The life history of domesticated genes illuminates the evolution of novel mammalian genes

Kordiš, D.

Department of Molecular and Biomedical Sciences,

J. Stefan Institute, Ljubljana, Slovenia

INTRODUCTION

- **Domesticated or exapted transposable element-derived sequences have contributed diverse and abundant regulatory and protein-coding sequences to the host genomes.**

- Molecular domestication of transposases, integrases, reverse transcriptases, and envelope proteins has **occurred repeatedly** during the evolution of **diverse major eukaryote lineages** and, **during neofunctionalization, some of the newly obtained functions are becoming essential for survival of the organism.**

- Vertebrates, especially mammals, possess numerous **single copy domesticated genes that have originated from intronless multicopy retroelements , DNA transposons or from their remains.**

- **Domestication** may require additional **mutations** that modify expression of the gene and the specificity of interaction of the recruited protein with nucleotide sequences or other proteins.

- During the domestication process, ***de novo*** **acquisition of the regulatory regions is a prerequisite for the survival of domesticated genes**.

MOLECULAR DOMESTICATION

Transposable element-derived sequences can evolve to become novel genes

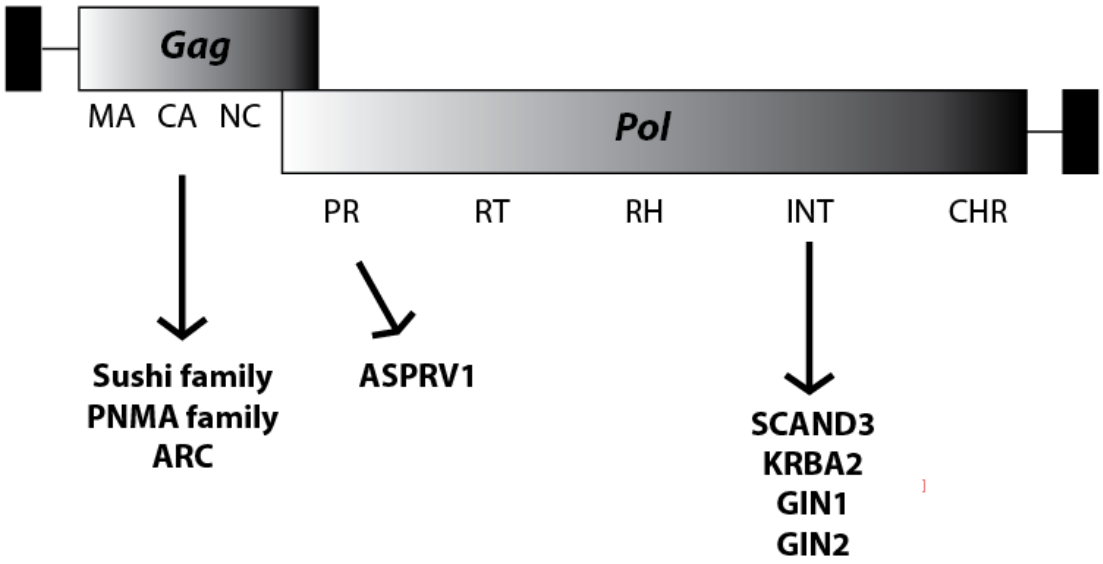
Molecular domestication or **exaptation** – the shift of the function/trait during evolution

Ancestral transposable element	Domesticated gene
Multi-copy	Single-copy
Can (autonomously) transpose	Cannot transpose
Present at different positions in the genomes of diverse species	At orthologous loci in different organisms

- **Domesticated genes are highly conserved in different species**
- **Domesticated genes often evolve under negative selection.**

FAMILIES OF METAVIRIDAE-DERIVED DOMESTICATED GENES

Analyzed **domesticated genes originated** from the **Metaviridae** (Ty3/Gypsy) group of LTR retrotransposons.



Numerous Metaviridae derived genes have been discovered in the human genome and are classified into ***five distinct families: SASPase (ASPRV1), Sushi (=Mart), SCAN, Paraneoplastic (PNMA), and ARC***.

However, their evolutionary history and dynamics have been only partially explored, due to absence of genome data or to the limited analysis of a single family of domesticated genes.

AIMS

- to gain a **comprehensive insight into the origin, distribution, diversity, and evolution of the domesticated genes in chordates**.
- We **traced the genesis and expansion of the domesticated genes through comparative genomic and phylogenomic analyses**, using publicly available whole-genome information from **more than 90 chordate genomes**.
- An **extensive phylogenomic analysis of all the domesticated genes** in chordate, and especially mammalian, genomes has provided crucial information as to **where and when Metaviridae gag, retroelement protease, and integrase domains were transformed into domesticated genes.**
- we elucidated the **timing of the domestication events, clarify their origins and evolution, and provide new insights into their regulatory and functional diversification.**

METHODS

- **Data mining** of genome databases,
- **Phylogenomic analysis** of domesticated genes,

- **Phylogenetic analysis of Metaviridae** in Deuterostomia (**progenitors of domesticated genes**),
- **Phylogenetic analysis of domesticated genes**,
- **Synteny analysis of domesticated genes**,

- Analysis of ***de novo*** **acquired regulatory regions in domesticated genes**,
- The analysis of **transcription factor binding sites in the promoters of human and mouse domesticated genes**,
- The analysis of **expression profiles of human and mouse domesticated genes**.

GENE STRUCTURES OF DOMESTICATED GENES

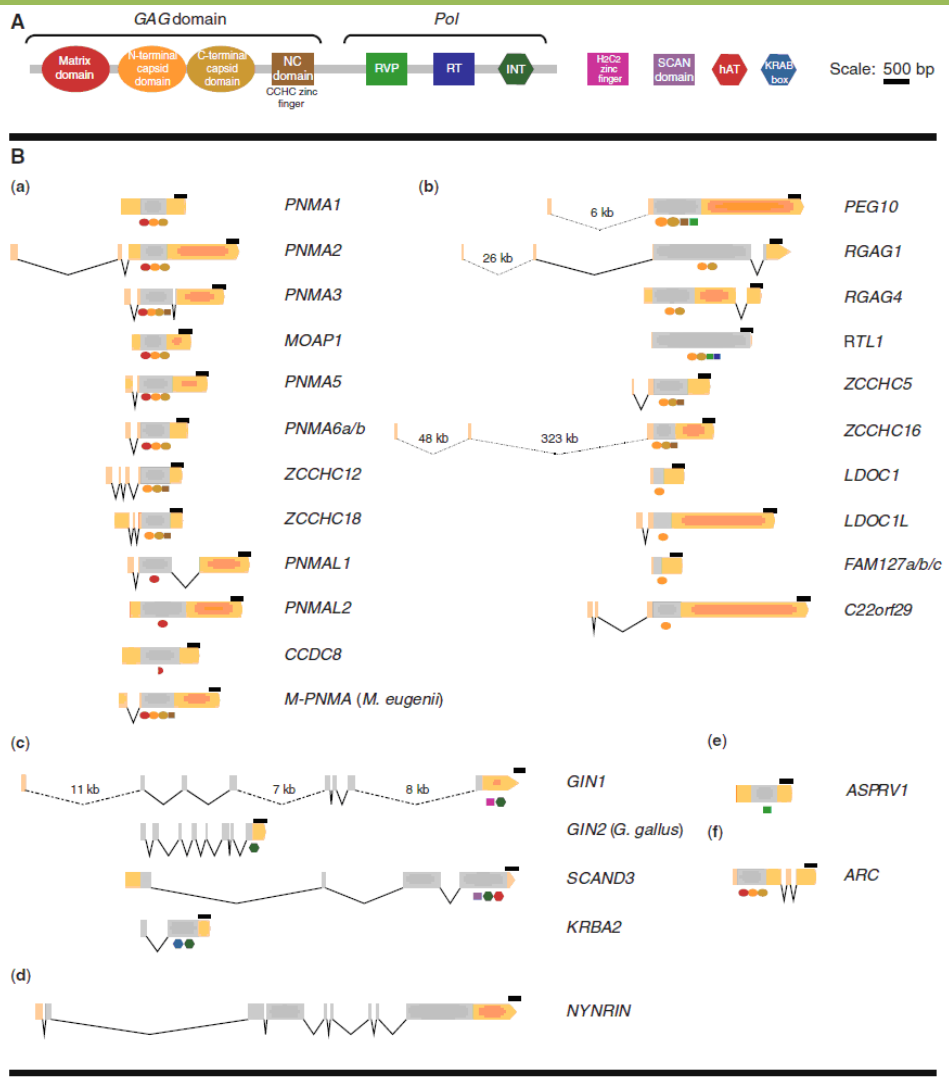
Gene structures of human domesticated genes

(A) Color-coded **protein domains** of Metaviridae retroelements and additional domains that are present in human domesticated genes.

(B) Exon–intron organization of human domesticated genes.

Exons are shown as boxes and **introns** as connecting lines. **Gray regions of exons** denote the protein-coding sequences, whereas **UTRs** are represented as orange boxes.

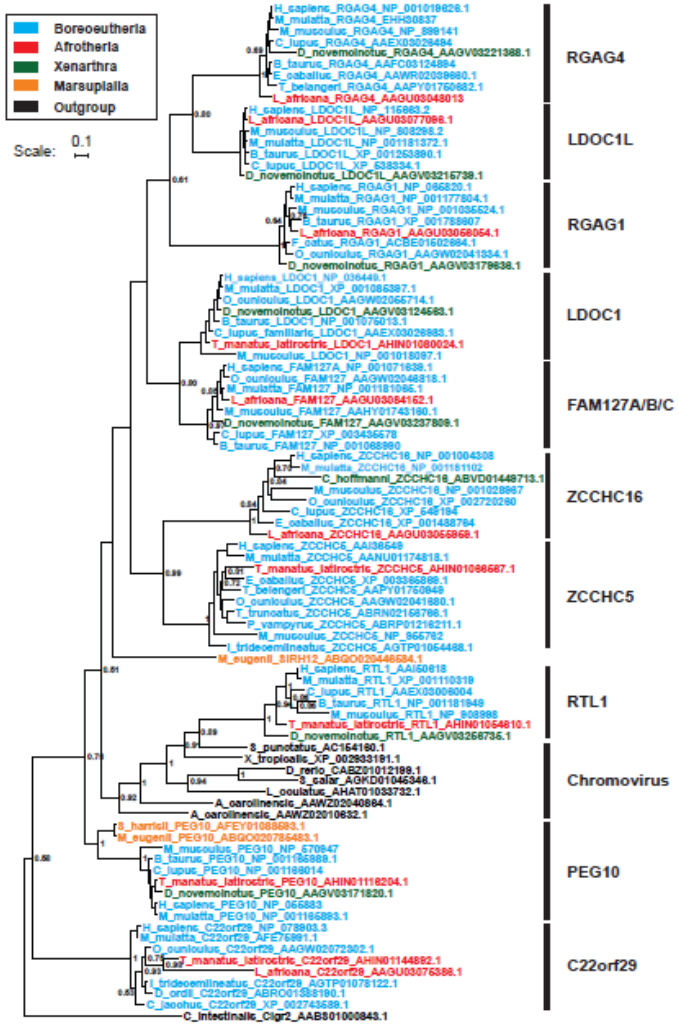
(C) Domesticated genes provided evidence for the **extensive intron gain in the ancestor of placental mammals** (Kordis, Biology Direct 2011).



PHYLOGENETIC ANALYSIS (SUSHI/Chromovirus FAMILY)

- Identification of **paralogs** and **orthologs**
- Identification of **ancestral retrotransposon sequences (progenitor sequences)**
- Helps to resolve the **evolutionary relationships within gene families**
- **Domesticated genes originated via several independent domestication events**

- MrBayes tree under Poisson+G4 model
- Alignment of the N-terminal capsid of gag domain.
- The scale bar corresponds to 0.1 substitutions per site.
- *C. intestinalis* Cigr2 retroelement was used to root this tree.

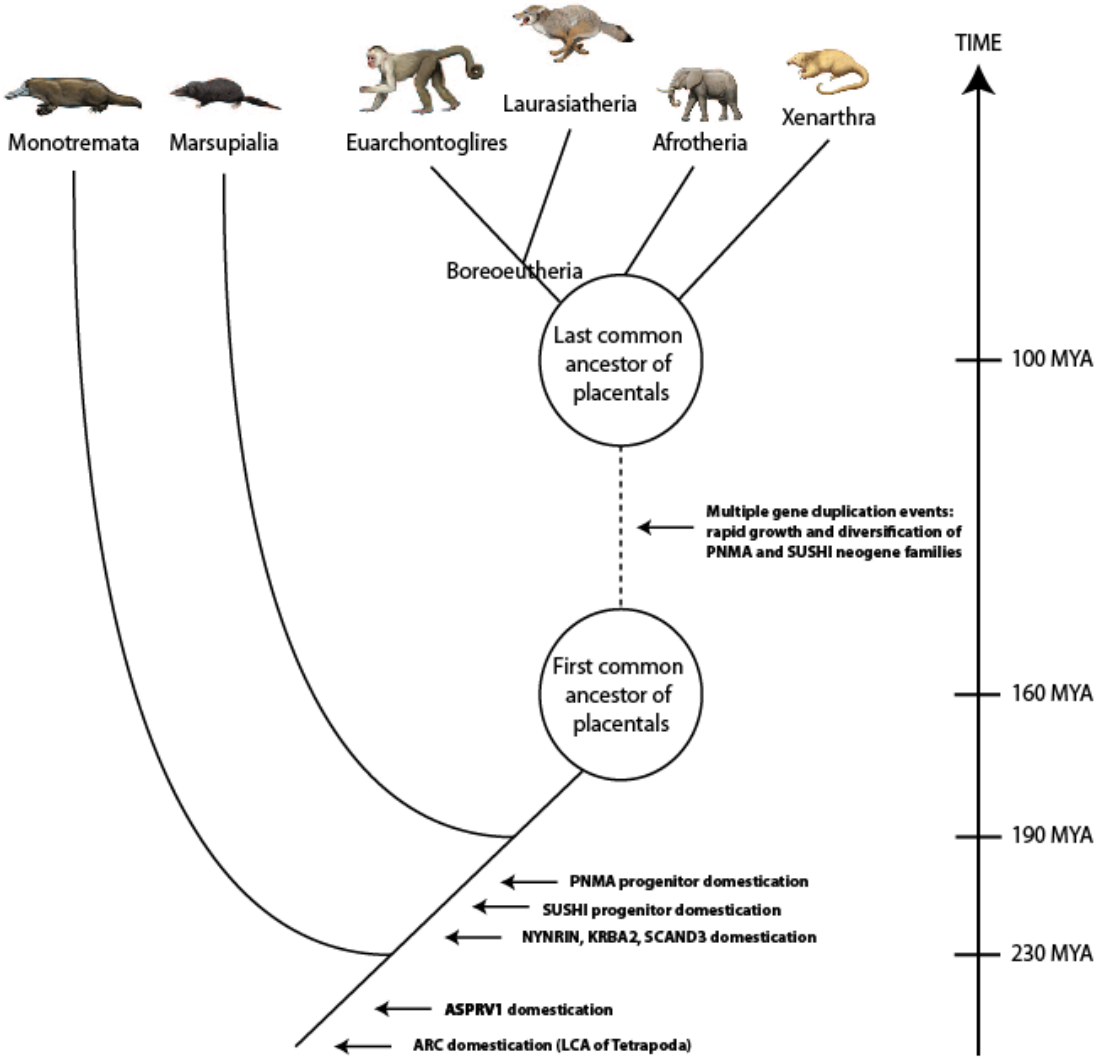


NUMEROUS INDEPENDENT ORIGINS OF DOMESTICATED GENES

Ancestral domain	Gene	LCA Chordata	LCA Tetrapoda	LCA mammals	LCA placentals
integrase					
	GIN1	●			
	GIN2	●			
	KRBA2			●	
	SCAND3			●	
	NYNRIN			●	
gag					
	ARC		●		
protease					
	ASPRV1			●	

Ancestral domain	Gene family	Gene	LCA Chordata	LCA Tetrapoda	LCA Theria	LCA placentals
gag	Sushi	RGAG1				●
		RGAG4				●
		PEG10			●	
		RTL1				●
		LDOC1				●
		LDOC1L				●
		FAM127				●
		C22orf29				●
		ZCCHC5				●
		ZCCHC16				●
		SIRH12			●	
gag	PNMA					
		PNMA1				●
		PNMA2				●
		MOAP1				●
		PNMA3				●
		PNMA5				●
		PNMA6A/B				●
		ZCCHC12				●
		ZCCHC18				●
		PNMAL1				●
		PNMAL2				●
		CCDC8				●
		M-PNMA			●	

GENESIS (BIRTH) OF DOMESTICATED GENES: BURST IN THE ANCESTOR OF PLACENTALS

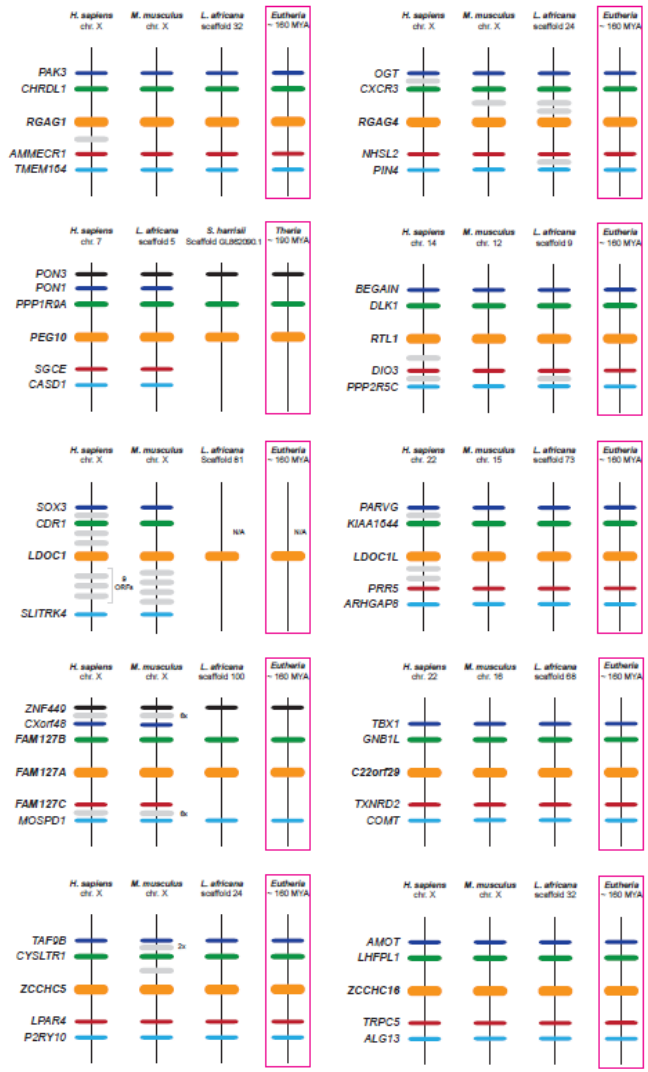


ANALYSIS OF CONSERVED SYNTENY

- From the analysis of syntenic positions, we established the **ancestral state of the genomic positions of the domesticated genes**

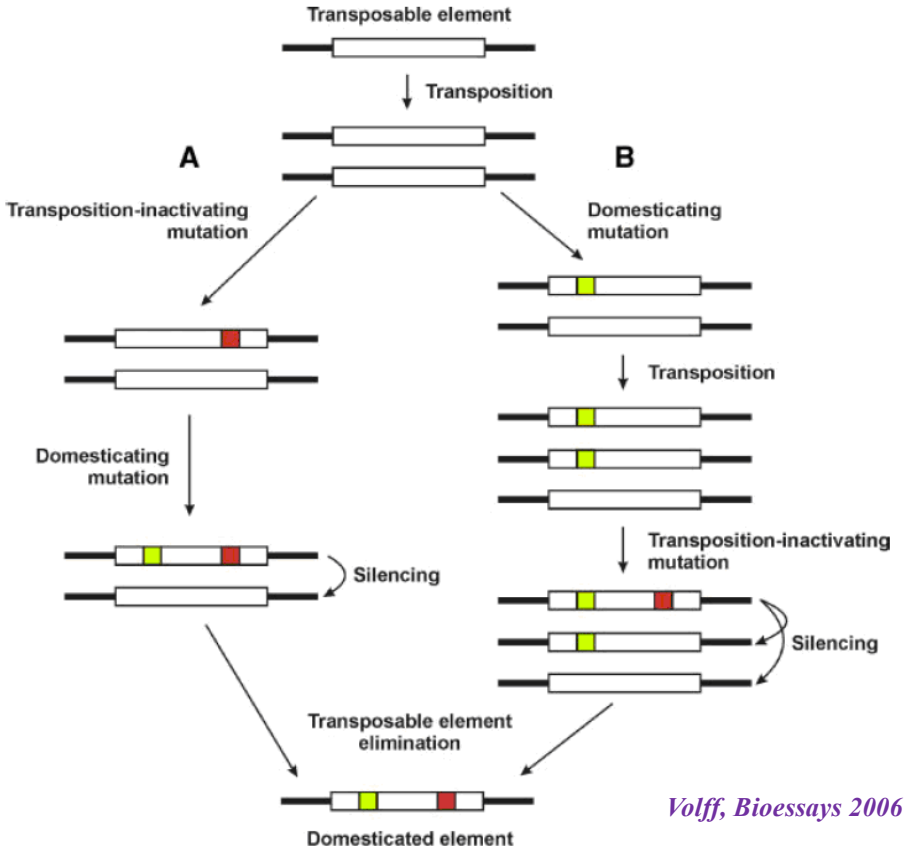
- Analysis of syntenic loci has enabled **clear orthology distinction** within domesticated genes and provides the **evidence for lineage-specific gene origins**

- Analysis of syntenic loci has shown that **diverse domesticated genes and their chromosomal positions were fully established in the ancestor of placental mammals**.



PROGENITORS OF DOMESTICATED GENES: ACTIVE RETROELEMENTS OR THEIR REMAINS?

Possible cascades of molecular events leading to the domestication of a transposable element



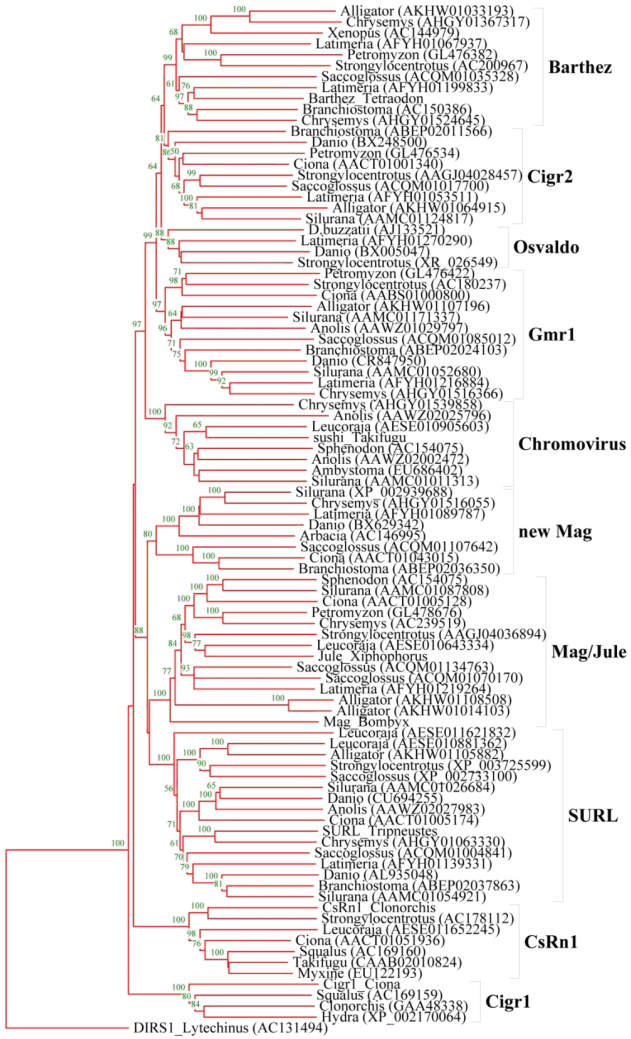
Volff, Bioessays 2006

Gag-like proteins might have caused the 'death' of their parental retrotransposons and might also protect the genome against infection by related viruses and retrotransposons [Lynch & Tristem, Curr. Biol. (2003)].

PHYLOGENETIC ANALYSIS OF METAVIRIDAE IN TETRAPODA

	Sauropsida				Mammals		Ancestral states			
Metaviridae clade	**Lepidosauria**	**Turtles**	**Archosauria (crocodiles)**	**Archosauria (birds)**	**Prototheria**	**Theria**	**Sauropsida**	**Synapsida**	**Amniota**	**Tetrapoda**
Chromovirus	+	*	*	Δ	Δ	Δ	+	+	+	+
Cigr2	Δ	Δ	+	Δ	Δ	Δ	+	+	+	+
Barthez	Δ	+	+	Δ	Δ	Δ	+	+	+	+
Gmr1	+	+	+	Δ	Δ	Δ	+	+	+	+
new Mag	Δ	+	Δ	Δ	Δ	Δ	+	+	+	+
SURL	+	+	+	Δ	Δ	Δ	+	+	+	+
Mag	+	+	+	Δ	Δ	Δ	+	+	+	+

+: full length elements present, Δ: elements absent, *: corrupted elements (retroelement remains)



Active Metaviridae progenitors of domesticated genes are still present in diverse reptilian genomes

Domesticated genes originated from transposable element remains

Domesticated genes were not responsible for the silencing of active Metaviridae lineages in mammals

As the synapsid ancestor also possessed a very rich TE repertoire, very similar to the sauropsids, this repertoire was extensively modified after the end Permian extinction in cynodont ancestors of modern mammals (Kordiš et al. 2006; Kordiš 2009).

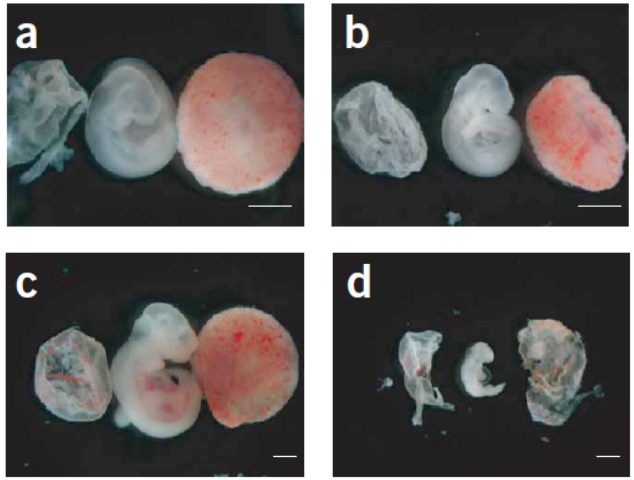
DGs AND THE ORIGIN OF DIVERSE PHENOTYPIC NOVELTIES

Neogene	Ancestral gene	Ancestral TE	Function
PEG10	*gag*	Sushi type LTR retrotransposon	apoptosis, cell proliferation, placenta formation
RTL1	*gag*	Sushi type LTR retrotransposon	feto-maternal interface, development of placenta
LDOC1	*gag*	Sushi type LTR retrotransposon	inhibition of NF-kappaB mediated response, potential tumor suppressor
MOAP1	*gag*	Barthez lineage LTR retrotransposon	BAX binding protein, modulator of apoptosis
GIN1	*integrase*	Gmr1 clade LTR retrotransposon	unknown
ASPRV1 (SASPase)	*protease*	Cigr2 clade LTR retrotransposon	expressed in epidermis, involved in prevention of wrinkle formation

Genesis of domesticated genes is connected to the origin of diverse phenotypic novelties in placental mammals

DOMESTICATED GENES ARE BECOMING ESSENTIAL GENES

WT **PEG10 knock-out**

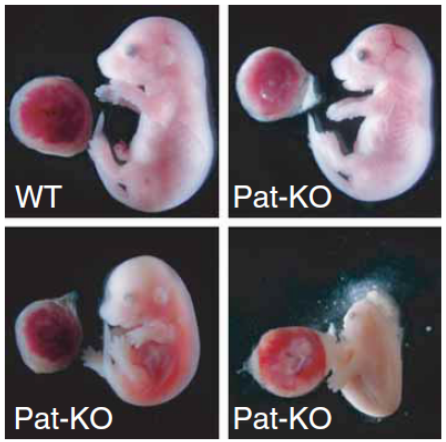


9.5 days

10.5 days

Deletion of *Peg10*, an imprinted gene acquired from a retrotransposon, causes early embryonic lethality

Ryuichi Ono[1,2], Kenji Nakamura[3], Kimiko Inoue[2,4], Mie Naruse[1], Takako Usami[5], Noriko Wakisaka-Saito[1,2,6,7], Toshiaki Hino[3], Rika Suzuki-Migishima[3], Narumi Ogonuki[4], Hiromi Miki[4], Takashi Kohda[1,2], Atsuo Ogura[2,4], Minesuke Yokoyama[2,3,8], Tomoko Kaneko-Ishino[2,7] & Fumitoshi Ishino[1,2]

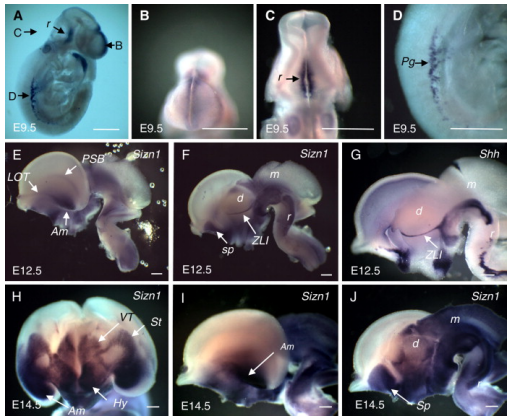


- **RTL1 plays an essential role in the development of the placenta and feto-maternal interface**

Role of retrotransposon-derived imprinted gene, *Rtl1*, in the feto-maternal interface of mouse placenta

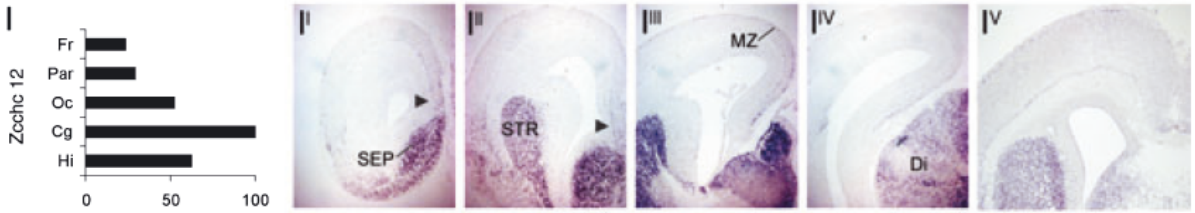
Yoichi Sekita[1], Hirotaka Wagatsuma[2], Kenji Nakamura[3], Ryuichi Ono[1], Masayo Kagami[4], Noriko Wakisaka[1,5], Toshiaki Hino[3], Rika Suzuki-Migishima[3], Takashi Kohda[1], Atsuo Ogura[6], Tsutomu Ogata[4], Minesuke Yokoyama[3,7], Tomoko Kaneko-Ishino[5] & Fumitoshi Ishino[1]

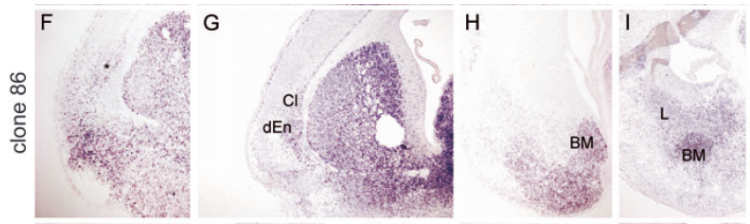
ROLE OF ZCCHC12/Sizn1 IN BRAIN DEVELOPMENT



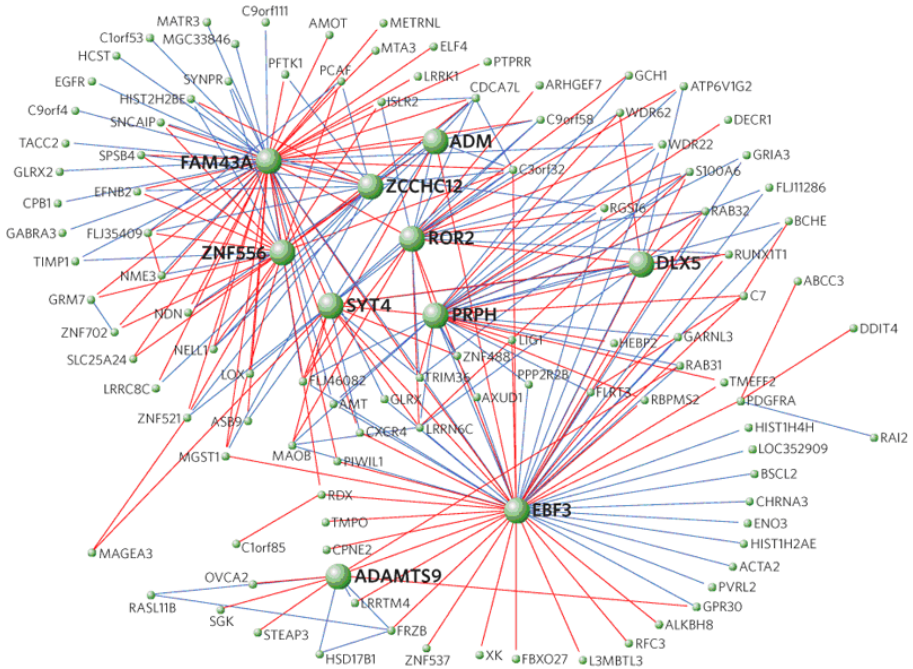
Cho et al., 2011

Genes that may be involved in the <u>cortical regionalization</u> during late neurogenesis in mouse: Differential gene expression across the embryonic cerebral cortex is assumed to play a role in the subdivision of the cortex into distinct areas with specific morphology, physiology and function.



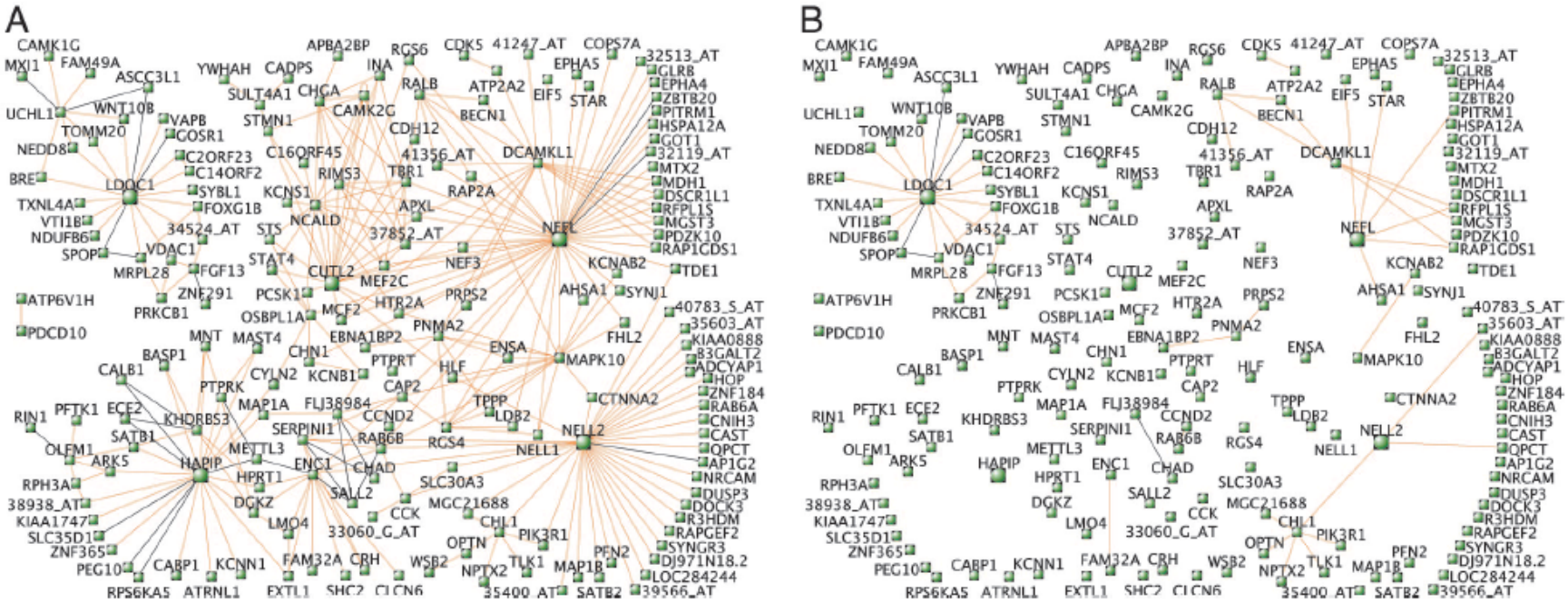


Muehlfriedel et al. 2007



Konopka et al. Nature 2009

DOMESTICATED GENES IN BRAIN INTERACTION NETWORKS



There is growing evidence that some domesticated genes (e.g.,**LDOC1**) are **involved in the gradual growth of CNS interaction networks in the particularly active regions of brain (neocortex)**—not only during the evolution of placentals, but in very recent times, that is, after the split of Homo and chimpanzee lineages.

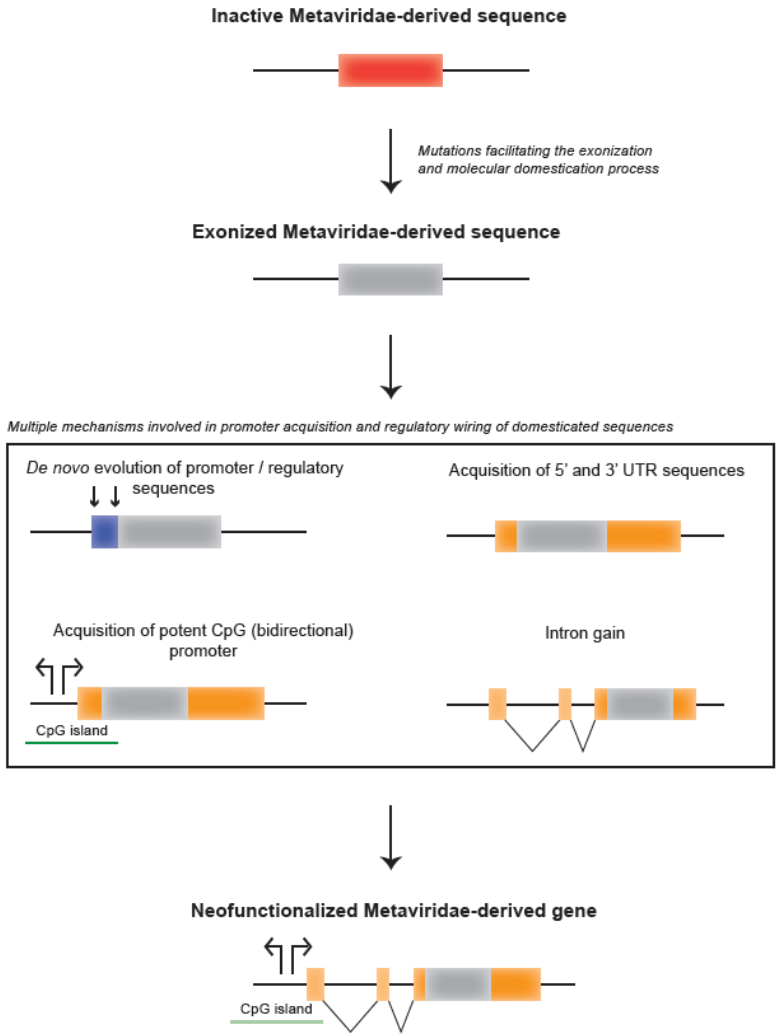
(B) Connections from A that are present in humans but absent in chimpanzees

LDOC1 (a human-specific hub) has been interrupted by an inversion in chimpanzees, effectively abolishing the entire "submodule" anchored by LDOC1 in cerebral cortex.

Oldham et al. PNAS 2006

MECHANISMS INVOLVED IN THE PROCESS OF NEOFUNCTIONALIZATION

- In the transition phase from retroelement remains to the first domesticated genes, **many nucleotide changes were necessary for the neofunctionalization**.
- One of the crucial steps in the process of neofunctionalization was the **exonization of retroelement domains (gag, protease, and integrase), which produced ready-to-use modules.**
- Retroelement remains in mammalian genomes will normally turn into pseudogenes, due to lack of a promoter, and **they can survive as a functional gene only if they recruit a new promoter sequence**.
- To become expressed at a significant level and in the tissues where it can exert a selectively beneficial function, a **new gene needs to acquire a core promoter and other structural elements that regulate its expression**.



REGULATORY EVOLUTION IN DOMESTICATED GENES

- Domesticated genes have acquired regulatory regions *de novo*
- *de novo* acquisition of promoters and 5'-UTRs in domesticated genes
- *de novo* acquisition of 3'-UTRs in domesticated genes
- Analysis of **transcription factor binding sites in promoters of domesticated genes** has shown a large diversity between genes or between human and mouse orthologous genes, indicating that ***cis*-regulatory evolution was responsible for the large differences in expression patterns of domesticated genes**.

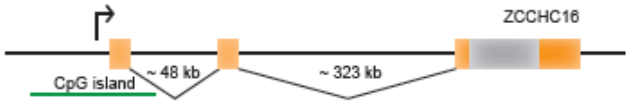
DIVERSE SOURCES OF PROMOTERS

a) *De novo* acquisition of promoter (CpG island-less promoter)

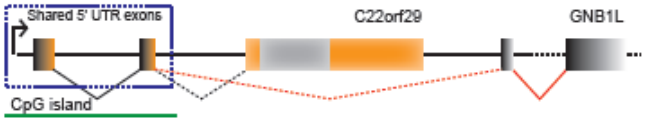
b) CpG-rich promoter

c) Bidirectional promoter (head to head gene pair)

d) Promoter capture via the evolution of 5' UTR exon/intron struct.



e) Shared promoter



de novo ACQUIRED PROMOTERS OF DOMESTICATED GENES

Table 1. Newly Recruited Promoters of RDDGs.

TE Progenitor	Gene Name	Presence of 5′-UTR Introns	CpG Island/Proto Promoter	Bidirectional Promoter	Not Associated with Promoter CpG Island
Chromovirus	*RGAG1*	Yes			•
	RGAG4	No	•		
	PEG10	Yes		•	
	RTL1	No			•
	LDOC1	No	•		
	LDOC1L	Yes	•		
	FAM127A/B/C	No	•		
	C22orf29	Yes	•		
	ZCCHC5	Yes			•
	ZCCHC16	Yes	•		
Barthez	*PNMA1*	No	•		
	PNMA2	Yes		•	
	MOAP1	Yes		•	
	PNMA3	Yes	•		
	PNMA5	Yes			•
	PNMA6A/B	Yes	•		
	ZCCHC12	Yes	•		
	ZCCHC18	Yes	•		
	PNMAL1	Yes	•		
	PNMAL2	No	•		
	CCDC8	No	•		
Gmr1	*GIN1*	Yes		•	
	GIN2	No		•	
	KRBA2	Yes			•
	SCAND3	No			•
Osvaldo	*ARC*	No	•		
Cigr2	*ASPRV1*	No			•
ERV	*NYNRIN*	Yes	•		

NOTE.—The type of promoter is marked with the black dot.

cis-REGULATORY EVOLUTION IN DOMESTICATED GENES

Chromovirus-related RDDGs

LDOC1:
Homo: AP-1, ATF-2, AP-2gamma, FOXL1, Egr-4, COMP1, AP-2beta, c-Jun, AP-2alpha, AP-2alphaA
Mus: Elk-1, Meis-1, Meis-1b, HOXA9, HOXA9B, Nkx5-1, SEF-1(1), p300, POU3F2, Nkx6-1

LDOC1L:
Homo: NF-1, Tal-1, CUTL1, Tal-1beta, E47, NRF-2, COMP1, STAT3, ITF-2
Mus: RFX1, PPAR-gamma1, PPAR-gamma2, HOXA9, HOXA9B, Meis-1, Meis-1a, c-Myb, Meis-1b, POU3F2

RGAG4:
Homo: E2F-4, E2F-3a, E2F-5, E2F-2, Egr-4, MZF-1, E2F, E2F-1, COMP1, Pax-4a
Mus: XBP-1, Cdc5, IRF-1, PPAR-gamma1, PPAR-gamma2, HTF, CUTL1, NCX/Ncx, HEN1, GR

FAM127A:
Homo: Elk-1, Nkx5-1, c-Ets-1
Mus: -

FAM127C:
Homo: Elk-1, Nkx5-1, c-Ets-1
Mus: -

PEG10:
Homo: C/EBPbeta
Mus: -

ZCCHC5:
Homo: Sox5, PPAR-gamma1, Olf-1, FOXO1a, PPAR-gamma2, HNF-3beta, Arnt, FOXO1
Mus: POU3F2, POU3F2 (N-Oct-5a), POU3F2 (N-Oct-5b), POU2F1, POU2F1a, Bach1, Pax-6, Brachyury, Evi-1, Roaz

ZCCHC16:
Homo: Sp1, RREB-1, NF-YA, HNF-3beta, NF-YB, CBF-A, CBF-B, CP1A, NF-Y, CBF(2)
Mus: PPAR-gamma1, PPAR-gamma2, C/EBPalpha, LCR-F1, RREB-1, HNF-4alpha1, HNF-4alpha2, FOXD1, NF-E2, NF-E2 p45

RTL1:
Homo: AML1a, Pax-5, Olf-1, MyoD, E4BP4, C/EBPalpha, FOXJ2 (long isoform), FOXJ2
Mus: FOXC1, ARP-1, HFH-1, RP58, STAT1, STAT1alpha, STAT1beta, STAT2, STAT3, STAT4

PNMA family

PNMA1:
Homo: E2F-3a, E2F-4, E2F-5, Brachyury, HSF1 (long), E2F-2, E2F-1, E2F, HSF1short, ATF
Mus: ZIC2/Zic2, Roaz, Nkx3-1v1, Nkx3-1v2, Nkx3-1v3, Nkx3-1v4, Nkx3-1, Msx-1, Zic1, RFX1

PNMA2:
Homo: AML1a, Pax-5, MyoD, Lmo2, AP-4, GATA-1, Egr-4, FOXL1, HEN1
Mus: NRSF form1, NRSF form2, GATA-1, ITF-2, Tal-1beta, HSF1 (long), HSF1 (short), YY1, MyoD, C/EBPalpha

PNMA3:
Homo: E2F-3a, E2F-4, E2F-5, SREBP-1c, E2F-2, SREBP-1b, E2F-1, E2F, SREBP-1a, HOXA5
Mus: ATF2, CRE-BP1, ATF, Ik-3, c-Jun, RP58, NF-kappaB1, Roaz, ISGF-3, p53

MOAP1:
Homo: AML1a, ATF-2, NF-kappaB, FOXL1, AREB6, IRF-2, NF-kappaB2, Meis-1a, NF-kappaB1, RSRFC4
Mus: Evi-1, ZID, c-Myb, STAT3, POU3F1, FOXJ2, FOXJ2 (long isoform), Nkx3-1, Nkx3-1 v1, Nkx3-1 v2

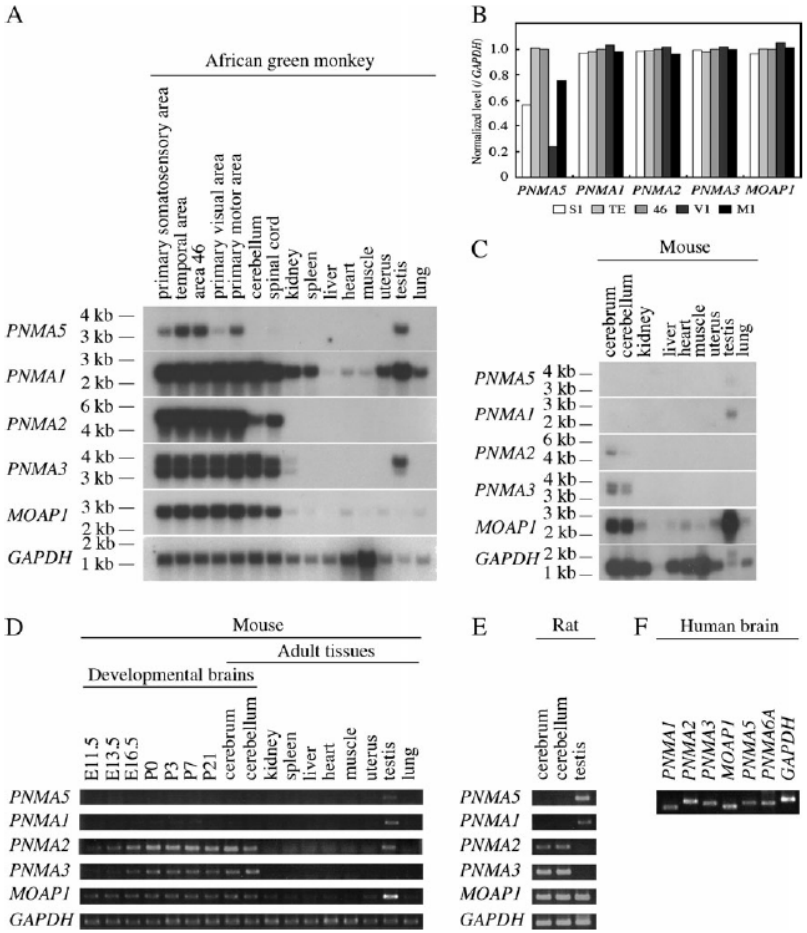
PNMA5:
Homo: COUP-TF1, LHX3b/Lhx3b, C/EBPbeta, AML1a, HNF-4alpha2, HNF-4alpha1, GATA-2, COUP-TF, CP2, LHX3a/Lhx3a
Mus: Nkx6-1, Evi-1, HSF1 short, HSF1 (long), Nkx2-5, NF-kappaB1, Pax-2, Pax-2a, SREBP-1a, SREBP-1b

Analysis of **TFBS in promoters** of domesticated genes has shown a **large diversity between genes or between human and mouse orthologous genes**, indicating that ***cis*-regulatory evolution was responsible for the large differences in expression patterns of domesticated genes**.

CONSEQUENCES OF THE *cis*-REGULATORY EVOLUTION: NEW EXPRESSION PATTERNS OF DOMESTICATED GENES

Tissue, species, spatiotemporal and development specific patterns in the expression of domesticated genes.

Expression analysis of the PNMA family genes in various tissues by northern hybridization and RT-PCR.



Takaji et al. (2009)

de novo ACQUISITION OF 3'UTRs IN DOMESTICATED GENES

de novo acquired 3'-UTRs in placental mammals show **large variation in length**, the shortest being present in the human SCAND3 gene (281 bp) and the longest in the mouse PEG10 gene (5,166 bp).

The great majority of human or mouse domesticated genes have long 3'-UTRs, eight are shorter than 1,000 bp, seven are in the range of 1,000 to 2,000 bp, six in the range of 2,000 to 3,000 bp, one is longer than 4 kb, and two longer than 5 kb.

What is the reason for such increased lengths of the 3'-UTRs of domesticated genes?

Searching for TEs in the unusually long 3'-UTRs with RepeatMasker has shown the **absence of species-specific repeats in the analyzed species**.

The consequence of the very long 3'-UTRs in some domesticated genes is that the **lengths of the 3' exons are greatly increased.**

Because **all the domesticated genes have recruited adjacent sequences as their 3'-UTRs**, these *de novo* acquired 3'-UTRs may play an important role in establishing **new regulatory functions.**

Table 2. Newly Recruited 3'-UTRs of RDDGs: 3'-UTR Lengths (in bp) in Human and Mouse RDDGs in Connection with Their Expression Profiles.

Gene Name	*Homo sapiens*	Expression Profile	*Mus musculus*	Expression Profile
RTL1	>2,000	TS	N/A	TS
PEG10	5,161	HK	5,166	TS
LDOC1L	4,267	HK	3,064	HK
C22ORF29	5,029	HK	N/A	N/A
ZCCHC5	924	TS	911	TS
ZCCHC16	1,584	TS	1,912	TS
RGAG4	2,268	HK/TS	2,521	HK/TS
LDOC1	836	HK	785	TS
Fam 127a	821	HK	N/A	
Fam127b	835	HK	N/A	
Fam127c	1,614	HK/TS	N/A	
RGAG1	1,013	TS	N/A	TS
PNMA1	795	HK	735	TS
PNMA2	2,981	TS	2,835	TS
PNMA3	2,023	TS	1,969	TS
MOAP1	991	HK	2,400	TS
PNMA5	1,428	TS	394	TS
PNMA6a	754	TS	N/A	
PNMAL1	2,070	HK	284	TS
PNMAL2	2,367	HK/TS	1,750	TS
ZCCHC12	515	TS	518	TS
ZCCHC18	519	TS	1,104	TS
CCDC8	865	HK/TS	N/A	TS
ARC	1,551	TS	1,668	TS
GIN1	1,898	HK	400	HK/TS
KRBA2	479	TS	N/A	
SCAND3	281	TS	N/A	
NYNRIN	1,842	HK	1,776	HK/TS
ASPRV1	568	TS	517	TS

NOTE.—HK, housekeeping gene; TS, tissue-specific gene; N/A, not available. Underlined expression profiles reflect the change of the expression profile between human and mouse orthologous genes.

CONCLUSIONS

- **We have mapped the life history of domesticated genes, from birth, their fixation in the genome, gain of regulatory elements and structural complexity to complete integration into the functional network of the cell.**
- We have provided direct evidence for the **main diversification of domesticated genes having occurred in the ancestor of placental mammals**.
- These **domesticated genes have originated in several steps** by **independent domestication events** and later diversified by gene duplications.
- We have demonstrated that **placental mammal-specific domesticated genes originated from retroelement remains.**
- Analysis of syntenic loci has shown that **diverse domesticated genes and their chromosomal positions were fully established in the ancestor of placental mammals**.
- During the domestication process, ***de novo* acquisition of regulatory regions is a prerequisite for survival of the domesticated genes**. The findings of this study thus provide a new view on the **origin and evolution of *de novo* acquired promoters, 5'- and 3'-UTRs, in diverse mammalian domesticated genes**.
- **The regulatory wiring of domesticated genes and their rapid fixation in the ancestor of placental mammals have played an important role in the origin of their innovations and adaptations, such as placenta and newly evolved brain functions**.
- Domesticated genes could thus constitute an **excellent system on which to analyze the mechanisms of regulatory evolution in placental mammals**.