Supplemental Figures for

Birth of protein-coding exons by ancient domestication of LINE-1 retrotransposon

Koichi Kitao, Kenji Ichiyanagi, So Nakagawa

Table of Contents:

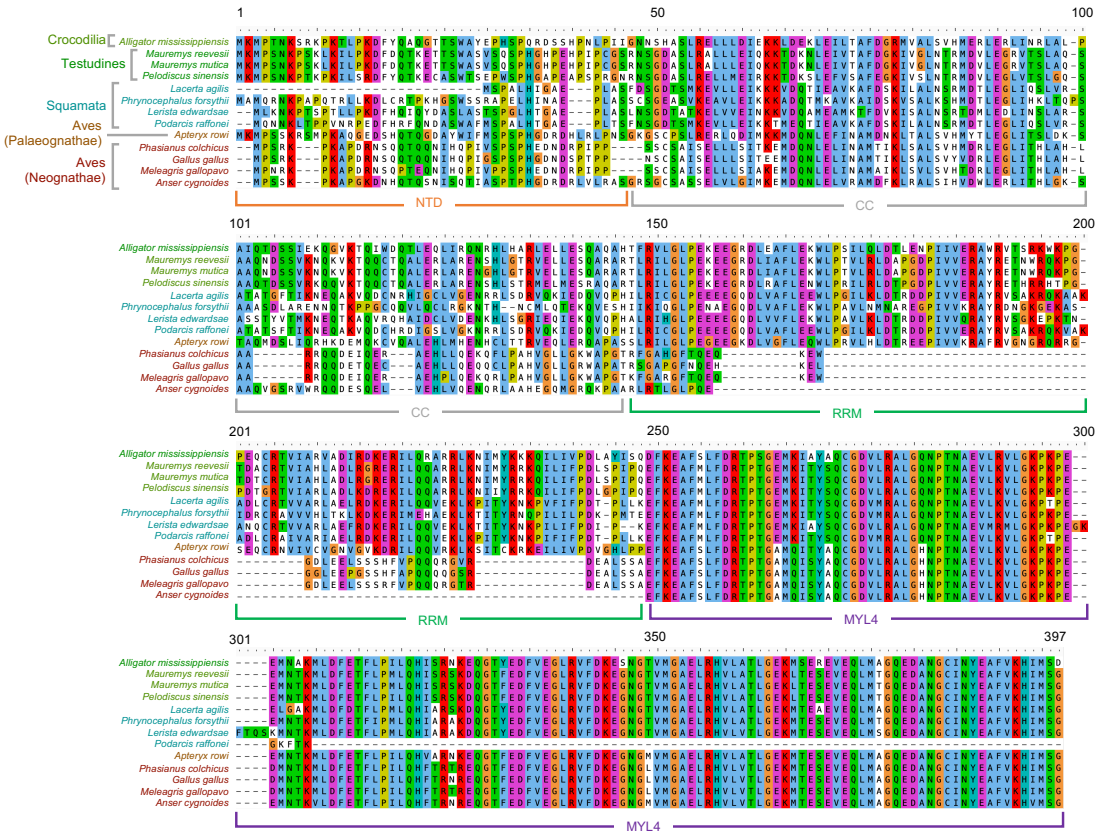

Figure S1. An alignment of Lyosin-like proteins obtained from NCBI BLASTP search. The accession numbers of the proteins were listed in **Dataset S1**.

A

Species	Genome	Clade	Mutations
Nothoprocta ornata	GCA_5F013398335.1	Palaeognathae	nonsense mutation
Nothoprocta pentlandii	GCA_5F013398315.1	Palaeognathae	nonsense mutation (2)
Nothoprocta perdicaria	GCF_5F003342845.1	Palaeognathae	nonsense mutation
Casuarius casuarius	GCA_5F003342895.1	Palaeognathae	deletion (1 base)
Rafetus swinhoei	GCA_5F019425775.1	Testudines	deletion (1 base)
Sphenodon punctatus	GCA_5F003113815.1	Rhynchocephalia	deletion (1 base)
Varanus komodoensis	GCF_5F004798865.1	Squamata	nonsense mutation, deletion (2 bases), insertion (23 bases)

B

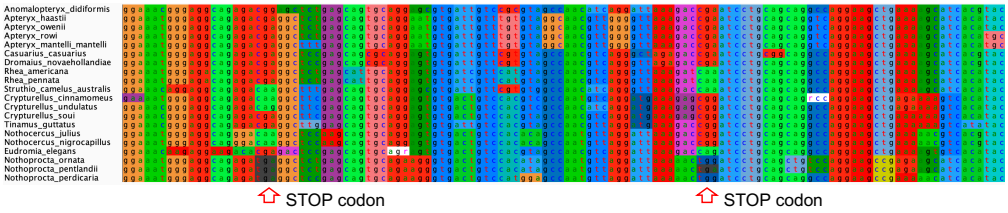

Figure S2. Mutations causing the ORF truncation of Lyosin exon L. (A) Summary of mutations observed in the species where the intact exon L was not detected despite positive hit on the BLAT search. (B) A shared nonsense mutation ("tag" stop codon) in the genus *Nothoprocta* (left arrow) and an additional species-specific nonsense mutation (right arrow).

A

		M0	M1 versus M2		M7 versus M8		No. of positive selection sites	
	No. of species	d_N/d_S	2ΔL	*p*-value	2Δl	*p*-value	M2	M8
Palaeognathae	15	0.5188	10.0666	0.0065	12.6683	0.0018	1	2
Crocodilia	4	1.2181	1.5544	0.4597	1.5544	0.4597	0	0
Testudines	23	0.4116	0.0000	1.0000	8.1628	0.0169	0	1
Squamata	14	0.5351	0.0321	0.9841	4.4942	0.1057	0	0

B

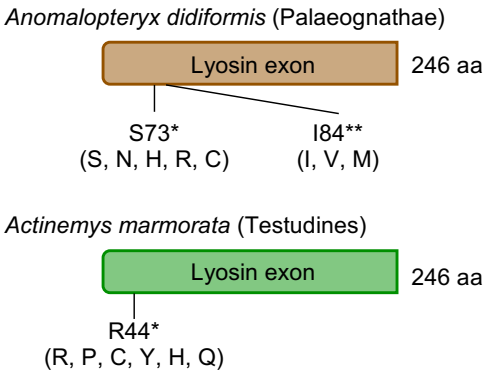

Figure S3. Analysis of the nonsynonymous and synonymous codon substitution (d_N/d_S) ratio. (A) 2ΔL indicates a two-fold difference in the natural log values of the maximum likelihood ratio from pairwise comparisons of the different models. The *p*-value indicates the confidence in rejecting the natural models (M1a or M7) in favor of the positive selection model (M2a or M8) using Pearson's chi-squared test. Codons under positive selection were identified with a posterior probability of 95% by Bayes empirical Bayes (BEB) analysis in M8. (B) The sites under positive selection were illustrated based on the Lyosin exon L of representative species. Amino acids other than the representative sequences were shown in brackets. *: supported in M8 model, **: supported in M2 and M8 models.

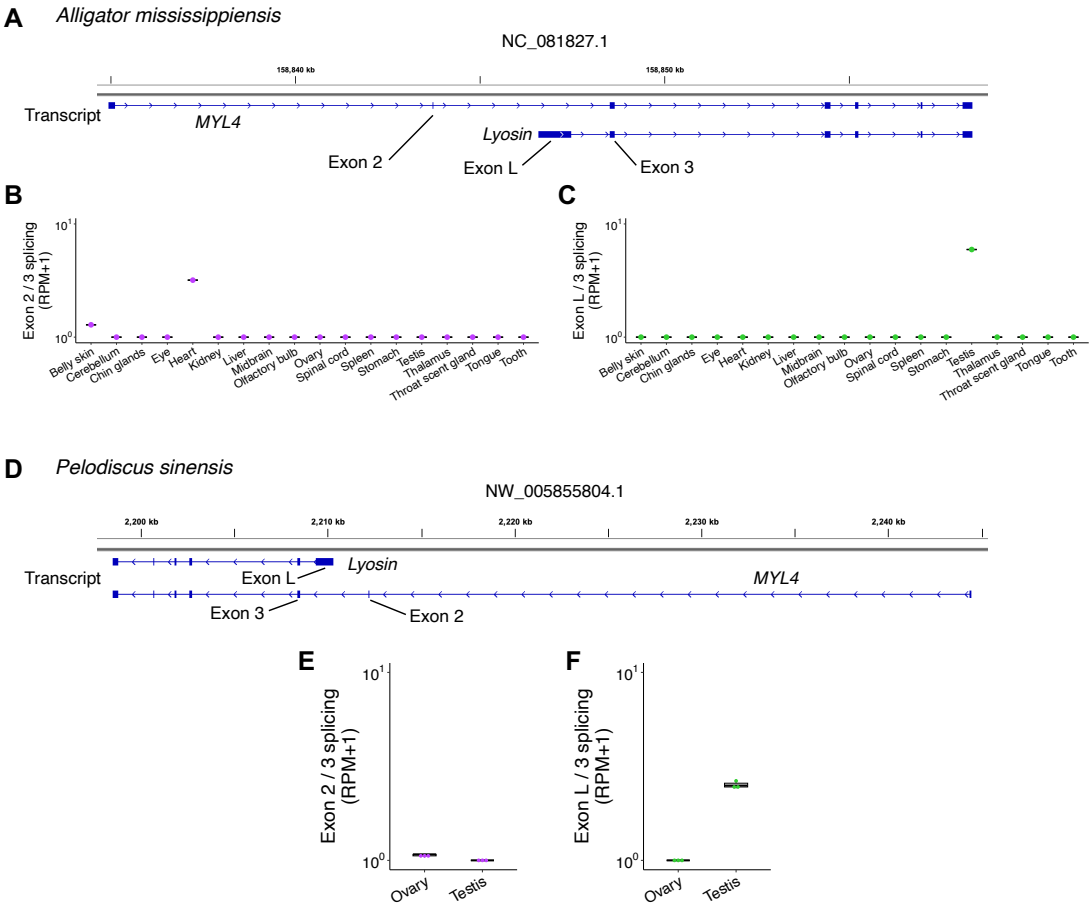

Figure S4. Tissue expression of the Lyosin transcript in the American alligator and Chinese soft-shelled turtle. (A) Genome browser view of the *MYL4* gene in American alligator (*Alligator mississippiensis*). The blue lines represent the assembled transcripts by RNA-seq. (B and C) Box plot and point representations of splice junction reads spanning exon 2 to 3 (B) and exon L to 3 (C) of the MYL4 gene in American alligator (*Alligator mississippiensis*). (D-E) The genome browser view and the normalized number of splice junction reads in Chinese soft-shelled turtle (*Pelodiscus sinensis*). RPM, reads per million.

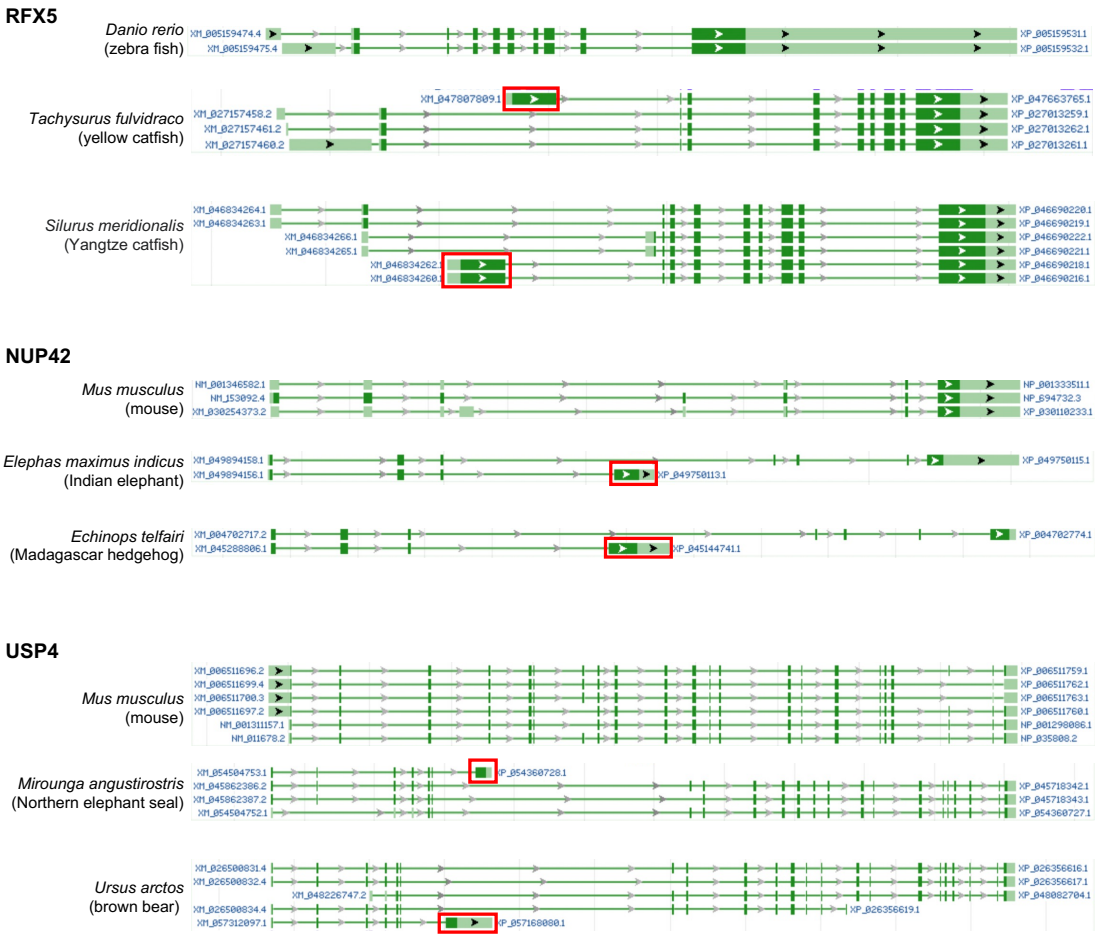

Figure S5. Other genes with ORF1p-like isoforms. Screenshots from the NCBI gene browser. The squared exons are non-canonical alternative exons encoding ORF1p-like proteins.

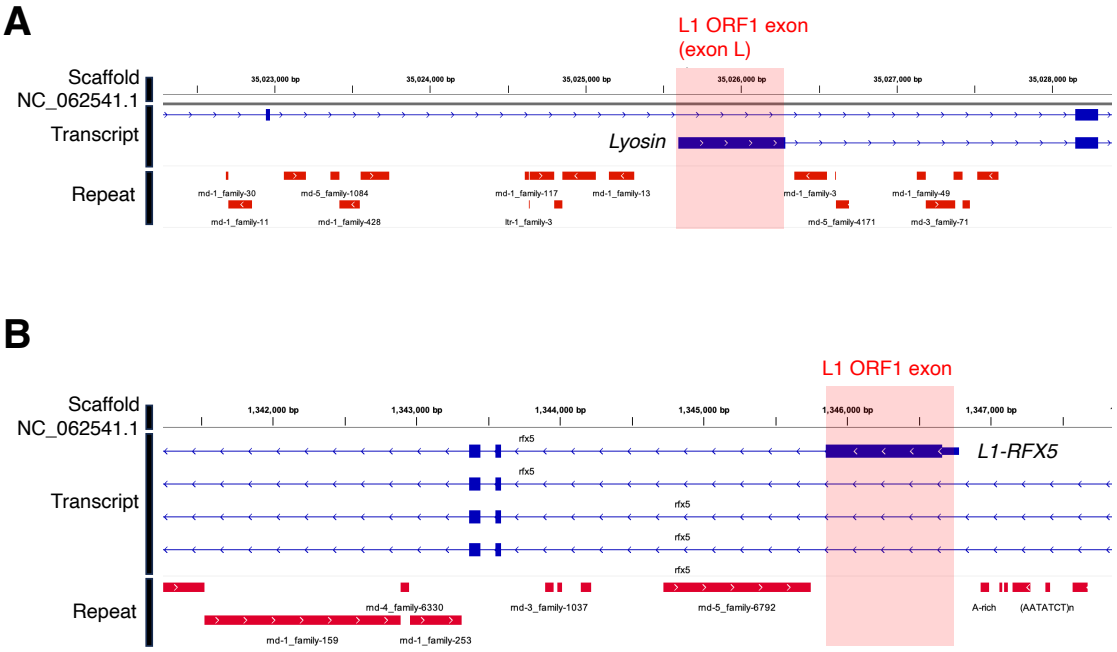

Figure S6. No overlap between the L1 ORF1 exon and de novo predicted repeats by RepeatModeler2. (A) The genome browser view of *Lyosin* exon L in green anole (*Anolis carolinensis*). The blue lines indicate the transcripts of the *MYL4* gene. The repeat consensus sequence was constructed by RepeatModeler version 2.0.4 with "-LTRStruct" option. RepeatMasker version 4.1.5 was used for masking the genome assembly with the constructed repeat consensus sequences. (B) The genome browser view of *L1-RFX5* and repeat sequences in *Tachysurus fulvidraco.*

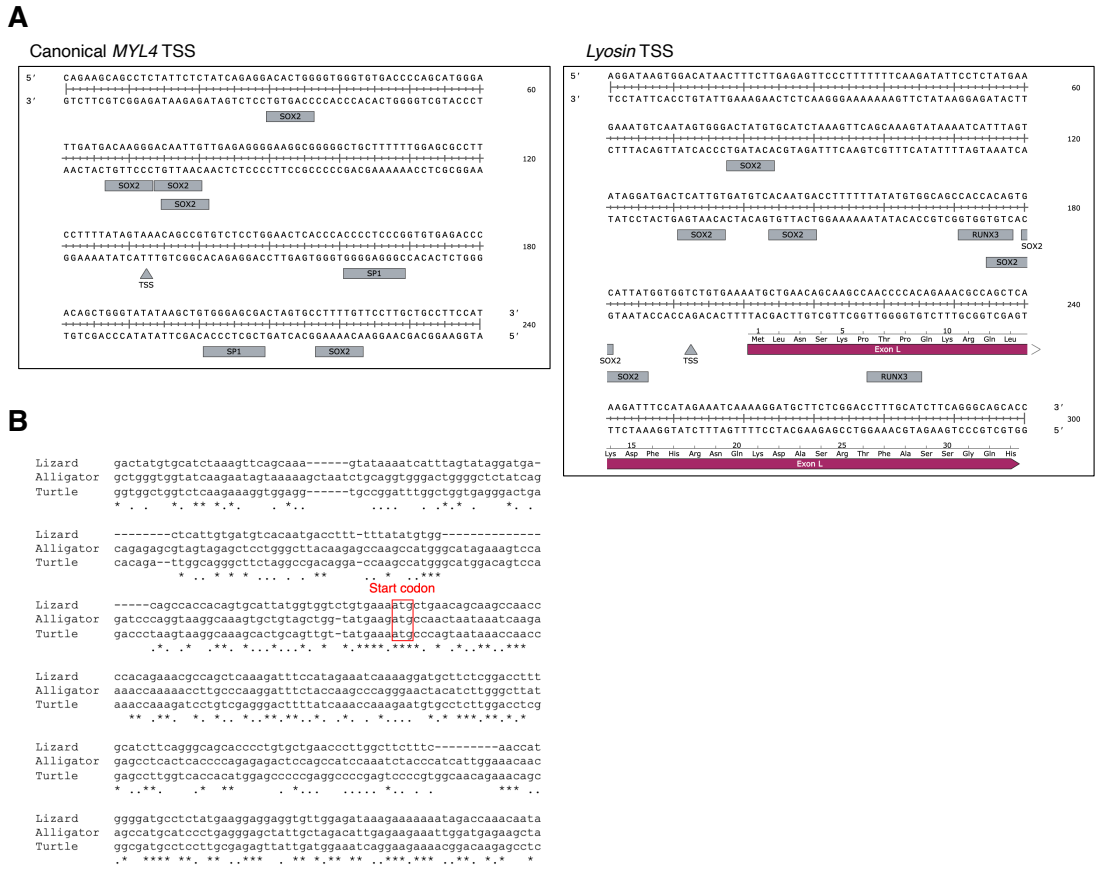

Figure S7. Transcription factor binding sites of the *MYL4* and *Lyosin* isoforms. (A) The transcription binding sites were mapped by the JASPAR database (https://jaspar.elixir.no/) using the latest version of *Homo sapiens* of RUNX3 (ID: MA0684.3), SP1 (ID: MA0079.5), SOX2 (ID: MA0143.5), and YY1 (ID: MA0095.2) in the Transcription start site (TSS) of *Anolis carolinensis MYL4* and *Lyosin*. TSS was inferred by RNA-seq. (B) The nucleotide sequence alignment of the upstream region of the exon L. The start codon of the exon L is indicated. Lizard, *Anolis carolinensis*; Alligator, *Alligator mississippiensis*; Turtle, *Pelodiscus sinensis*.

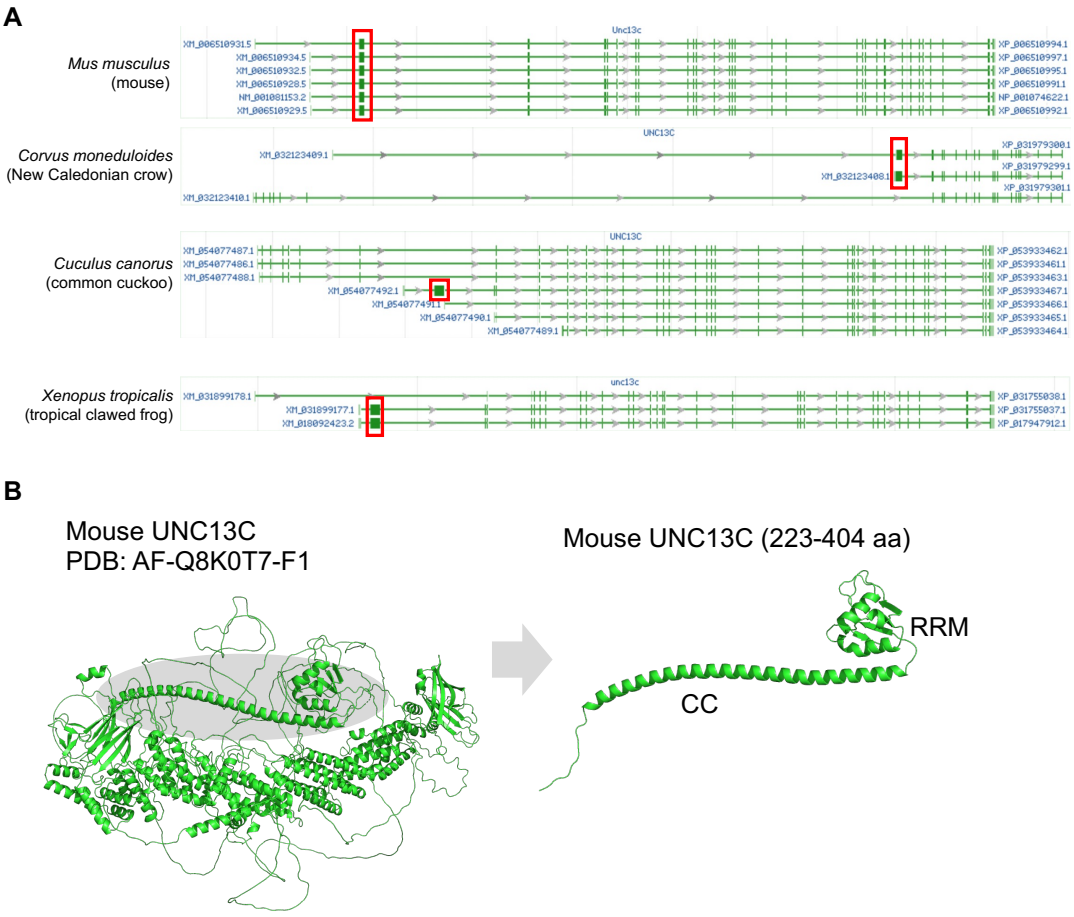

Figure S8. A canonical exon of *UNC13C* encodes a protein similar to L1 ORF1p. (A) Exons encoding ORF1p-like amino acids are enclosed in squares. In mouse, all splicing variants of *Unc13c* contain the exon encoding ORF1p-like amino acids. In the three species below, splicing variants without ORF1p-like exons are annotated. (B) The 3D structure of mouse UNC13C predicted by Alphafold2 (PDB: AF-Q8K0T7-F1). On the left is the predicted structure of full-length UNC13C, while on the right is the partial structure corresponding to the L1 ORF1p-like region. CC, coiled-coil; RRM, RNA recognition motif.