12-2023

FIRST GENOMIC RESOURCE FOR AN ENDANGERED NEOTROPICAL MEGA-HERBIVORE: THE COMPLETE MITOCHONDRIAL GENOME OF THE FOREST-DWELLER (BAIRD'S) TAPIR (*TAPIRUS BAIRDII*)

Caroline C. Ennis

FIRST GENOMIC RESOURCE FOR AN ENDANGERED NEOTROPICAL MEGA-HERBIVORE: THE COMPLETE MITOCHONDRIAL GENOME OF THE FOREST-DWELLER (BAIRD'S) TAPIR (*TAPIRUS BAIRDII*)

Caroline C. Ennis

Completed: December 2021 (Published: June 2022, PeerJ)

Submitted for Departmental Honors in Biological Sciences Fall 2022

Advisor: J. Antonio Baeza

ABSTRACT

Baird's tapir, or the Central American Tapir *Tapirus bairdii* (family Tapiridae), is one of the largest mammals native to the forests and wetlands of southern North America and Central America and is categorized as `endangered' on the 2014 IUCN Red List of Threatened Species. This study reports, for the first time, the complete mitochondrial genome of *T. bairdii* and examines the phylogenetic position of *T. bairdii* amongst closely related species in the same family and order to which it belongs using mitochondrial protein-coding genes (PCG's). The circular, double-stranded, A-T rich mitochondrial genome of T. bairdii is 16,697 bp in length consisting of 13 protein coding genes (PCG's), two ribosomal RNA genes (rrnS (12s ribosomal RNA and rrnL (16s ribosomal RNA)), and 22 transfer RNA (tRNA) genes. A 33 bp long region was identified to be the origin of replication for the light strand (OL), and a 1,247 bp long control region (CR) contains the origin of replication for the heavy strand (OH). A majority of the PCG's and tRNA genes are encoded on the positive, or heavy, strand. The gene order in *T. bairdii* is identical to that of *T. indicus* and *T. terrestris,* the only two other species of extant tapirs with assembled mitochondrial genomes. An analysis of Ka/Ks ratios for all the PCG's show values <1, suggesting that all these PCGs experience strong purifying selection. A maximum-likelihood phylogenetic analysis supports the monophyly of the genus Tapirus and the order Perissodactyla. The complete annotation and analysis of the mitochondrial genome of *T. bairdii* will contribute to a better understanding of the population genomic diversity and structure of this species, and it will assist in the conservation and protection of its dwindling populations.

TABLE OF CONTENTS

LIST OF TABLES AND FIGURES

INTRODUCTION

In the family Tapiridae, Baird's tapir (*Tapirus bairdii*), also known as the Central American or Mesoamerica tapir, is the largest mammal native to southern North America and Central America (**Fig. 1**) and is one of four extant species within the genus *Tapirus* among odd-toed ungulates (order Perissodactyla). *Tapirus bairdii* can be found in forests and wetlands ranging from southern Mexico to Colombia [1]. They play an important role as seed dispersers in tropical forests, including areas highly threatened by disturbance [2], and represent a food source for rural-dwelling people [3]. Baird's tapir browses the forest in areas with nearby freshwater bodies, high tree and shrub diversity that provide good food quality, low predation (including hunting) pressure, and limited human presence [4]. Humans remain the primary predator of Baird's Tapir, contributing to the decline in population over the last four decades [5]. This mega-herbivore is classified as 'endangered' on the 2014 IUCN Red List of Threatened Species, with estimates suggesting approximately 4,500 individuals remaining in the field. Furthermore, populations of *T. bardii* are dwindling due to anthropogenic activities causing habitat loss and fragmentation that include, among others, urban development, pollution, over hunting, as well as local and global climate change [6].

Despite their endangered status, very few genetic and genomic resources exist for *Tapirus* in general, and a complete mitochondrial genome assembly and annotation does not exist for Baird's tapir. An early study, using a fragment of the mitochondrial Cytochrome Oxidase subunit II (*cox2*) to examine the phylogenetic relationships among representatives of the genus *Tapirus*, suggested the divergence of three distinct *Tapirus* lineages (South American, Central American, and Asian) occurred 20–30 million years ago [7]. Two more recent studies that used DNA microsatellite markers in wild and captive *T. bairdii* populations revealed that *T. bairdii* was at increased risk of losing genetic variability due to inbreeding [8,9]. Lastly, a recent study focusing on the phylogeography of *T. bairdii* and *T. pinchaque* reported a close phylogenetic relationship between *T. bairdii* and fossil tapir species from North America, supporting previously established phylogenetic relationships based on morphometric data [10,11]. Interestingly though, the molecular study of Ruiz-García et al. (2012) [11] contradicts an earlier

morphometric study by Hulbert (1995) [12] that indicated a close relationship between *T. bairdii* and the South American tapir *T. terrestris*. A fifth species within *Tapirus* has been proposed and further genetic analysis may contribute to its confirmation [13]. Overall, the phylogenetic position of *T. bairdi* is still not completely resolved and might evolve as more phylogenetic data become available.

This study, for the first time, generated a genomic resource, for this species. Specifically, we focused on assembling and characterizing in detail the mitochondrial genome of *T. bairdii*. The information we generated was used to explore the phylogenetic position of *T. bairdii* among closely related species based on the protein coding genes (PCGs). Among others, we have analyzed nucleotide composition, codon usage, and selective constraints in PCG's. Also, the secondary structure of tRNA genes and the control region was investigated in detail. Lastly, the phylogenetic position of *T. bairdii* amongst members of the Perissodactyla was examined using PCGs. The complete and detailed characterization of the mitochondrial genome is a stride towards improving understanding of the evolutionary relationships of *Tapirus bairdii*.

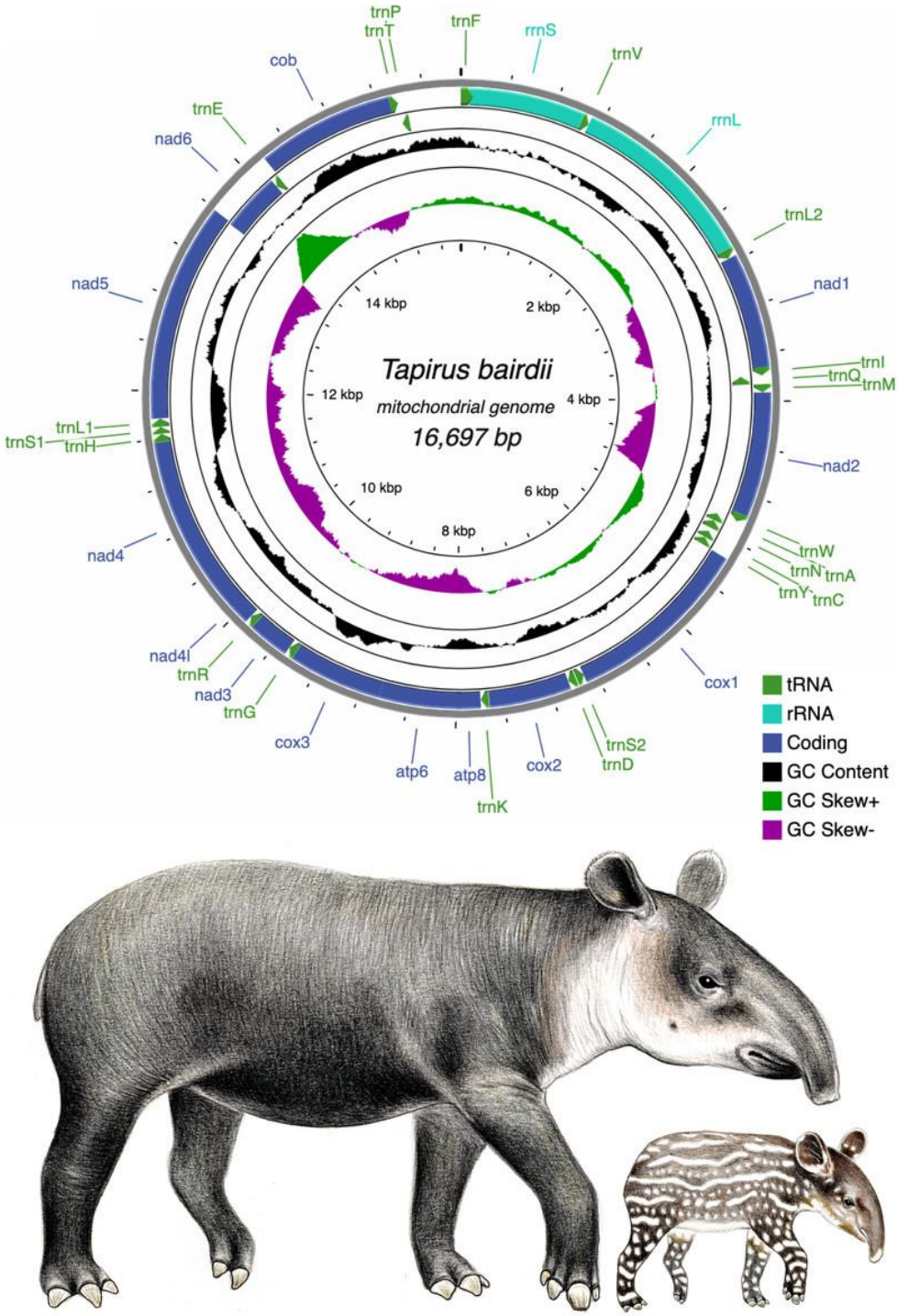


Figure 1: Circular genome map of the mitochondrial DNA of *Tapirus bairdii*. The mitochondrial genome is comprised of 13 protein coding genes (PCGs), 22 transfer RNA genes (tRNA), and 2 ribosomal RNA genes (rRNA). The inner circles show the GC content and GC skew along the sequence. Illustrations of *Tapirus bairdii* copyright 1990 Stephen D. Nash. Used with permission.

MATERIALS AND METHODS

Collection, tissue sampling, and DNA genomic sequencing

We requested a blood sample fixed in ethanol (95%) from a specimen of *T. bardii* exhibited at Parque Xcaret, Playa del Carmen, Quintana Roo, Mexico. This sample was transported to the Laboratorio de Bioconservación y Manejo, Instituto Politécnico Nacional, Cuidad de Mexico, Mexico, for its posterior laboratory treatment. The Qiagen Blood and Tissue Kit (Qiagen, Hilden, Germany) was used in extraction of total genomic DNA from the tissue sample according to manufacturer's instructions [14]. Following extraction, the Savannah River Ecology Laboratory at the University of Georgia, Aiken, performed next generation sequencing from the extracted DNA sample [14].

An Illumina HiSeq sequencer (Illumina, San Diego, CA, USA) that used 2 × 200 cycle sequenced Illumina paired-end (PE) shotgun library [14]. The PE reads were prepared using standard protocol of the Nextera™ DNA Sample Prep Kit (Epicentre®). The pairs generated totaled 5,507,122 and were provided in FASTQ format by the facility [14]. DNA-seq data have been deposited in the NCBI Sequence Read Archive (SRA) under Bioproject ID: PRJNA785336, Biosample accession: SAMN23553527, and SRA accession: SRR17086167. Methods of extraction and sequencing the mitochondrial DNA followed standard protocol and has been utilized previously for other organisms [14, 15].

Mitochondrial genome assembly

The mitochondrial genome of *T. bardii* was *de novo*-assembled using the pipeline GetOrganelle v1.6.4 [16]. The mitochondrial genome of the congeneric *T. terrestris*, available in GenBank (KJ417810), was used as a reference. The run used k-mer sizes of 21, 55, 85, and 115.

Annotation and analysis of the assembled mitochondrial genome

The newly assembled mitochondrial genome of *T. bairdii* was first annotated using the MITOS and MITOS2 web servers (http://mitos2.bioinf.uni-leipzig.de) [17] with the vertebrate genetic code (code 2). Corrections to the start and stop codons were made using the server ExPASy (https://web.expasy.org) and the MEGAX software [18]. Visualization was performed using the CGView Server (beta) (http://cgview.ca/) [19] using the manually corrected annotation.

The codon usage and open reading frames of the protein-coding genes (PCG) were analyzed. The codon usage of the 13 PCG's was predicted using the Codon Usage web server (https://www.bioinformatics.org/sms2) [20] and visualized using the EZcodon tool in the web server EZmito (http://ezmito.unisi.it/ezcodon) [21], both set to the vertebrate genetic code. tRNA genes were identified by MITOS and MITOS2, and the secondary structures were visualized in the Forna web server (http://rna.tbi.univie.ac.at/forna/) [22].

Selective pressure on the PCG's was examined by comparing rates of nonsynonymous and synonymous substitutions. The values of K_A (number of nonsynonymous substitutions per nonsynonymous site: $K_A = d_N = S_A/L_A$), K_S (number of synonymous substitutions per synonymous site: $K_S = d_S = S_S/L_S$), and ω (K_A/K_S) were found for each PCG using the KaKs_calculator 2.0 [23-25]. The estimated ω values were based on a comparison between *T. bairdii* and its closely related species, *T. indicus* (KJ417810). The γ-MYN model was used in order to account for variable mutation rates across sites within each PCG sequence. When $\omega = 1$, the PCG's are assumed to be under neutral selection, for $\omega > 1$ positive (diversifying) selection is assumed, and $\omega < 1$ indicates negative selective constraints (purifying selection).

The long non-coding region, understood to be the control region, was analyzed. Repeats within the region were found using the Tandem Repeat Finder Version 4.09 web server (https://tandem.bu.edu/trf/trf.basic.submit.html) [26] and the BioPHP Microsatellite Repeats Finder web server (http://insilico.ehu.es/mini_tools/microsatellites/) [27]. Manual comparison to known annotations of mammalian CR consensus sequences revealed conserved domains and blocks. The RNAstructure web server (https://rna.urmc.rochester.edu/RNAstructureWeb) provided predictions of secondary structures based on the lowest free energy [28]. A short non-coding region, understood to be the origin of replication of the light strand (O_L), was also analyzed in the RNA-structure web server, with an attention to the presence of stem and loop structures.

Phylogenetic position of Tapirus bairdii

Using PCGs, we examined the phylogenetic position of *T. bairdii* among other representatives of the order Perissodactyla and superfamily Tapiroidea. The newly assembled mitochondrial genome along

with the mitochondrial genomes of 61 other specimens belonging to the Perissodactyla available in the GenBank database were used for the phylogenetic analysis conducted using the MitoPhAST V02 pipeline [29]. Outgroups included three species belonging to the Artiodactyla (*Lama guanicoe*, *Vicugna vicugna*, and *Camelus ferus* [Fam. Camelidae]). MitoPhAST first extracted all 13 PCG nucleotide sequences from the species available in GenBank as well as from *T. bairdii*. Clustal Omega aligned the PCG nucleotide sequences after translation into amino acid sequences [14, 30]. Poorly aligned regions were removed with trimAl [31] before the dataset was partitioned and the best fitting models of sequence evolution were selected with ProtTest [32]. Finally, the concatenated and partitioned PCG amino acid alignments were used to perform a maximum likelihood phylogenetic tree search in the software IQ-TREE [33]. The robustness of the ML tree topology was ascertained by 1,000 bootstrap pseudoreplicates of the tree search.

RESULTS AND DISCUSSION

The pipeline GetOrganelle assembled the complete mitochondrial chromosome of *T. bairdii* with an average coverage of 562x (sequence available at GenBank (OM935749)). The full mitochondrial genome of *T. bairdii* is 16,697 bp in length and contains 13 protein-coding genes (PCG's), two ribosomal RNA genes (*rrnS* (12S ribosomal RNA) and *rrnL* (16S ribosomal RNA)), and 22 transfer RNA (tRNA) genes. All of the PCG's are encoded on the heavy (H) or positive strand, excluding *nad6* which is found on the light (L) strand. Both ribosomal RNA genes and fifteen of the tRNA genes are also encoded on the L strand (**Fig. 1**, **Table 1**). The gene order, and distribution on each strand, is identical to that reported in the congeneric *Tapirus indicus* [34], as well as in the rhinoceros *Diceros bicornis* [35] and horses *Equus kiang* [36] and *Equus caballus* [37], all members of Perissodactyla. Mitochondrial genome arrangements in mammals tend to remain stable, therefore the identical mitochondrial genome arrangement of *Tapirus* to members of Perissodactyla is expected and exhibits the typical vertebrate arrangement [38].

Table 1. Mitochondrial genome of Tapirus bairdii. Arrangement and annotation.

Name	Type	Start	Stop	Strand	Size (bp)	Start Codon	Stop Codon	Anti-codon	Continuity
trnF(gaa)	tRNA	1	68	+	68			GAA	0
rrnS	rRNA	69	1038	+	970				0
trnV(tac)	tRNA	1039	1105	+	67			TAC	0
rrnL	rRNA	1106	2684	+	1579				0

trnL2(taa)	tRNA	2685	2759	+	75			TAA	2
nad1	Coding	2762	3718	+	957	ATG	TAA		-1
trnI(gat)	tRNA	3718	3787	+	70			GAT	-3
trnQ(ttg)	tRNA	3785	3857	-	73			TTG	2
trnM(cat)	tRNA	3860	3928	+	69			CAT	0
nad2	Coding	3929	4970	+	1042	ATA	T		0
trnW(tca)	tRNA	4971	5040	+	70			TCA	5
trnA(tgc)	tRNA	5046	5114	-	69			TGC	1
trnN(gtt)	tRNA	5116	5188	-	73			GTT	2
O_L		5189	5221	+	33				-1
trnC(gca)	tRNA	5221	5286	-	66			GCA	0
trnY(gta)	tRNA	5287	5353	-	67			GTA	1
cox1	Coding	5355	6899	+	1545	ATG	TAA		-3
trnS2(tga)	tRNA	6897	6965	-	69			TGA	7
trnD(gtc)	tRNA	6973	7039	+	67			GTC	0
cox2	Coding	7040	7723	+	684	ATG	TAA		3
trnK(ttt)	tRNA	7727	7793	+	67			TTT	1
atp8	Coding	7795	7998	+	204	ATG	TAA		-43
atp6	Coding	7956	8636	+	681	ATG	TAA		-1
cox3	Coding	8636	9419	+	784	ATG	T		0
trnG(tcc)	tRNA	9420	9488	+	69			TCC	0
nad3	Coding	9489	9834	+	346	ATA	T		1
trnR(tcg)	tRNA	9836	9903	+	68			TCG	0
nad4l	Coding	9904	10200	+	297	ATG	TAA		-7
nad4	Coding	10194	11571	+	1378	ATG	T		0
trnH(gtg)	tRNA	11572	11640	+	69			GTG	0
trnS1(gct)	tRNA	11641	11699	+	59			TCT	0
trnL1(tag)	tRNA	11700	11769	+	70			TAG	0
nad5	Coding	11770	13590	+	1821	ATA	TAA		-17
nad6	Coding	13574	14101	-	528	ATG	TAA		0
trnE(ttc)	tRNA	14102	14170	-	69			TTC	5
cob	Coding	14176	15315	+	1140	ATG	AGA		0
trnT(tgt)	tRNA	15316	15384	+	69			TGT	0
trnP(tgg)	tRNA	15385	15450	-	66			TGG	0
CR		15451	16697	+	1247				0

The mitogenome is compact with intergenic spaces and overlaps primarily between 1 and 7 bp, with two relatively long gene junctions at *atp8-atp6* (overlap = 43 bp) and *nad5-nad6* (overlap = 17 bp). A 1,247 bp long non-coding region was assumed to be the D-loop/Control Region (CR).

All 13 PCG's in the mitogenome of *T. bairdii* exhibit the typical vertebrate mitochondrial start codons ATG (*n* = 10 PCGs) or ATA (*n* = 3 PCGs) (Table 1). Eight of the PCG's end with the typical termination codon of TAA. The *cob* gene terminates with AGA, which is also identified to be a conventional mtDNA stop codon in vertebrates, including other representatives of the Perissodactyla: *T. indicus* [34] and *Diceros bicornis* [35]. The other four genes (*nad2*, *cox3*, *nad3*, *nad4*) terminate with an incomplete stop codon T, as reported before in *T. indicus* [34] as well as in representatives of the sister

clade Rhinocerotidae [35]. Incomplete stop codons in mitochondrial genomes appear to be completed with A residues via post-transcriptional polyadenylation [39].

The PCGs in *T. bardii* exhibit an A+T bias with an overall base composition of *A* = 32.2%, *T* = 29.8%, *C* = 25.6%, and *G* = 12.3%. This bias is also exhibited by other members of the Perissodactyla, including *T. indicus* [34] and *Equus kiang* [36]. The most frequently used codons in the PCG's of *T. bairdii* are CTA (Leu, *N* = 286 times used, 7.52% of total), ATA (Met, *N* = 203 times used, 5.34% of total), ATT (Ile, *N* = 180 times used, 4.74% of total), ACA (Thr, *N* = 169 times used, 4.45% of total), and ATC (Ile, *N* = 165 times used, 4.34% of total). The least frequently used codons are AGA (End, *N* = 1 time used, 0.03% of total), CGG (Arg, *N* = 1 time used, 0.03% of total), TGG (Trp, *N* = 2 times used, 0.05% of total), CAG (Gln, *N* = 4 times used, 0.11% of total), and TCG (Ser, *N* = 4 times used, 0.11% of total) (**Fig. 2**).

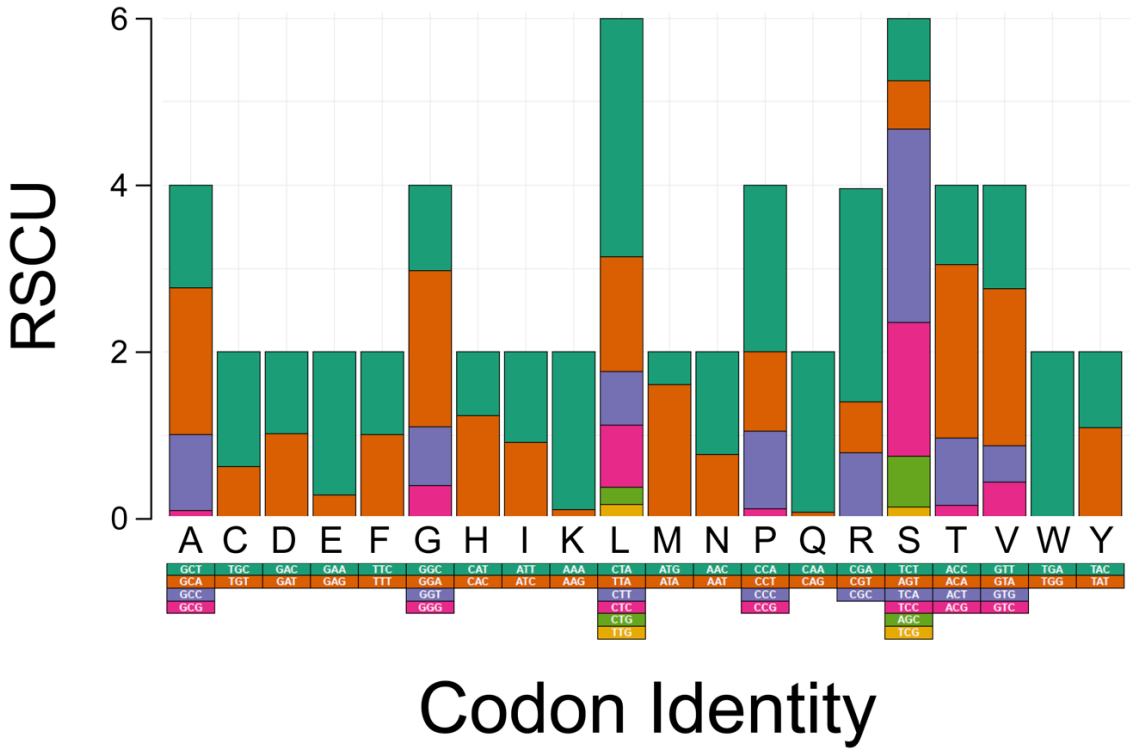


Figure 2: Codon usage analysis of the protein coding genes in the mitochondrial DNA of *Tapirus bairdii*.

All of the K_A/K_s ratios for all of the 13 the PCG's of *T. bairdii* show values <1, indicating all the PCG's are under purifying selection (**Fig. 3**). The *atp8* K_A/K_s ratio (K_A/K_s = 0.3791) is much greater than

that of the other 12 genes under purifying selection, suggesting a weaker constraint in the *atp8* gene. The K_A/K_s ratios observed for *nad4l*, *cox1*, *cox2*, *cox3* (0.0000, 0.0037, 0.0027, 0.0016, respectively) were much lower in comparison, suggesting stronger purifying selection and evolutionary constraints in the aforementioned genes. Selective pressure analyses in mitochondrial PCG's have not been conducted before in other species within the family Tapiridae and, in general, PCG selective pressures have been poorly studied in the Perissodactyla. A single previous study examining PCG's of the mitochondrial genome in members of the genus *Equus* also found an overall pattern of purifying selection [40].

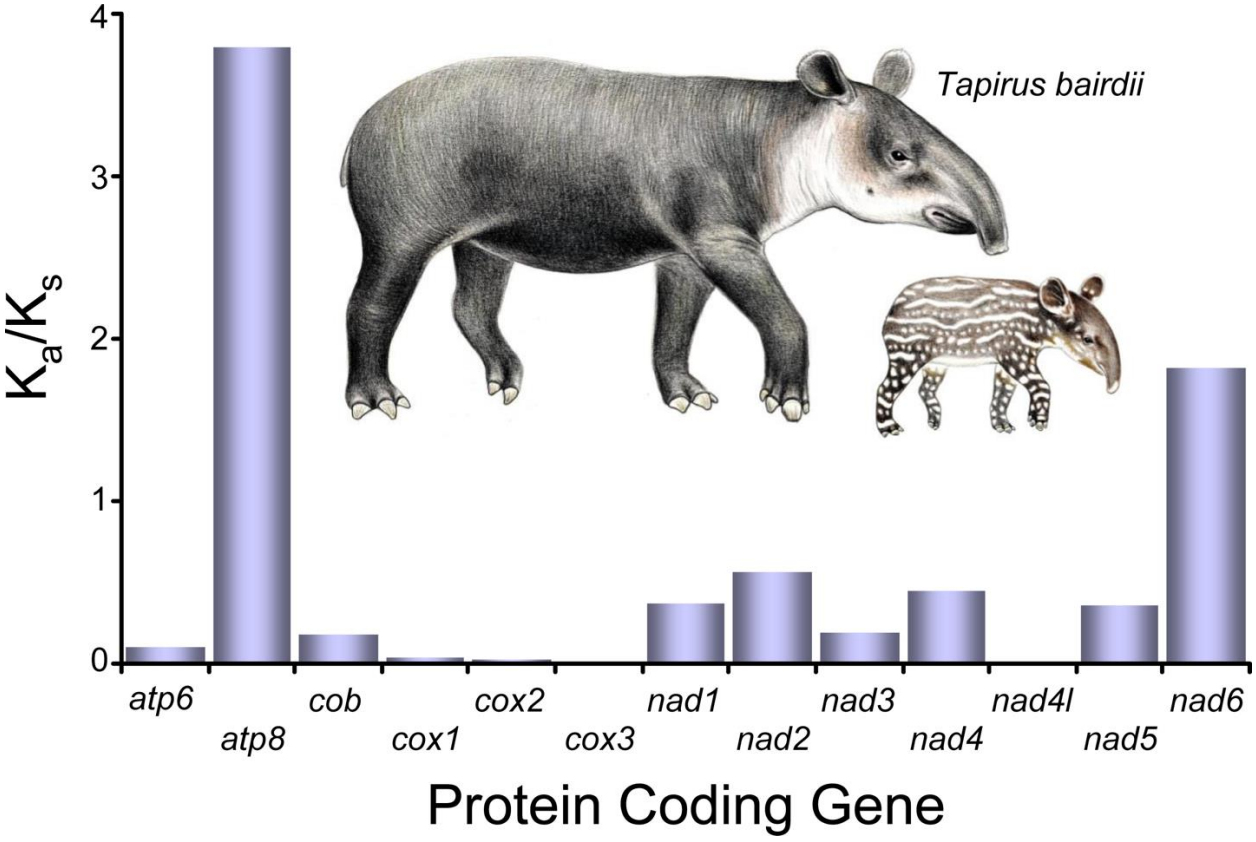


Figure 3: Selective pressure analysis in the PCG's of *Tapirus bairdii*. K_A/K_S ($x10^{-1}$) values were calculated using the γ-MYN model. Illustrations of *Tapirus bairdii* copyright 1990 Stephen D. Nash. Used with permission.

The tRNA genes in *T. bairdii* range from 59 to 75 bp in length and all, except *trnS1*, exhibit the typical 'cloverleaf' secondary structure (**Fig. 4**). The *trnS1* gene was predicted to be missing the D-arm by MITFI (implemented in the MITOS software), similar to that reported before for the same tRNA gene in other members of the Perissodactyla, *i.e.*, *Tapirus terrestris* [7] and *Rhinoceros unicornis* [41]. The loss of

stem-loop structures, specifically in *trnS1*, is a common occurrence in almost all metazoan mitochondrial genomes [42], so the missing D-arm and shortened length of *trnS1* in. *T. bardii* is not unanticipated. Aminoacylation and EF-Tu binding of D-arm-lacking tRNAs have been identified as factors that assist in translation [42]. The anticodon pattern is similar to that of members of Perissodactyla and that of the closely related Asian tapir [34], with the exception of *trnS1* codon which differs at the first position.

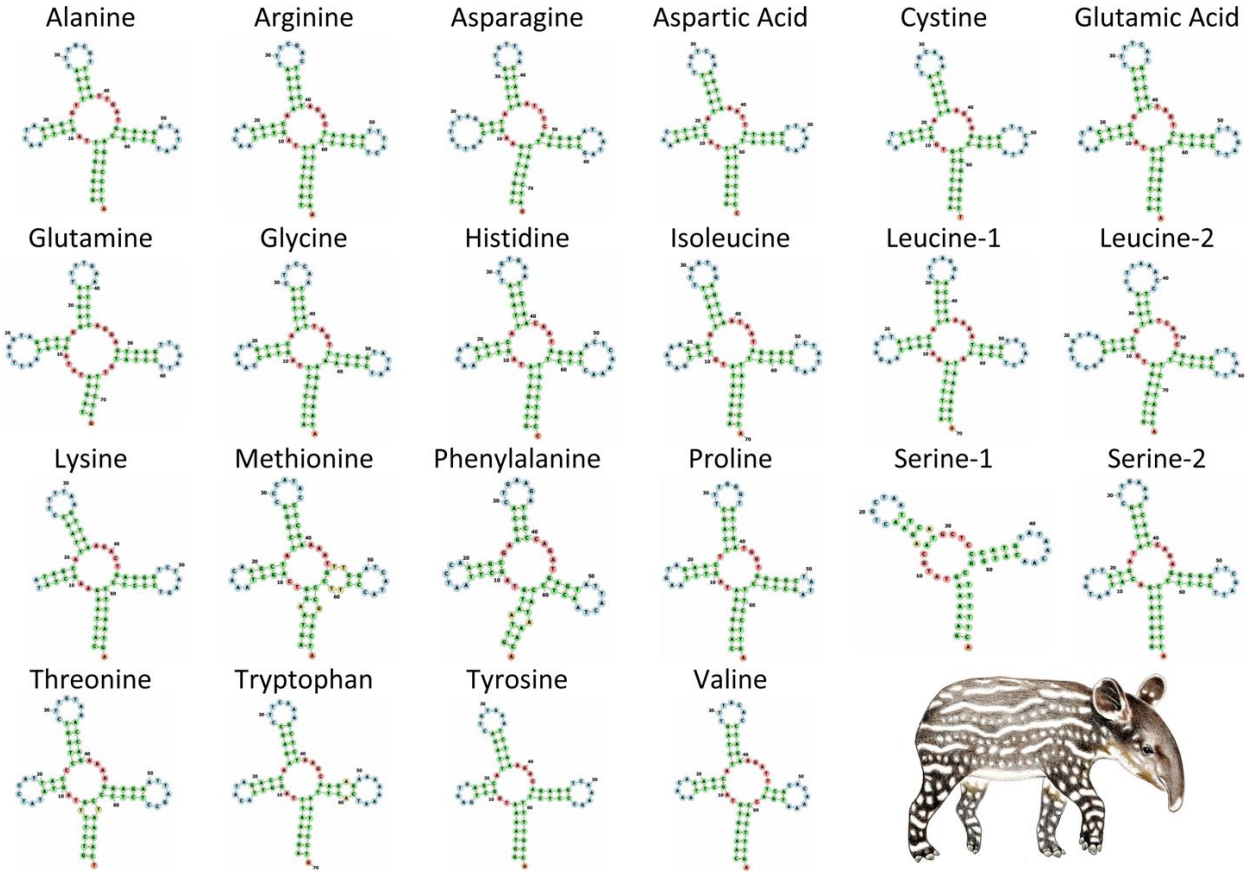
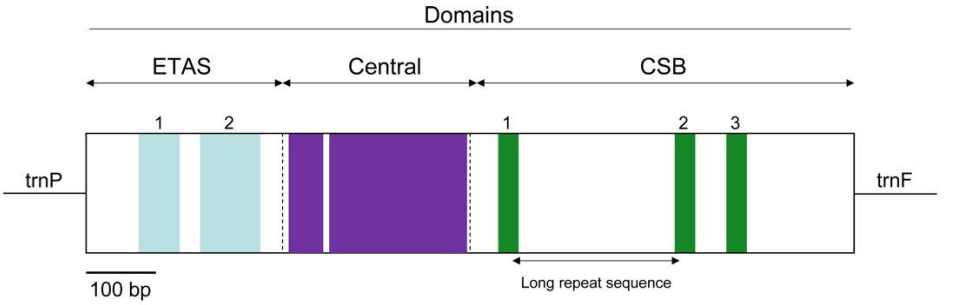


Figure 4: Secondary structure of tRNA's in the mitochondrial genome of *Tapirus bairdii*. Secondary structure of tRNA's visualized using the Forna web-server. Illustration of *Tapirus bairdii* copyright 1990 Stephen D. Nash. Used with permission.

The *rrnS* (12s) and *rrnL* (16s) genes located on the heavy strand are 970 bp and 1,579 bp in length, respectively. The *rrnS* gene is located between the *trnF* and *trnA* genes, with a base composition of *A* = 38.6%, *T* = 23.1%, *C* = 22.0%, and *G* = 16.4%. The *rrnL* gene is located nearby between the *trnV* and *trnL2* genes, with a base composition of *A* = 38.1%, *T* = 24.5%, *C* = 21.6%, and *G* = 15.8%. Both genes show an A-T composition bias, as does the entire mitochondrial genome. Nucleotide composition analysis of the rRNA genes in other species of *Tapirus* has not been conducted. In the closely

related *Rhinoceros unicornis*, a slight A-T bias was found for the *rrnL* gene, while a slight A-C bias was found for the *rrnS* gene [41]. The base composition for both rRNA genes of *R. unicoris* of thymine and cytosine was found to be between 21% and 25%, similar to the composition found in *T. bairdii*.



[AACACCATTGAATGATAAAACTGCTATCGTGTGTCATATCAGTATTAAAAA
TTTTCTTTTTTCCCCCCCCCCGTAGGTACCCCTATGTATATCGTGCATTAAA
TTGTTTGCCCCATGCATATAAGCATGTACATTTGAATTATTAATCTTGCATA
AAAACATCGAATGATTAATTCAACATAATACTGGCAACCAAACATGAATATC
GTCACTCCAGGAATAGAATGGTTGATCTTACATAGTACATATTATTATTGGT
CGTACATACCCCATTAAGTCAAATCATTTCTCGACAACACGCATATCACCTC
CCATGTAAATT**][CTTAATTACCAACTCCCGAGAAATCACCAATCCTTGCGC**
GATCTGCATTTCATTCTCGCTCCGGGCCCATTAACTGTGGGGGTAGTTAAAC
TGAGCTGTATCCGGCATCTGGTTCTTACTTCAGGGCCATCTCACCTAAAATC
GCCTATTCTTTCCTCTTAAATAAGACATCTCGATGGACTAATGACTAATCAG
CCCATGCTCACACATAACTGTGATGTCATGCATTTGGTATTTTTTATAATTT
GGGGATGCTATGACTCAGCTATGGCCGTCTGAGGCCTTAACACATTCAAGCA
AATTGTAGCTGGACTTA][AATTGAACATGATTTACCCGCATCAGATAACCA
TAAGGTGTTATTCAGTCAATGGTCACAGGACATACGTATACACACGCTTACA
$_{1}$CATACGTATACACACGCTTACA$_{2}$CATACGTATACACACGCTTACA$_{3}$CATACG
TATACACACGCTTACA$_{4}$CATACGTATACACACGCTTACA$_{5}$CATACGTATACAC
ACGCTTACA$_{6}$CATACGTATATACACGCTTACA$_{7}$CATACGTATATACACGCTTA
CA$_{8}$CATACGTATATACACGCTTACA$_{9}$CATACGTATATACACGCTTACA$_{10}$CAT
ACGTATATACACGCTTACA$_{11}$CATACGTATACACACGCTTACACACC$_{12}$CATT
AAGTACATGATTATCTTAGCAAACCCCCCCTTCCCCCCATTAAACCTCGCGT
CCATGTATTCTCTAAAGCCTTGCCAAACCCCAAAAACAAAGCCAAATACACG
AAGTCTACAAAGTTTAACTTTTTCAATTCAGCAAACCGCCCTAAACTAATAC
AAACATGCTACTTCAACCAATAAAATTTATGTAGACAGACATCCCCCTAGAT
CTGAAAAATTTTTTTTTAAAACGTTCTCAATCACTTATAAAACAATAATAAA
CCCAAAAATTAACT]

Figure 5: Visual representation of the control region (CR) in the mitochondrial genome of *Tapirus bairdii*. The CR is divided into the extended terminal association sequence (ETAS), central, and conserved sequence block (CSB) domains. Locations of the ETAS 1 and ETAS 2, CSB1, CSB2, CSB3 blocks, as well as the large highly conserved regions within the central domain are shown. The long repetitive motif is indicated in underline and a possible secondary structure is depicted for the region.

The full 1,247 bp long putative CR ranges from position 15,451 to 16,697 and is located between the *trnP* and *trnF* genes. The CR has a slight A-T skew, with an overall nucleotide composition of *A* = 33.2%, *T* = 25.6%, *C* = 28.7%, and *G* = 12.5%, which has been observed in other organisms [43]. Stem-loop structures are located within this putative control region as well as microsatellite repeats. The microsatellite repeats-finder web server found 22 microsatellites within the CR most of them with AC or TA dinucleotide repeats (Table S1). A large tandem repeat 5′-(CAT ACG TAT ACA CAC GCT TAC A)$_{12}$-3′ is found to begin at position 16,152 in the CR. The RNA-Structure web server produced 20 possible secondary structures all containing variable numbers and sizes of stem-loops throughout the entire sequence. These predictions ranged in values of Gibbs free energy (ΔG) from ΔG = −111.8 kcal/mol to ΔG = −111.4 kcal/mol (**Fig. S1**). The O_H region, within the control region (CR), is the origin of replication for the heavy strand (H-strand). The O_L, origin of replication for the light strand (L-strand), of *T. bairdii* is a 33 bp long sequence with a stable stem-loop secondary structure and found within the WANCY cluster, similar to other vertebrate species [44]. A common characteristic for vertebrate mitochondrial genomes, the WANCY cluster, contains a series of five tRNA's (*trnW*, *trnA*, *trnN*, *trnC*, and *trnY*) with a conserved order flanking the O_L region. Secondary structure prediction found 2 possible stem-loop secondary structures for the O_L, ΔG = −11.9 kcal/mol and ΔG = 3.2 kcal/mol (**Fig. S2**).

The three functional domains of the control region found in mammals, namely the extended terminal association sequences (ETAS), central, and conserved sequence block (CSB) domains were also detected observed in the same region of *T. bairdii* (**Fig. 5**) [45]. The length of each domain (ETAS = 322 bp, Central = 316 bp, CSB = 609 bp) is within the normal range of mammalian control regions, notably similar to that of *Tapirus indicus* [34, 45] The CSB domain was A-T rich, with 56.3% A-T composition. The large tandem repeat found to be from position 16,152 to 16,420–between CSB-1 and CSB-2–is similar

to others in Perissodactyla [37,43]. The function of the CSB is still unclear, but their occurrence in vertebrate mitogenomes in general suggests that they play a critical role in genome replication and transcription [46]. A high degree of conservation was observed in the Central domain between members of Perissodactyla and *T. bairdii* (after a multiple alignment of CR's), which is expected amongst vertebrates [45]. Conserved blocks within ETAS (1 and 2) and CSB (1–3) were also identified based on the alignment of sample Perissodactyla and comparison to other mammalian sequences [45]. Higher nucleotide variability within the ETAS and CSB domains follows expectations due to higher substitution rates within these regions compared to the Central domain [47]. Further analysis is needed in the organization and annotation of the CR's of Perissodactyla to better understand the variation between the species.

The ML phylogenetic analysis confirmed the monophyly of the order Perissodactyla considering that all the species belonging to the superfamily Tapiroidea, including *T. bardii* and congeneric species, superfamily Rhinocerotidae, and family Equidae clustered into a single fully supported clade (**Fig. 6**). Within this monophyletic Perissodactyla, the family Equidae occupied a basal position, sister to a second clade composed of representatives from the families Tapiridae and Rhinocerotidae. Within the superfamily Tapiroidea, *T. bardii* was sister to *T. terrestris*. *Tapirus indicus* was sister to the fully supported *T. bardii* + *T. terrestris* clade. Ancestral mtDNA retrieved from fossil and/or subfossil specimens belonging to the genus *Tapirus* most likely will permit a deeper exploration of the phylogenetic relationships in these iconic mammals. The phylogenetic relationships among the different species of Rhinocerotidae are identical to those found by previous studies [35, 48].

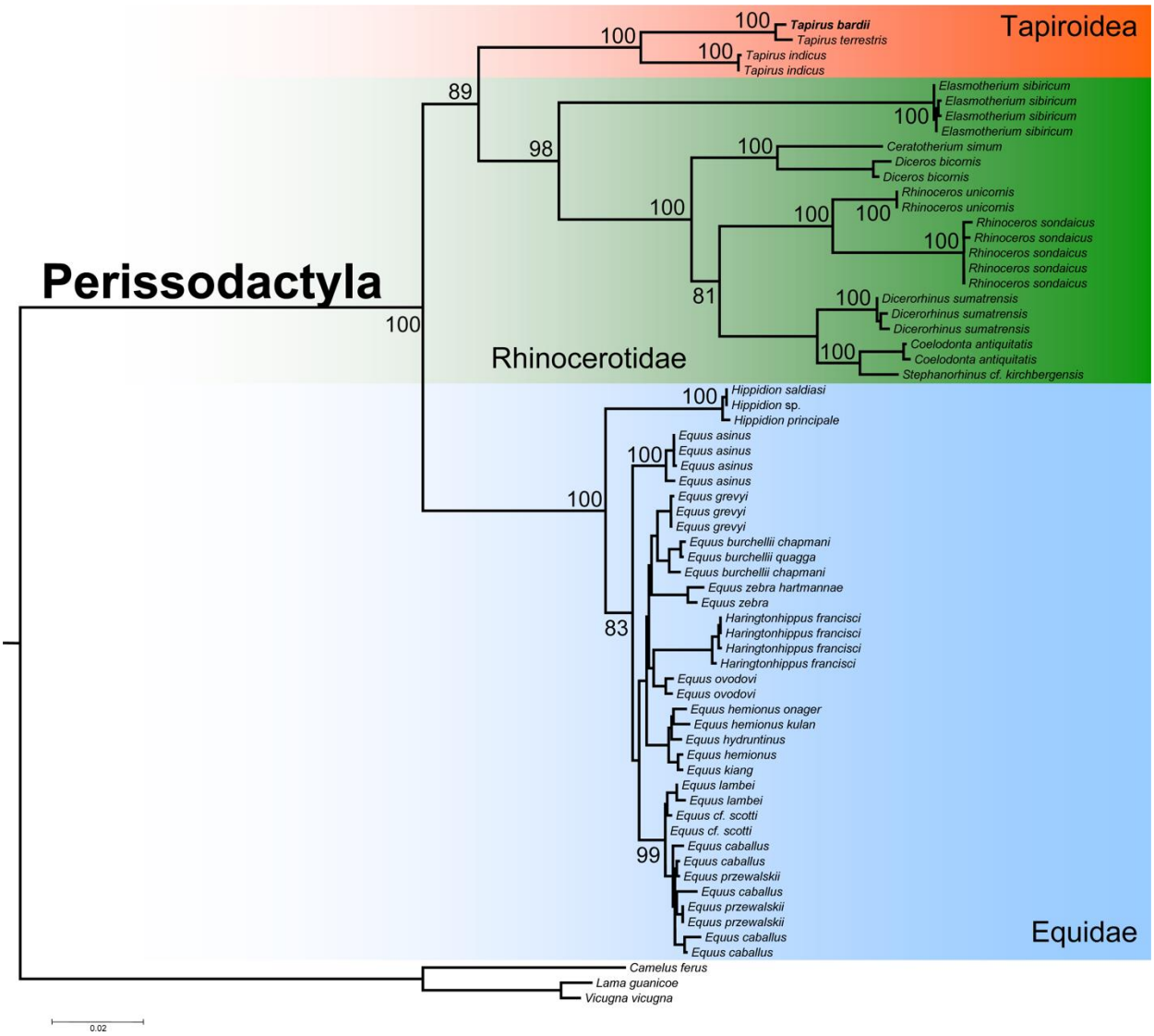


Figure 6: Phylogenetic analysis of *Tapirus bairdii* and related species in the order Perissodactyla. Total evidence phylogenetic tree obtained from ML analysis based on a concatenated alignment of amino acids of the 13 protein-coding genes present in the mitochondrial genome of representatives of the order Perissodactyla. Numbers above or below the branches represent bootstrap values.

Given raising concerns about population decline experienced by multiple populations of *T. bairdii* across its range of distribution [1] it is fundamental to improve our understanding of population abundance, deme dynamics, as well as gaining knowledge on the presence/absence of this species in human altered habitats. Direct invasive sampling may not represent the optimal solution to understand the demography of *T. bairdii* considering this species has become elusive and given that invasive sampling can

disrupt and or stress individuals and populations that may already be experiencing moderate or major local anthropocentric impact [49]. We propose this newly assembled genome can be used as a reference for the retrieval (using bioinformatics strategies) of mitochondrial markers for bioprospecting and biomonitoring of *T. bairdii* when using indirect surveillance strategies such as environmental DNA (eDNA) in the form of scats or blood from insects (iDNA) [50]. Such efforts are currently being tested in other large herbivorous vertebrates with major conservation problems (*e.g.*, in moose) [51] and we believe this study is a step forward towards to implementation of indirect surveillance in *T. bairdii*.

CONCLUSIONS

This study assembled, for the first time, the full mitochondrial genome of the Central American Tapir, *T. bairdii*. This large proboscis-bearing mammal is threatened by deforestation and overhunting, contributing to population decline. The complete annotation and analysis of the mitochondrial genome will contribute to the understanding of selective pressures and evolutionary relationships of *T. bairdii* as well as providing more knowledge for use in conservation efforts of this iconic endangered mega-mammal from the Neotropics.

Data Availability

The DNA-seq data are available at the NCBI Sequence Read Archive (SRA) under Bioproject: PRJNA785336, Biosample accession: SAMN23553527, and SRA accession: SRR17086167. The complete mitochondrial chromosome of T. bairdii is available at GenBank: OM935749.

Acknowledgements

The authors are grateful to Dr. Vincent P. Richards of Clemson University for bioinformatics support. This study was supported by Creative Inquiry and Clemson Thinks^2 at Clemson University. We also appreciate support by Parque Xcaret, Playa del Carmen, Quintana Roo, Mexico for the tissue sample provided.

REFERENCES CITED

1. Schank CJ, Cove MV, Arima EY, Brandt LSE, Brenes-Mora E, Carver A et al (2020). Population status, connectivity, and conservation action for the endangered baird's tapir. *Biological Conservation*, *245*, 108501. https://doi.org/10.1016/j.biocon.2020.108501
2. Paolucci LN, Pereira RL, Rattis L, Silverio DV, Marques NC, Macedo MN et al (2019). Lowland tapirs facilitate seed dispersal in degraded Amazonian forests. Biotropica, 51(2), pp.245-252.
3. Foerster CR, Vaughan C (2006). Home range, habitat use, and activity of baird's Tapir in Costa Rica. Biotropica, 34(3), 423–437. https://doi.org/10.1111/j.1744-7429.2002.tb00556.x.
4. Naranjo EJ (2009). Ecology and conservation of baird's Tapir in Mexico. Tropical Conservation Science, 2(2), 140–158. https://doi.org/10.1177/194008290900200203
5. Pérez-Flores J, Arias-Domínguez H, Arias-Domínguez N (2020). First documented predation of a Baird's tapir by a jaguar in the Calakmul region, Mexico. Neotropical Biology and Conservation. doi:10.3897/neotropical.15.e57029.
6. García M, Jordan C, O'Farril G, Poot C, Meyer N (2014). Baird's Tapir. *IUCN Red List of Threatened Species*. https://doi.org/10.2305/iucn.uk.2016-1.rlts.t21471a45173340.en.
7. Ashley MV, Norman JE, Stross L (1996) Phylogenetic analysis of the perissodactylan family Tapiridae using mitochondrial cytochrome *c* oxidase (COII) sequences. J Mammal Evol 3, 315–326. https://doi.org/10.1007/BF02077448.
8. Norton, J.E. and Ashley, M.V. (2004a), Genetic variability and population structure among wild Baird's tapirs. Animal Conservation, 7: 211-220. https://doi.org/10.1017/S1367943004001295
9. Norton, J.E. and Ashley, M.V. (2004b), Genetic variability and population differentiation in captive baird's Tapirs (*Tapirus bairdii*). Zoo Biol., 23: 521-531. https://doi.org/10.1002/zoo.20031
10. Ferrero, Brenda & Noriega, Jorge. (2009). A new upper Pleistocene tapir from Argentina: Remarks on the phylogenetics and diversification of neotropical Tapiridae. Journal of Vertebrate Paleontology 27: 504-511.
11. Ruiz-García M, Vasquez C, Pinedo-Castro M, Sandoval S, Castellanos A, Kaston F, et al (2012). Phylogeography of the Mountain Tapir (*Tapirus pinchaque*) and the Central American Tapir

(*Tapirus bairdii*) and the origins of the three Latin-american TAPIRS by means of mt-Cyt-B Sequences. Current Topics in Phylogenetics and Phylogeography of Terrestrial and Aquatic Systems. https://doi.org/10.5772/35361

12. Hulbert R (1995). The giant tapir, *Tapirus haysii*, from Leisey Shell Pit 1A and other Florida Invingtonian localities. Bulletin of Florida Museum of Natural History. 37. 515 - 551.
13. Cozzuol, M. A., Clozato, C. L., Holanda, E. C., Rodrigues, F. H., Nienow, S., de Thoisy, B., Redondo, R. A., & Santos, F. R. (2013). A new species of tapir from the Amazon. Journal of Mammalogy 94, 1331–1345. https://doi.org/10.1644/12-mamm-a-169.1
14. Vivas-Toro, I., Ortega, J., & Baeza, J. A. (2021). The complete mitochondrial genome of the Honduran white bat Ectophylla alba (Allen 1982) (Chiroptera: Phyllostomidae). Gene, 802, 145868. https://doi.org/10.1016/j.gene.2021.145868
15. López-Cuamatzi, I.L., Ortega, J. & Baeza, J.A. The complete mitochondrial genome of the 'Zacatuche' Volcano rabbit (Romerolagus diazi), an endemic and endangered species from the Volcanic Belt of Central Mexico. Mol Biol Rep 49, 1141–1149 (2022). https://doi.org/10.1007/s11033-021-06940-7
16. Jin JJ, Yu WB, Yang JB, Song Y, Depamphilis CW, Yi TS, Li DZ (2020) GetOrganelle: a fast and versatile toolkit for accurate de novo assembly of organelle genomes. Genome Biol 21:1–31. https:// doi. org/ 10. 1186/ s13059- 020- 02154-5
17. Bernt M, Donath A, Jühling F, Externbrink F, Florentz C, Fritzsch G et al (2013) MITOS: improved de novo metazoan mitochondrial genome annotation. Molecular Phylogenetics and Evolution 69:313-319. doi:10.1016/j.ympev.2012.08.023.
18. Artimo P, Jonnalagedda M, Arnold K, Baratin D, Csardi G, De Castro E et al (2012) ExPASy: SIB bioinformatics resource portal. Nucleic Acids Research 40:597-603. doi: 10.1093/nar/gks400.
19. Stothard P, Wishart DS (2004). Circular genome visualization and exploration using CGView. *Bioinformatics*, *21*(4), 537–539. https://doi.org/10.1093/bioinformatics/bti054

20. Stothard, P (2000) The sequence manipulation suite: JavaScript programs for analyzing and formatting protein and DNA sequences. Biotechniques 28:1102-1104. doi: 10.2144/00286ir01.
21. Cucini C., Leo C., Iannotti N., Boschi S., Brunetti C., Pons J., Fanciulli P. P., Frati F., Carapelli A., Nardi F. (2021) EZmito: a simple and fast tool for multiple mitogenome analyses, Mitochondrial DNA Part B, 6(3), 1101-1109. Doi: 10.1080/23802359.2021.1899865
22. Kerpedjiev P, Hammer S, Hofacker IL (2015) Forna (force-directed RNA): Simple and effective online RNA secondary structure diagrams. Bioinformatics 31:3377-3379 doi: 10.1093/bioinformatics/btv372.
23. Wang DP, Wan HL, Zhang S, Yu J (2009) γ-MYN: a new algorithm for estimating Ka and Ks with consideration of variable substitution rates. Biol Direct 4:1-18. doi:10.1186/1745-6150-4-20.
24. Wang D, Zhang Y, Zhang Z, Zhu J, Yu J (2010) KaKs_Calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies. Genom Proteom Bioinf 8:77-80. doi: 10.1016/S1672-0229(10)60008-3.
25. Conrad, I., Craft, A., Thurman, C. L., & Baeza, J. A. (2021). The complete mitochondrial genome of the red-jointed brackish-water fiddler crab Minuca minax (LeConte 1855) (Brachyura: Ocypodidae): New family gene order, and purifying selection and phylogenetic informativeness of protein coding genes. Genomics, 113(1 Pt 2), 565–572. https://doi.org/10.1016/j.ygeno.2020.09.050
26. Benson G (1999). Tandem repeats finder: A program to analyze DNA sequences. Nucleic Acids Research, *27*(2) 573–580.
27. Bikandi J, Millán RS, Rementeria A, Garaizar J (2004) In silico analysis of complete bacterial genomes: PCR, AFLP–PCR and endonuclease restriction. Bioinformatics 20:798-799. doi: 10.1093/bioinformatics/btg491.
28. Bellaousov S, Reuter JS, Seetin MG, Mathews DH (2013). Rnastructure: Web servers for rna secondary structure prediction and analysis. Nucleic Acids Research, 41:W471-W474. https://doi.org/10.1093/nar/gkt290.

29. Tan MH, Gan HM, Schultz MB, Austin CM (2015) MitoPhAST, a new automated mitogenomic phylogeny tool in the post-genomic era with a case study of 89 decapod mitogenomes including eight new freshwater crayfish mitogenomes. Mol Phylogenetics Evol 85:180-188. doi:10.1016/j.ympev.2015.02.009.
30. Sievers F, Higgins DG (2014) Clustal omega. Current Protocols Bioinformatics 48:3-13. doi: 10.1002/0471250953.bi0313s48.
31. Capella-Gutiérrez S, Silla-Martínez J, Gabaldón T (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics, 25: 1972–1973, https://doi.org/10.1093/bioinformatics/btp348
32. Abascal F, Zardoya R, Posada D (2005) ProtTest: selection of best-fit models of protein evolution. Bioinformatics 21:2104-2105. doi: 10.1093/bioinformatics/bti263.
33. Nguyen LT, Schmidt HA, Von Haeseler A, Minh BQ (2015) IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol 32:268-274. doi: 10.1093/molbev/msu300.
34. Muangkram Y, Wajjwalku W, Kaolim N, Buddhakosai W, Kamolnorranath S, Siriaroonrat B, et al (2015). The complete mitochondrial genome of the Asian tapirs (*Tapirus indicus*): the only extant Tapiridae species in the old world. Mitochondrial DNA. Part A, DNA mapping, sequencing, and analysis, 27(1), 413–415. https://doi.org/10.3109/19401736.2014.898283
35. Willerslev E, Gilbert MT, Binladen J, Ho SYW, Campos PF, Ratan A, et al (2009). Analysis of complete mitochondrial genomes from extinct and extant rhinoceroses reveals lack of phylogenetic resolution. *BMC Evolutionary Biology*, *9*(1), 95. https://doi.org/10.1186/1471-2148-9-95.
36. Luo Y, Chen Y, Liu F, Jiang C, Gao Y (2011). Mitochondrial genome sequence of the TIBETAN wild ASS (EQUUS kiang). Mitochondrial DNA, 22(1-2), 6–8. https://doi.org/10.3109/19401736.2011.588221

37. Xu X, Arnason U. (1994). The complete mitochondrial DNA sequence of the horse, *Equus caballus*: extensive heteroplasmy of the control region. *Gene*, *148*(2), 357–362. https://doi.org/10.1016/0378-1119(94)90713-7.
38. Boore JL (2001) Mitochondrial gene arrangement source guide, version 6.0. Department of the Environment Joint Genome Institute, Walnut Creek (available from http://www.jgi.doe.gov/Mitochondrial Genomics.html).
39. Ojala D, Montoya J, Attardi G. (1981) tRNA punctuation model of RNA processing in human mitochondria. Nature 290, 470–474. https://doi.org/10.1038/290470a0.
40. Achilli A, Olivieri A, Soares P, Lancioni H, Kashani BH, Perego UA, et al (2012). Mitochondrial genomes from modern horses reveal the major haplogroups that underwent domestication. Proceedings of the National Academy of Sciences, 109(7), 2449–2454. https://doi.org/10.1073/pnas.1111637109.
41. Xu X, Janke A, Arnason U (1996). The complete mitochondrial DNA sequence of the greater Indian rhinoceros, *Rhinoceros unicornis*, and the Phylogenetic relationship among Carnivora, Perissodactyla, and Artiodactyla (+ Cetacea). *Molecular biology and evolution*, *13*(9), 1167–1173. https://doi.org/10.1093/oxfordjournals.molbev.a025681.
42. Watanabe Y, Suematsu T, Ohtsuki T (2014). Losing the stem-loop structure from metazoan mitochondrial tRNAs and co-evolution of interacting factors. *Frontiers in Genetics*, *5*, 109. https://doi.org/10.3389/fgene.2014.00109.
43. Muangkram Y, Armano A, Wajjwalku W, Pinyopummintr T, Thongtip N, Kaolim N, et a1 (2017). Genetic diversity of the captive Asian tapir population in Thailand, based on mitochondrial control region sequence data and the comparison of its nucleotide structure with Brazilian tapir, Mitochondrial DNA. Part A, DNA mapping, sequencing, and analysis, 28(8), 597-601. doi: 10.3109/24701394.2016.1149828

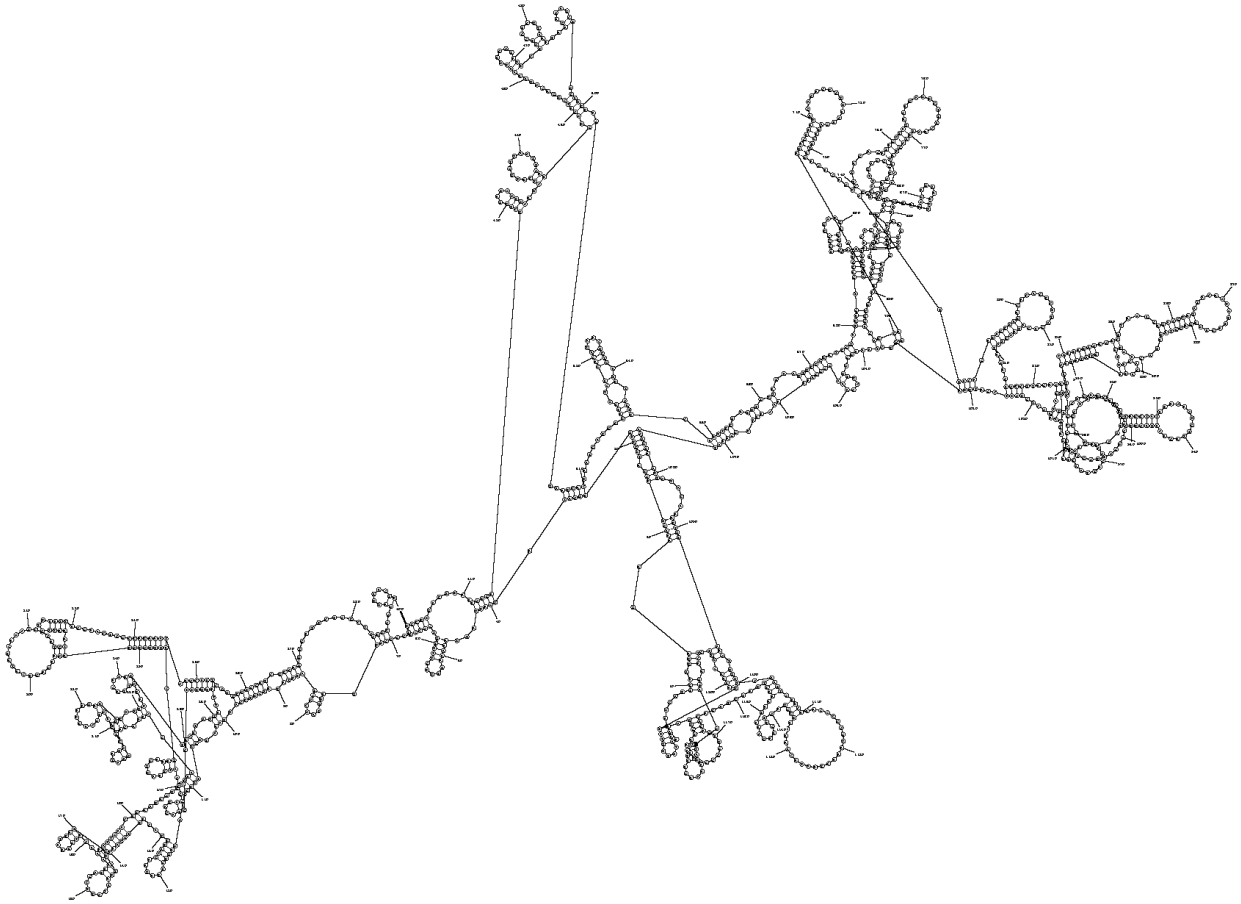
44. Seutin G, Lang B, Mindell D, Morais R. (1994). Evolution of the WANCY region in amniote mitochondrial DNA. *Molecular Biology and Evolution*. 11. 329-40. 10.1093/oxfordjournals.molbev.a040116.
45. Sbis`a, E., Tanzariello, F., Reyes, A., Pesole, G., Saccone, C., (1997). Mammalian mitochondrial D-loop region structural analysis: identification of new conserved sequences and their functional and evolutionary implications. Gene 205 (1–2), 125–140. https://doi.org/10.1016/s0378-1119(97)00404-6
46. Satoh, TP, Miya M, Mabuchi, K, Nishida M.(2016) Structure and variation of the mitochondrial genome of fishes. BMC Genomics **17**;719 https://doi.org/10.1186/s12864-016-3054-y
47. Pesole, G., Gissi, C., De Chirico, A. et al (1999). Nucleotide Substitution Rate of Mammalian Mitochondrial Genomes. J Mol Evol (48), 427–434. https://doi.org/10.1007/PL00006487
48. Margaryan A, Sinding MHS, Liu S, Vieira, FG, Chan YL, Nathan SK, et al (2020). Recent mitochondrial lineage extinction in the critically endangered Javan rhinoceros. Zoological Journal of the Linnean Society, 190(1), 372-383.
49. Le Breton, T.D., Zimmer, H.C., Gallagher, R.V. (2019) Using IUCN criteria to perform rapid assessments of at-risk taxa. Biodivers Conserv 28: 863–883. https://doi.org/10.1007/s10531-019-01697-9
50. Schnell IB, Sollmann R, Calvignac-Spencer S, et al. (2015) iDNA from terrestrial haematophagous leeches as a wildlife surveying and monitoring tool – prospects, pitfalls and avenues to be developed. *Frontiers in Zoology* 12:24. https://doi.org/10.1186/s12983-015-0115-z
51. Lyet A, Pellissier L, Valentini A. et al (2021) eDNA sampled from stream networks correlates with camera trap detection rates of terrestrial mammals. Sci. Rep. 11: 11362. https://doi.org/10.1038/s41598-021-90598-5

SUPPLEMENTAL INFORMATION

Table S1. Position and identity of the microsatellites repeat in the Control Region

Microsatellite Repeats Finder

Position	Cycle	Repeats	Sequence
28	2	3	GTGTGT
56	2	3	TTTTTT
62	3	3	CCCCCCCCC
246	3	3	TATTATTAT
525	2	3	CACACA
556	2	3	TTTTTT
710	2	3	ACACAC
732	2	3	ACACAC
754	2	3	ACACAC
776	2	3	ACACAC
798	2	3	ACACAC
820	2	3	ACACAC
839	2	3	TATATA
861	2	3	TATATA
883	2	3	TATATA
905	2	3	TATATA
927	2	3	TATATA
952	2	3	ACACAC
962	2	3	ACACAC
996	2	3	CCCCCC
1005	2	3	CCCCCC
1189	2	4	TTTTTTTT



ENERGY = -111.8 CR

Figure S1. Secondary structure of the Control Region

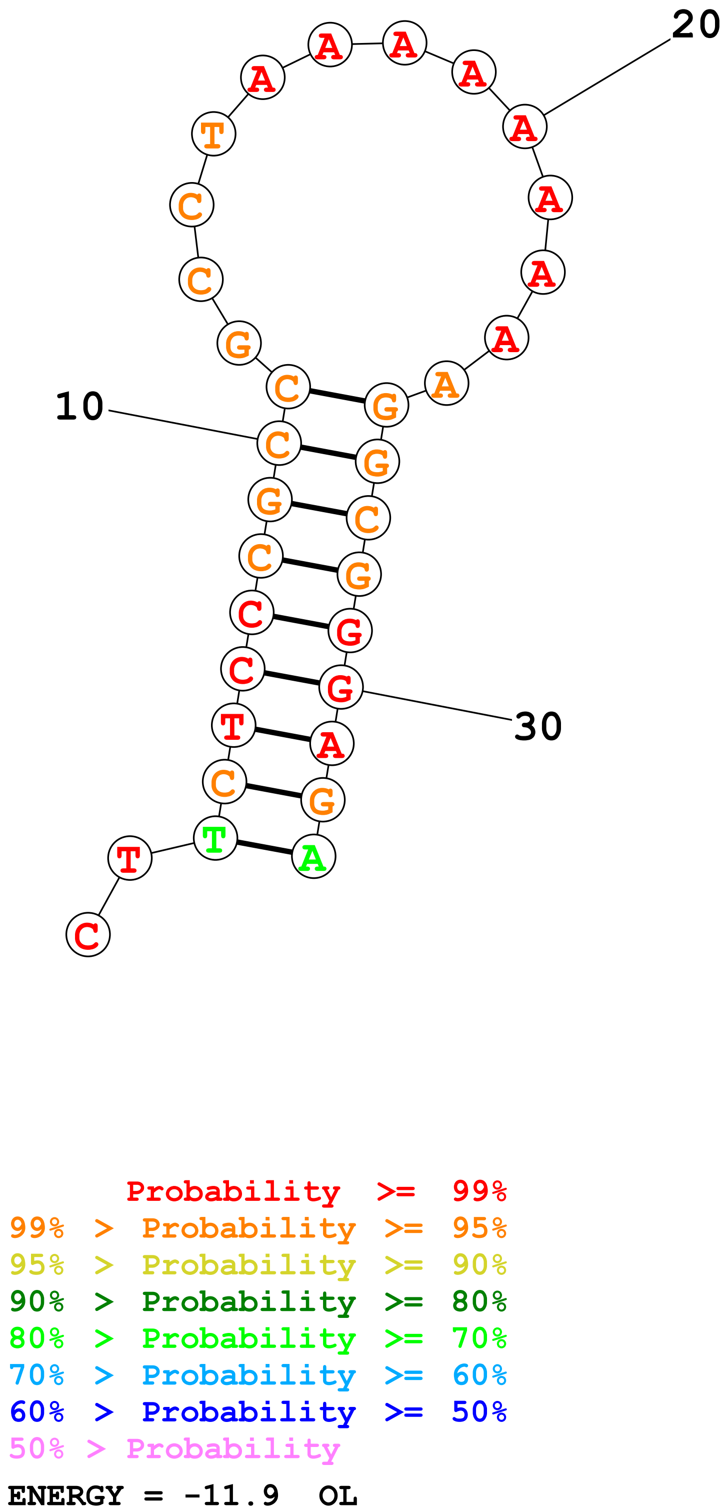


Figure S2. Secondary structure of the Origin of Replication in the Light Strand