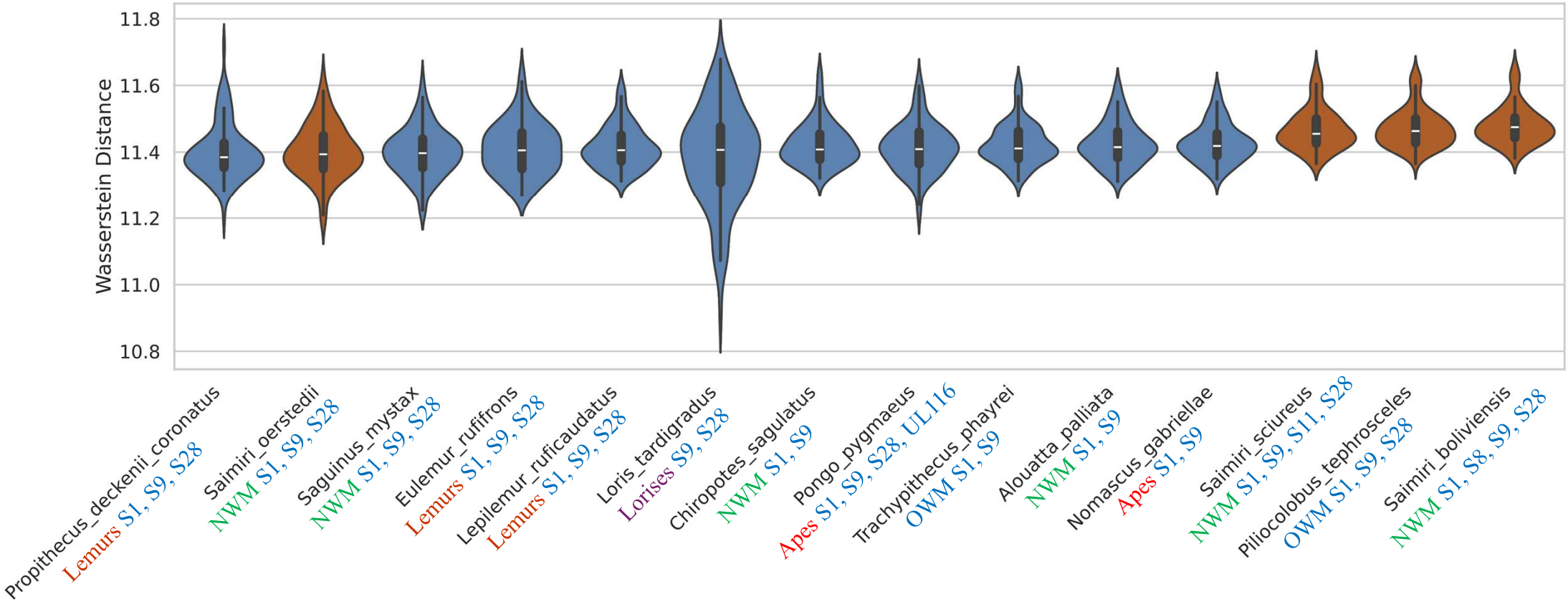

Primates
Colobus_angolensis_palliatus
Piliocolobus_gordonorum
Pygathrix_nigripes
Rhinopithecus_bieti
Rhinopithecus_roxellana
Trachypithecus_francoisi
Trachypithecus_germaini
Gorilla_gorilla_gorilla
Pan_paniscus
Pan_troglodytes
Pongo_abelii
Pongo_pygmaeus
Callithrix_geoffroyi
Mico_argentatus
Leontocebus_fuscicollis
Leontocebus_nigricollis
Brachyteles_hypoxanthus
Chiropotes_israelita
Chiropotes_sagulatus
Cebus_albifrons.1
Cebus_imitator
Saimiri_boliviensis
SBHV4_S1
Saimiri_oerstedii
Saimiri_sciureus
0.264
Cebus_unicolor
Lophocebus_aterrimus
Papio_ursinus.1
Cercopithecus_cephus
Cercopithecus_ascanius
Chlorocebus_sabaeus
Allochrocebus_lhoesti
Erythrocebus_patas.1
Cercopithecus_mitis
Mandrillus_leucophaeus
Cercocebus_atys
Cercocebus_lunulatus
Allenopithecus_nigroviridis
Cercopithecus_hamlyni
Mandrillus_sphinx
Macaca_assamensis.1
Cercopithecus_albogularis
Miopithecus_talapoin
Chlorocebus_aethiops.1
Theropithecus_gelada
Pygathrix_nemaeus
Nasalis_larvatus
Piliocolobus_tephrosceles
Piliocolobus_badius
Colobus_polykomos
Colobus_guereza
Symphalangus_syndactylus
Hylobates_pileatus
Hoolock_leuconedys
Homo_sapiens
Lagothrix_lagotricha
Brachyteles_arachnoides
Aotus_nancymaae
Sapajus_apella
Cebus_olivaceus
S-value
100
87
<74


Q matrix (Iqtree 2)
100 randomized trials
Primate 1
Collected DNAs
Primate N
Collected DNAs
A pool of k-mers
A pool of k-mers
Equally extracted k-mer pool (Primate 1 – N),
Minimizer extraction, Totally 2,000 k-mers
2,000 k-mers
Hamming
Distance
matrix
2,000 k-mers
Substitution rate
(cost)
Fuzzy mapping
to 2,000 k-mers
Probability distribution
SBHV4 similarity to
Primate species
Violin plot
Wasserstein distance
(All pairwise primate combinations)
Neighbor-Joining (NJ)
100 Newick file generation
Representative tree
with branch frequencies*
*Mantel Test:
Wasserstein vs. NJ Path Distance
Benchmark comparison
with Iqtree2 (Maximun-likelifood)


ML
0.05
Aotus_azarai
Aotus_nancymaae
Aotus_vociferans
Saimiri_boliviensis
Saimiri_macrodon
New world monkeys (NWM)
Aotus_griseimembra
Aotus_trivirgatus
Saimiri_sciureus
Saimiri_cassiquiarensis
Saimiri_oerstedii
Saimiri_ustus
NWM
Cephalopachus_bancanus
Tarsius_lariang
Tarsius_wallacei
Tarsiers
Hylobates_muelleri
Nomascus_annamensis
Nomascus_gabriellae
Apes
Macaca_thibetana
Macaca_leonina
Macaca_nigra
Macaca_maura
Macaca_siberu
Macaca_fascicularis
Macaca_mulatta
Macaca_arctoides
Macaca_nemestrina
Old world monkeys (OWM)
Homo_sapiens
Apes
Macaca_cyclopis
Macaca_radiata
OWM
Gorilla_beringei
Hoolock_leuconedys
Gorilla_gorilla_gorilla
Hylobates_agilis
Hylobates_pileatus
Apes
Macaca_tonkeana
Macaca_silenus
Macaca_assamensis
Macaca_fuscata_fuscata
OWM
Hylobates_klossii
Hylobates_moloch
Symphalangus_syndactylus
Nomascus_leucogenys
Pan_paniscus
Pan_troglodytes
Pongo_abelii
Pongo_pygmaeus
Apes
Nomascus_concolor
Nomascus_siki
Carlito_syrichta
Tarsiers
S-value
100
87
<74
Counts
100
50
1

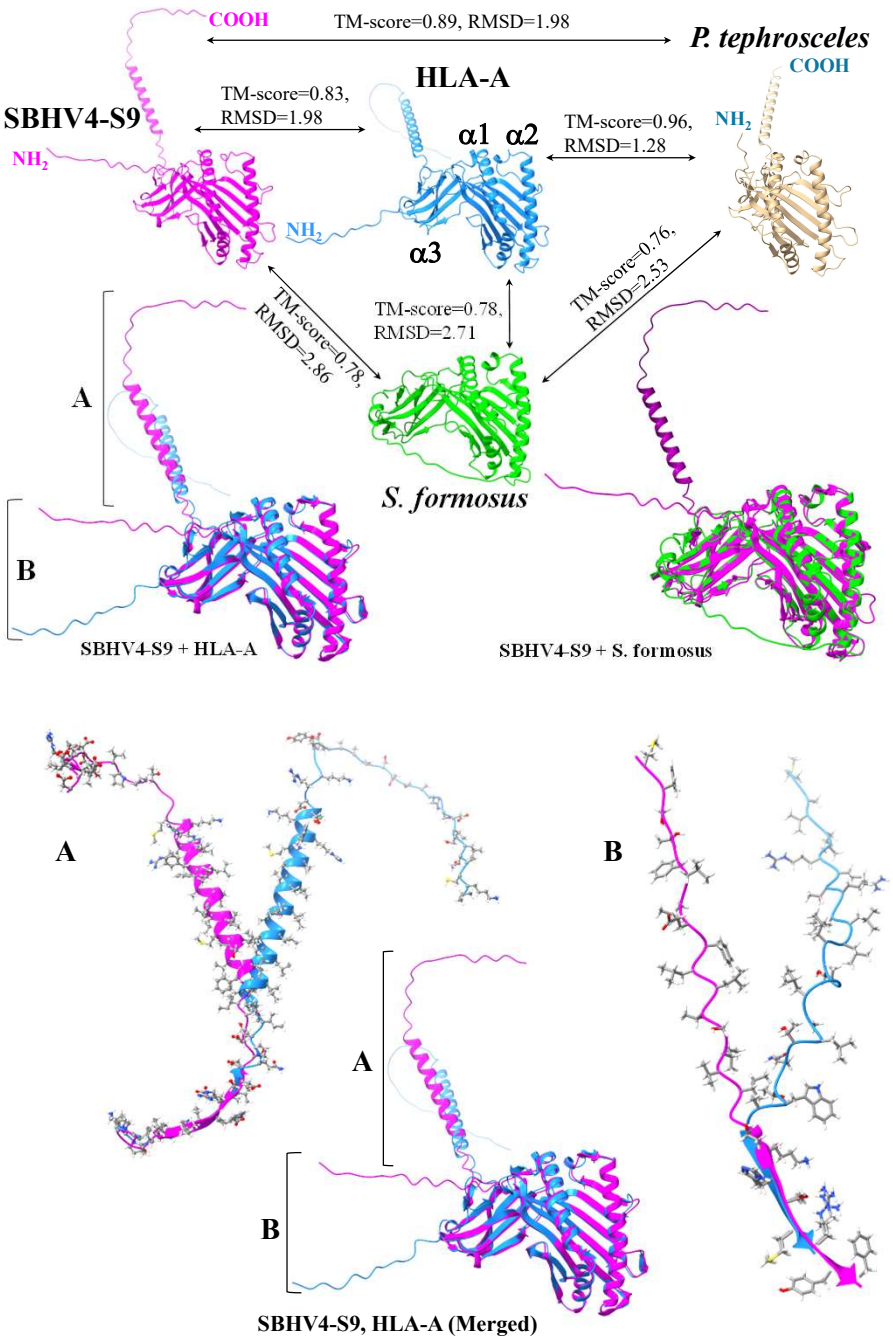

NJ with Wasserstein distance
4.60
Aotus_trivirgatus
Aotus_griseimembra
Aotus_azarai
Aotus_vociferans
Aotus_nancymaae
Saimiri_macrodon
Saimiri_boliviensis
Saimiri_oerstedii
Saimiri_cassiquiarensis
Saimiri_ustus
Saimiri_sciureus
NWM
Hylobates_klossii
Symphalangus_syndactylus
Hylobates_moloch
Nomascus_leucogenys
Nomascus_siki
Nomascus_concolor
Pongo_abelii
Pongo_pygmaeus
Pan_troglodytes
Pan_paniscus
Homo_sapiens
Gorilla_beringei
Gorilla_gorilla_gorilla
Hoolock_leuconedys
Hylobates_muelleri
Nomascus_annamensis
Nomascus_gabriellae
Hylobates_pileatus
Hylobates_agilis
Apes
Macaca_thibetana
Macaca_assamensis
Macaca_tonkeana
Macaca_silenus
Macaca_fuscata_fuscata
Macaca_cyclopis
Macaca_nemestrina
Macaca_arctoides
Macaca_mulatta
Macaca_fascicularis
Macaca_radiata
Macaca_leonina
Macaca_maura
Macaca_siberu
Macaca_nigra
OWM
Tarsius_wallacei
Tarsius_lariang
Cephalopachus_bancanus
Carlito_syrichta
Tarsiers

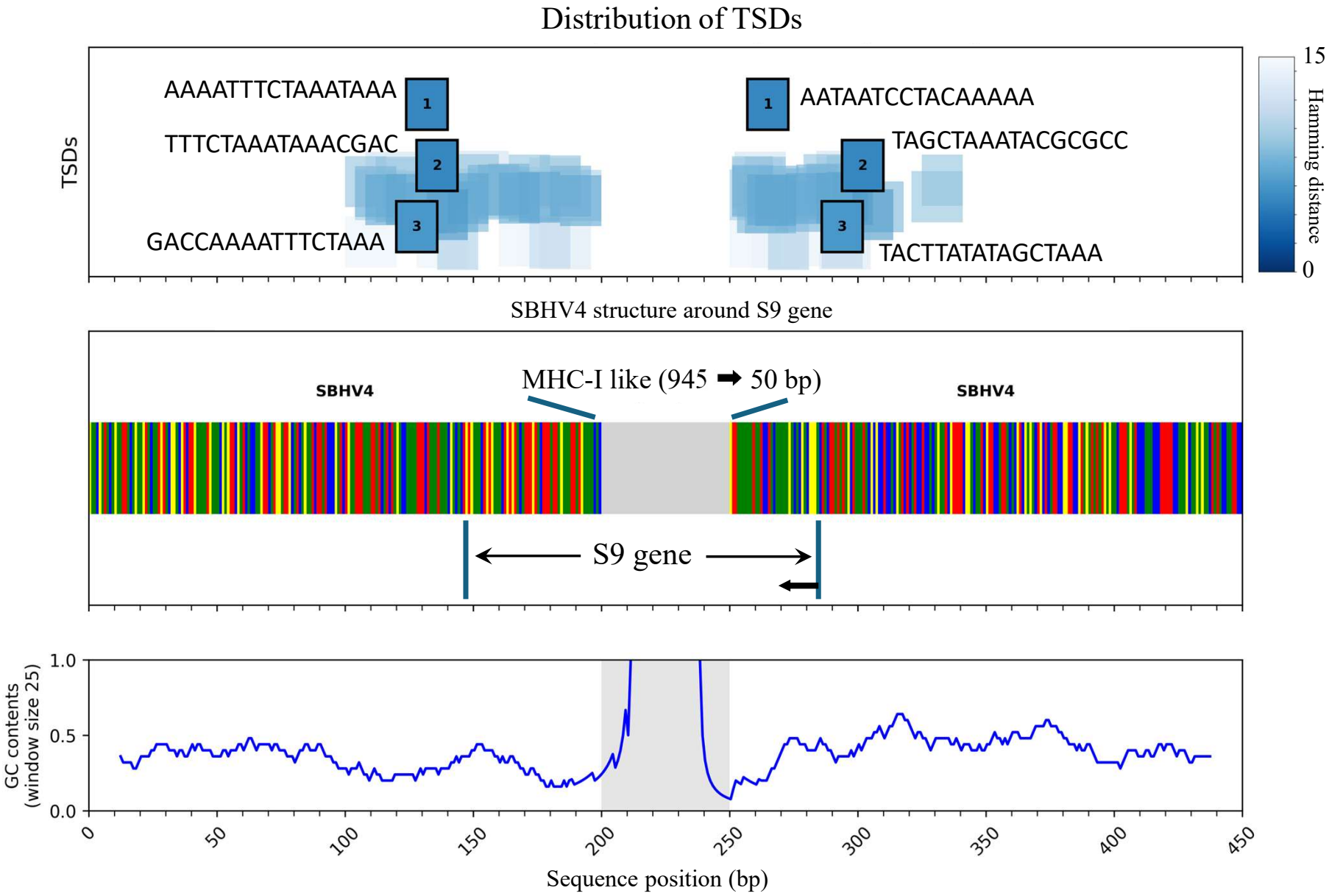
Wasserstein distances to SBHV4



Dunn's post-hoc test results

	Piliocolobus_tephrosceles	Saimiri_boliviensis	Saimiri_oerstedii	Saimiri_sciureus
Piliocolobus_tephrosceles	N/A	1.00e+00	**8.36e-10**	1.00e+00
Saimiri_boliviensis	1.00e+00	N/A	**1.24e-13**	5.34e-01
Saimiri_oerstedii	**8.36e-10**	**1.24e-13**	N/A	**1.65e-08**
Saimiri_sciureus	1.00e+00	5.33e-01	**1.65e-08**	N/A


COOH
TM-score=0.89, RMSD=1.98
P. tephrosceles
HLA-A
COOH
SBHV4-S9
TM-score=0.83,
RMSD=1.98
α1 α2
TM-score=0.96,
RMSD=1.28
NH2
NH2
NH2
α3
TM-score=0.76,
RMSD=2.53
TM-score=0.78,
RMSD=2.86
TM-score=0.78,
RMSD=2.71
A
S. formosus
B
SBHV4-S9 + HLA-A
SBHV4-S9 + S. formosus
A
B
A
B
SBHV4-S9, HLA-A (Merged)


Distribution of TSDs
TSDs
AAAATTTCTAAATAAA
TTTCTAAATAAACGAC
GACCAAAATTTCTAAA
AATAATCCTACAAAAA
TAGCTAAATACGCGCC
TACTTATATAGCTAAA
1
2
3
15
Hamming distance
0
SBHV4 structure around S9 gene
SBHV4
MHC-I like (945 ➡ 50 bp)
SBHV4
S9 gene
GC contents
(window size 25)
1.0
0.5
0.0
0
50
100
150
200
250
300
350
400
450
Sequence position (bp)