Phylogenomics
(Phylogenetics based on genome-scale data)

Could genome-scale data easily resolve phylogenetic problems?

→ Not necessarily!

Phylogenomics

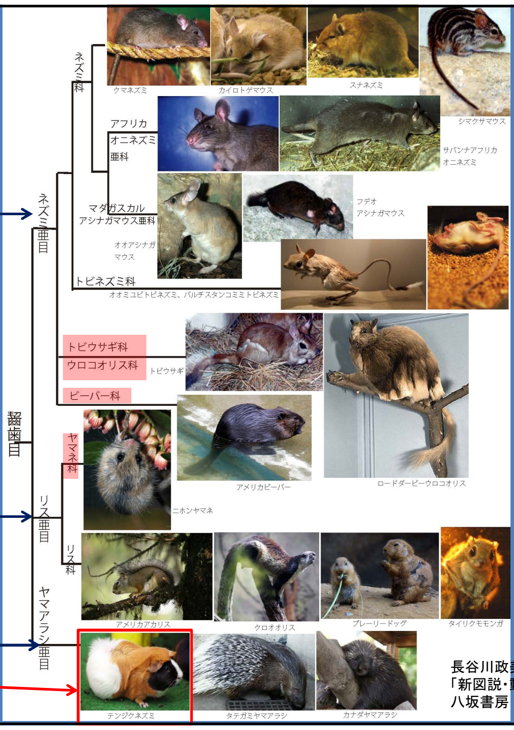
- The longer the sequences, the smaller the sampling error becomes.
  → Apparently strong conclusion can be obtained.
- However, if the estimation is biased, an erroneous tree can be supported with a high confidence. → Bias of tree estimation caused by a model misspecification is an important problem in phylogenetics.

Systematic errors in phylogenetic inference

- Long-branch-attraction (LBA)

Tree of Rodentia

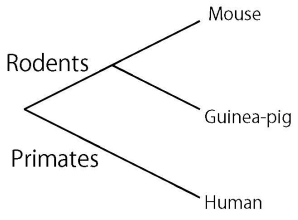
- Myomorpha
- Sciuromorpha
- Hystricomprpha
  Guinea-pig



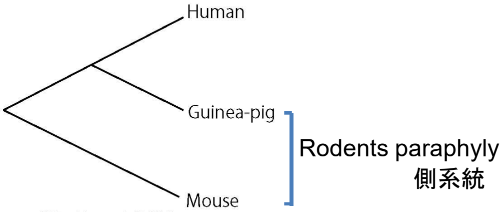
長谷川政美(2011)
「新図説・動物の起源と進化」
八坂書房

Rodents polyphyly?（齧歯類は多系統か？）

Tree-1 (Traditional)

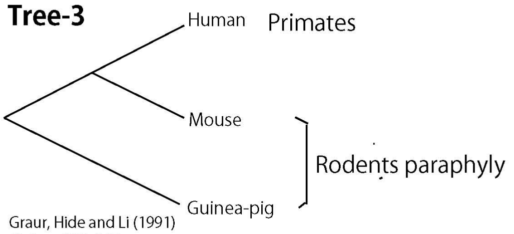


Tree-2



D'Erchia et al. (1996)
Nature 381, 597

Tree-3

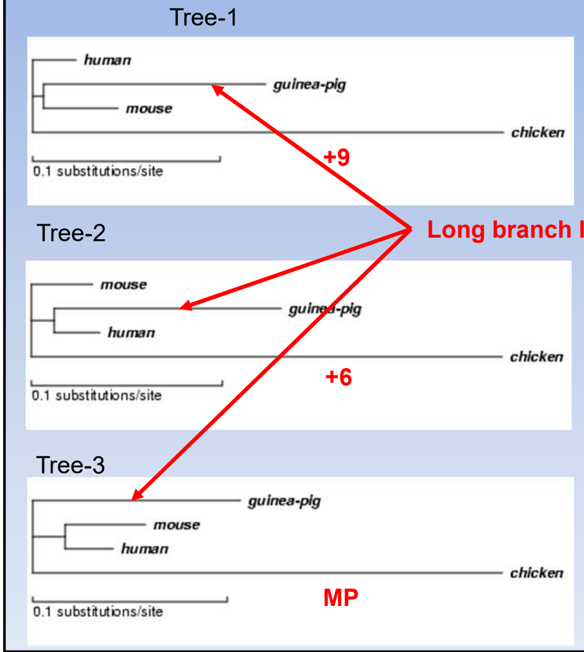


Graur, Hide and Li (1991)
Nature 351, 649

	Parsimony（節約法）			**Maximum-likelihood（最尤法）**		
	Tree-1	Tree-2	Tree-3	Tree-1	Tree-2	Tree-3
Crystalli	MP	2	3	ML	-0.4	-5.6
Lactalbu	5	MP	4	ML	-4.2	-5.5
Hbα	4	8	MP	-5.6	-10.5	ML
Hbβ	1	5	MP	-0.5	-3.3	ML
NGF	1	5	MP	ML	-9.1	-2.2
Factor9	1	MP	3	-4.0	ML	-3.8
Ribonucl	3	MP	2	-3.5	ML	-3.0
Insulin	1	1	MP	-1.6	ML	-1.6
LipLipas	9	6	MP	-3.1	ML	-0.2
Lipocort	4	5	MP	-8.2	-10.5	ML
total	17	20	MP	-4.7 ±16.0	-16.1 ±14.0	ML
BP	0.0303	0.0108	0.9589	0.3608	0.0515	0.5877

Hasegawa et al. (1992) *Nature* 355, 595

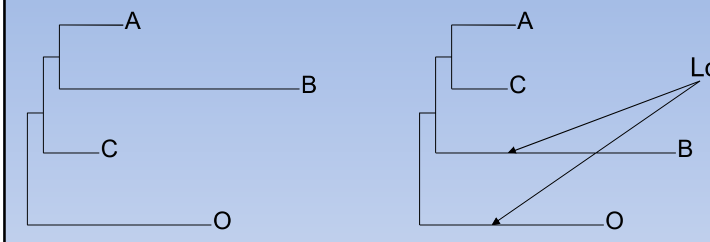
ML analysis of lipoprotein lipase (JTT-F model)



Long branch leading to guinea-pig

	$\Delta \ln L$ Without Γ	With Γ
Tree-1	-1.5± 3.3	-0.3±2.1
Tree-2	-0.2±3.9	<-2,144.8>
Tree-3	<-2,159.6>	-0.6±1.8

Long branch attraction artifact

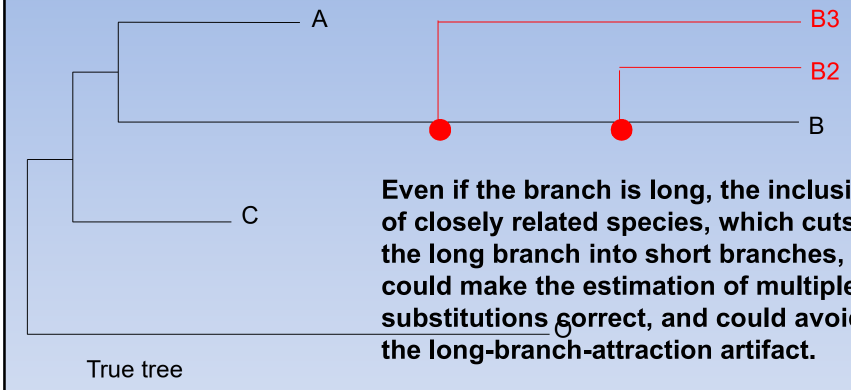


Long branch attraction
(Felsenstein 1978)

True tree
真の系統樹

Estimated tree
推定された系統樹

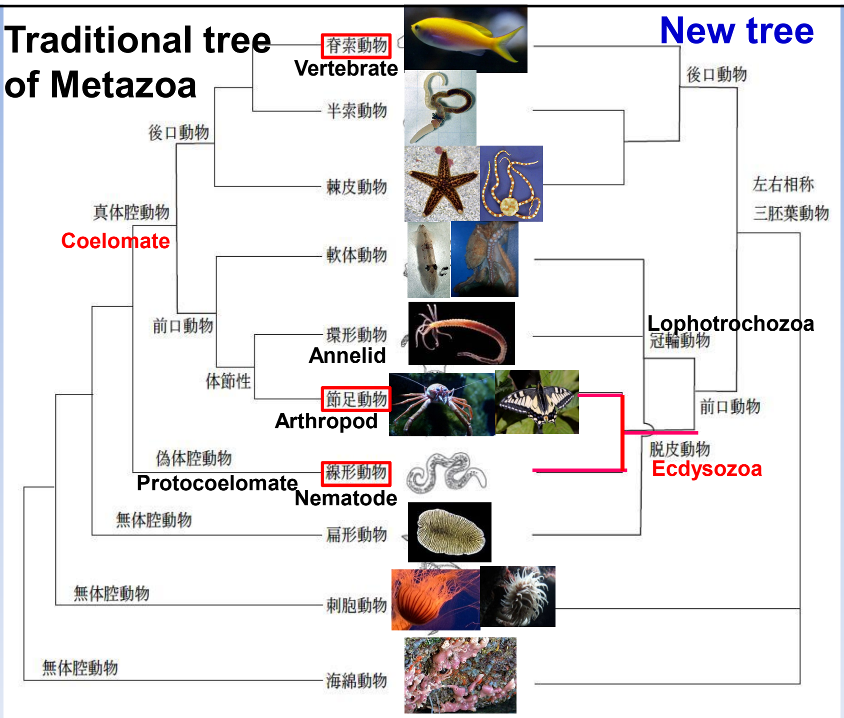
The longer the branch with many substitutions, the more extreme the underestimation of multiple-substitutions becomes, and accordingly two long branches tend to make a cluster erroneouly. The parsimony method ignores the multiple substitutions, and therefore the method is affected most seriouly. In the case of ML, when a simple model is used, the multiple substitutions are underestimated, but as a more realistic model is applied and as the species sampling becomes denser, the estimation becomes better.

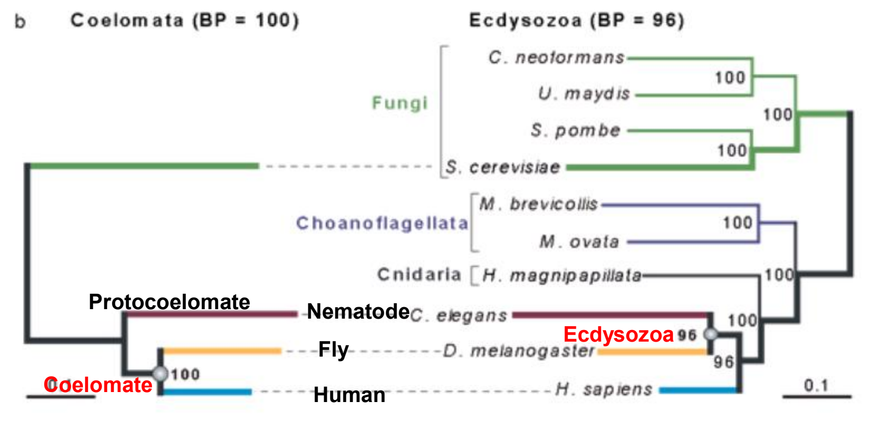


Even if the branch is long, the inclusion of closely related species, which cuts the long branch into short branches, could make the estimation of multiple substitutions correct, and could avoid the long-branch-attraction artifact.

Traditional tree of Metazoa

New tree





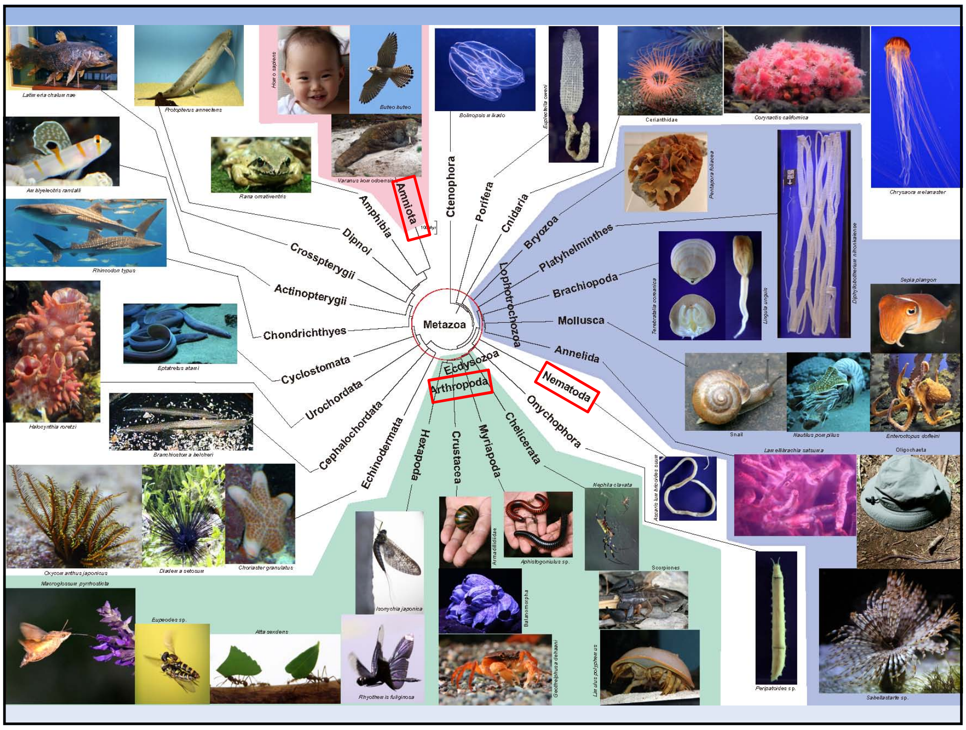
Blair et al. (2002)

Philippe et al. (2005)

ML tree with the JTT+Γ model

Long-branch-attraction

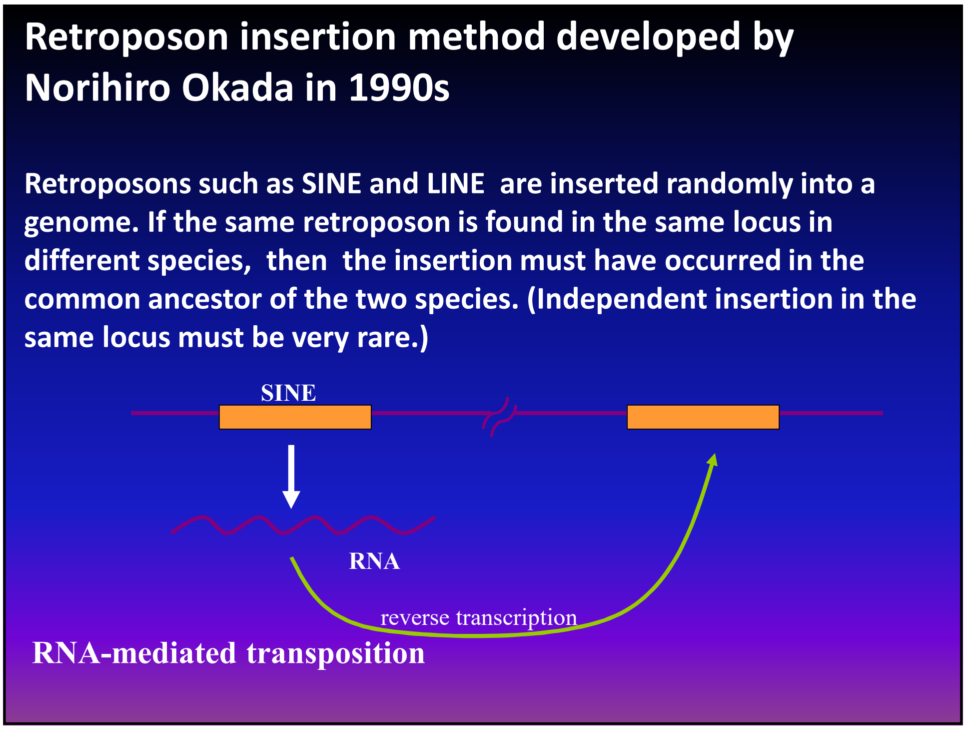
Delsuc, Brinkmann & Philippe (2005)



Systematic errors in phylogenetic inference

- Long-branch-attraction (LBA)
- **Heterogeneous tempo & mode of evolution among different genes**

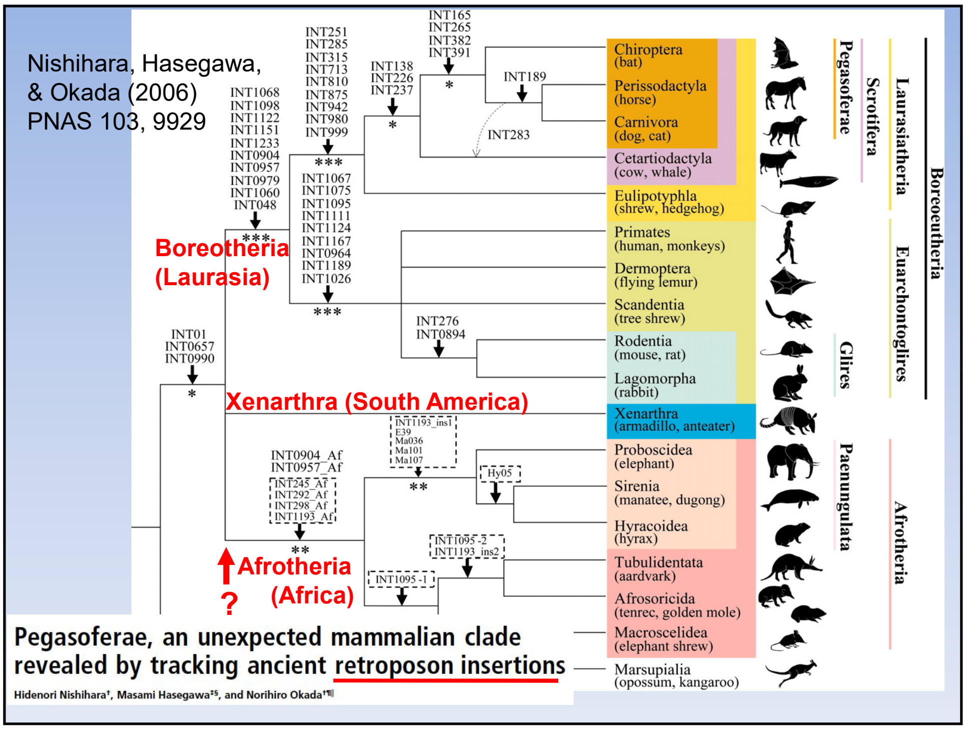
Diversity of Eutheria
真獸類的多樣性



Retroposon insertion method developed by Norihiro Okada in 1990s

Retroposons such as SINE and LINE are inserted randomly into a genome. If the same retroposon is found in the same locus in different species, then the insertion must have occurred in the common ancestor of the two species. (Independent insertion in the same locus must be very rare.)

RNA-mediated transposition

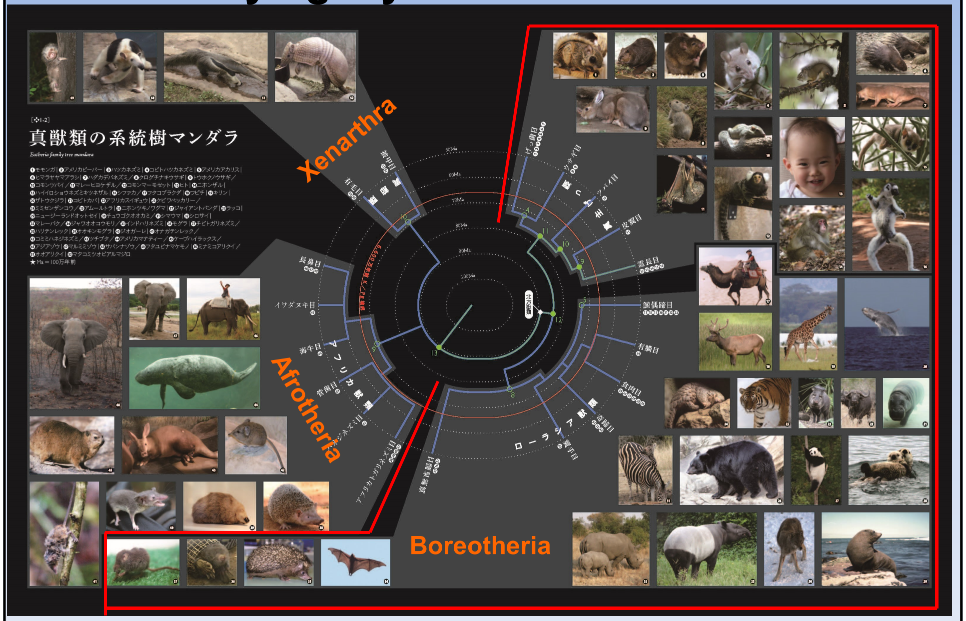


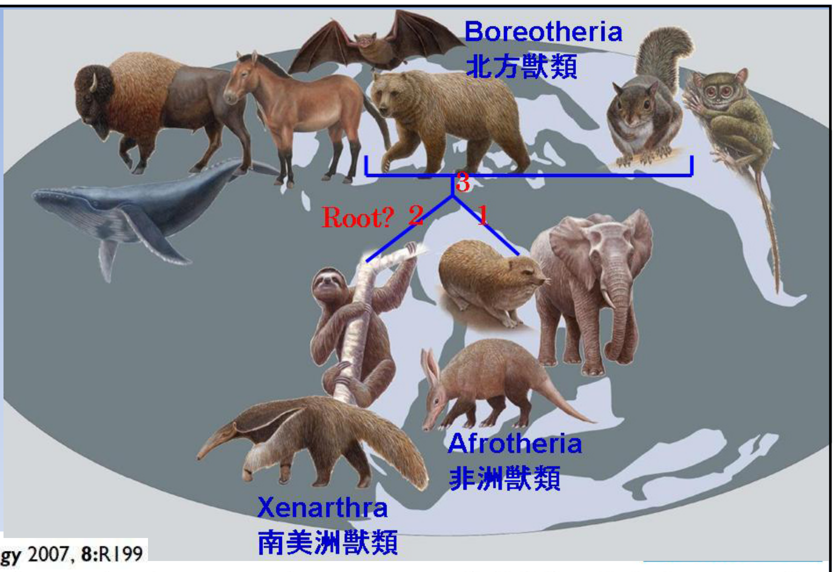
Nishihara, Hasegawa, & Okada (2006) PNAS 103, 9929

Pegasoferae, an unexpected mammalian clade revealed by tracking ancient retroposon insertions

Hidenori Nishihara†, Masami Hasegawa‡§, and Norihiro Okada†¶‖

Phylogeny of Eutheria





Genome ***Biology*** 2007, **8:**R199

Rooting the eutherian tree: the power and pitfalls of phylogenomics

Hidenori Nishihara[*†], Norihiro Okada[*] and Masami Hasegawa[†‡]

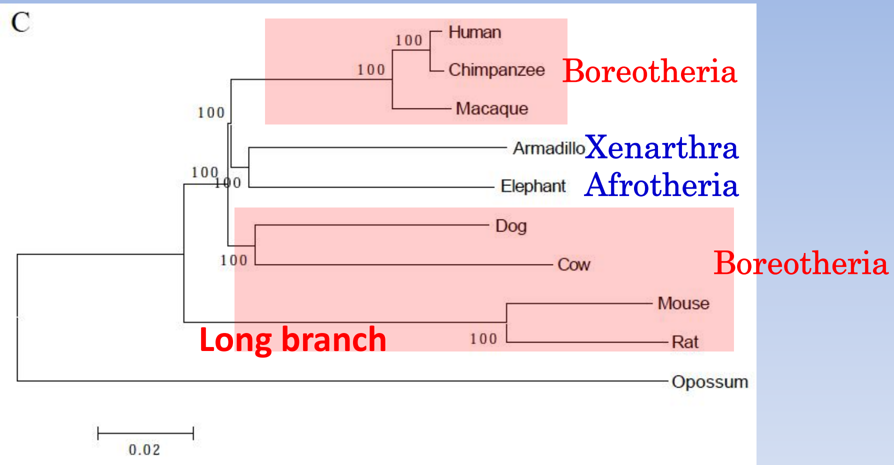
Addresses: *Graduate School of Bioscience and Biotechnology, Tokyo Institute of Technology, 4259-B-21 Nagatsuta-cho, Midori-ku, Yokohama 226-8501, Japan. †Department of Statistical Modeling, Institute of Statistical Mathematics, 4-6-7 Minami-Azabu, Minato-ku, Tokyo 106-8569, Japan. ‡School of Life Sciences, Fudan University, Handan Road 220#, Shanghai 200433, China.

Collection of exon sequences from database

1. Extraction of all exon sequences >200 bp from the human genome database
2. Removal of duplicated (paralog) sequences from the human data
3. Search of the armadillo (Xenarthra) and elephant (Afrotheria) genomic data for homologs of the human exons
4. Collection of the homologous exons from other mammalian genomic data
5. Alignment of all the sequences and removal of ambiguous nucleotide sites

→ 1,011,870 bp of 2,789 protein-encoding genes from 9 eutherian species and 1 marsupial (opossum)

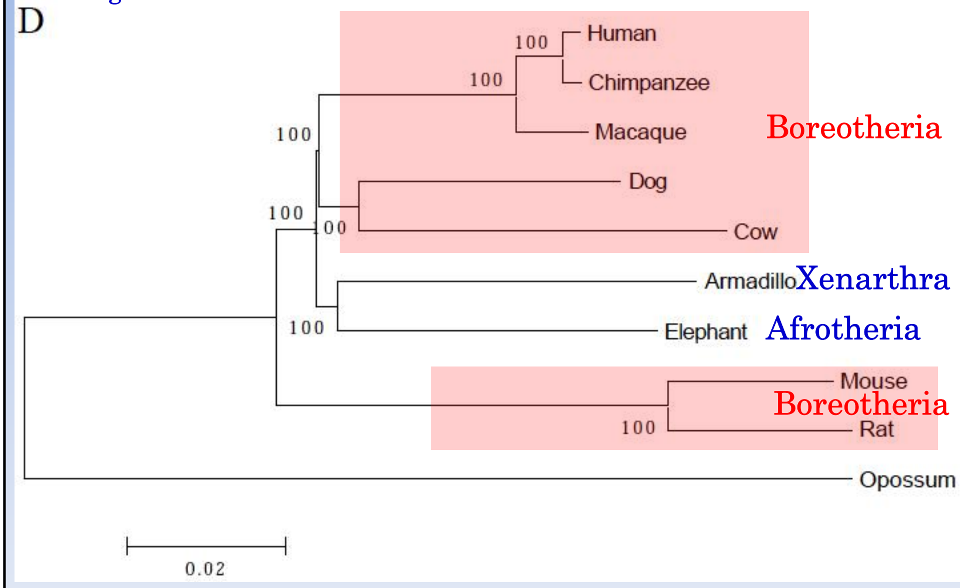
NJ tree of concatenated 1,011,870 nucleotides

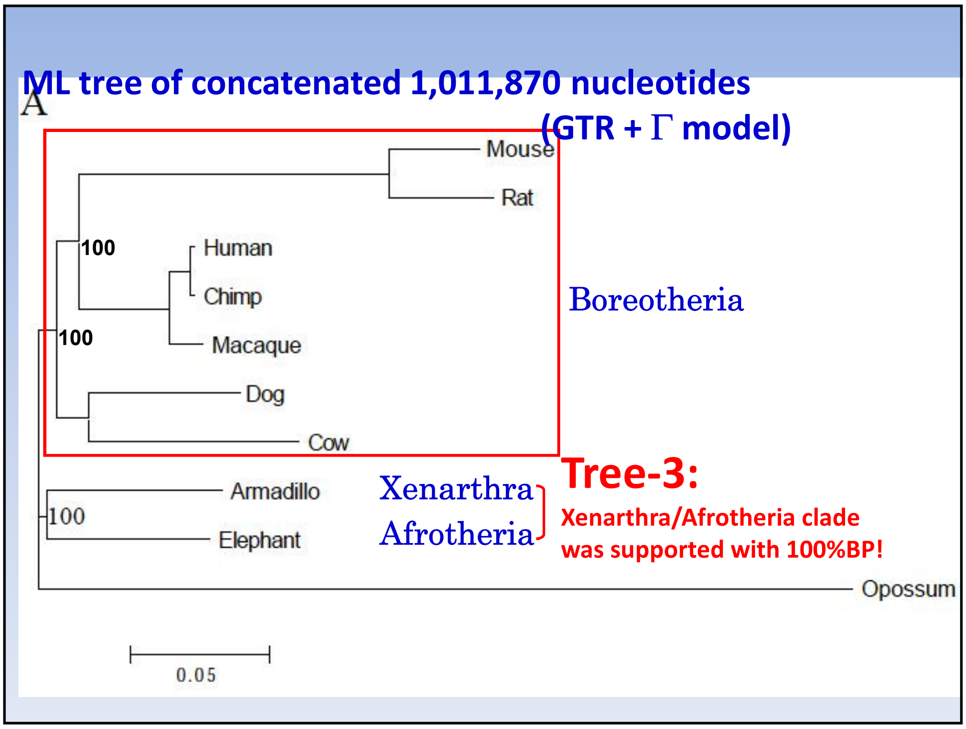


Boreotheria monophyly (supported strongly by the retroposon insertion and ML analyses) was not reproduced by the NJ and MP analyses. This is probably due to the long-branch-attraction artifact of the long mouse/rat lineage.

NJ tree of concatenated 337,290 amino acids

Even when amino acid sequences are used, the NJ and ML give a wrong tree.





ML tree of concatenated 1,011,870 nucleotides (GTR + Γ model)

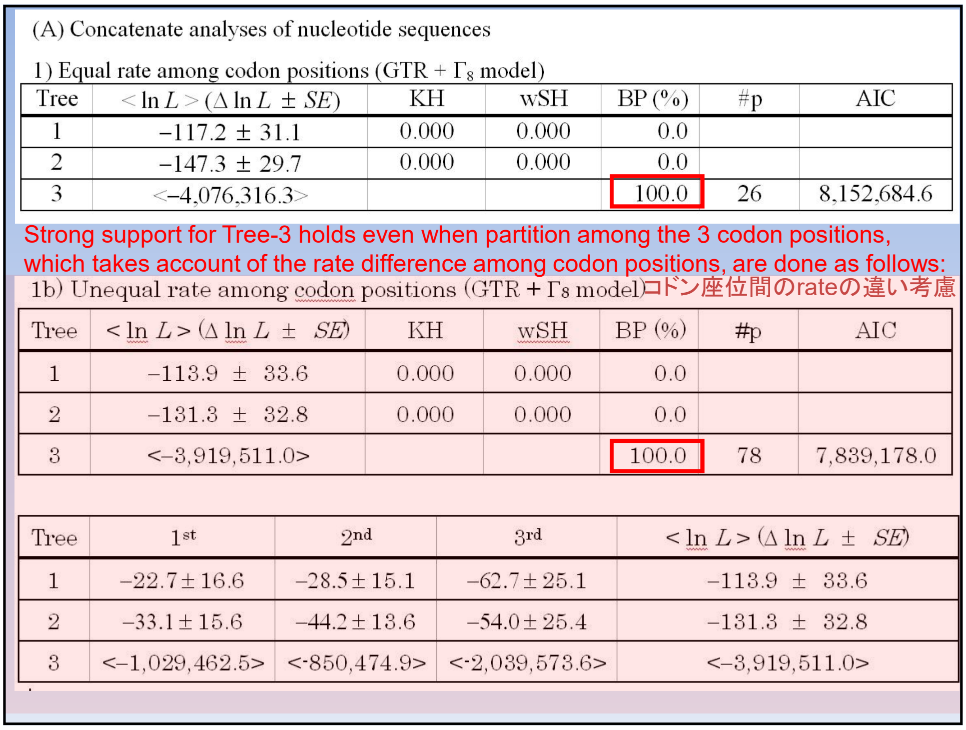
Boreotheria

Xenarthra

Afrotheria

Tree-3:

Xenarthra/Afrotheria clade was supported with 100%BP!



(A) Concatenate analyses of nucleotide sequences

1) Equal rate among codon positions (GTR + Γ_8 model)

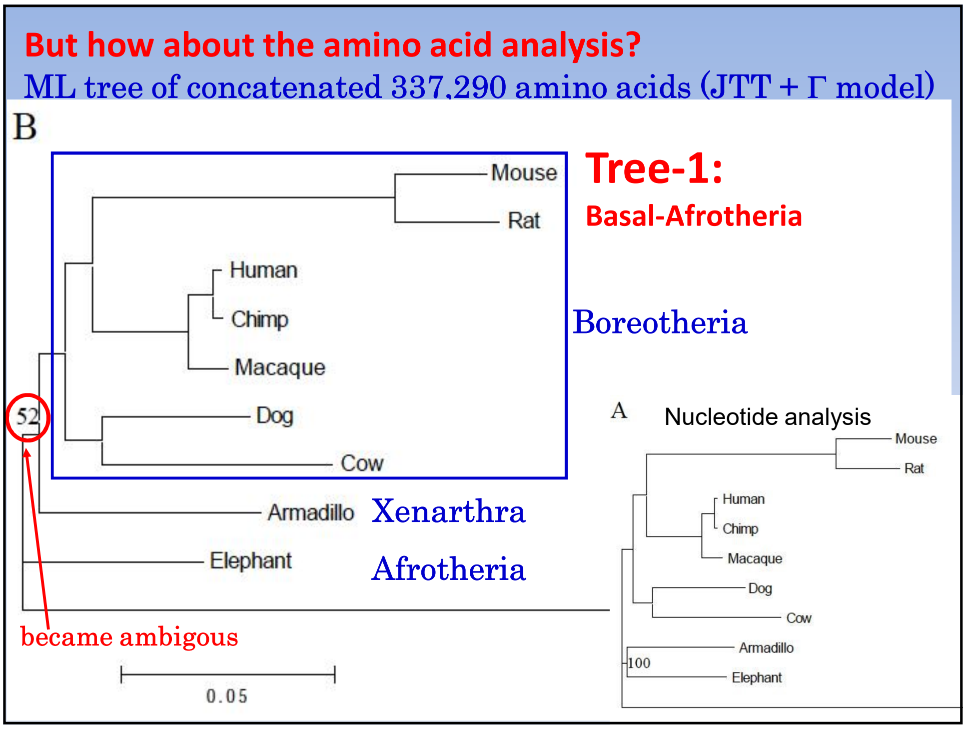
Tree	< ln *L* > (Δ ln *L* ± *SE*)	KH	wSH	BP (%)	#p	AIC
1	−117.2 ± 31.1	0.000	0.000	0.0		
2	−147.3 ± 29.7	0.000	0.000	0.0		
3	<−4,076,316.3>			100.0	26	8,152,684.6

Strong support for Tree-3 holds even when partition among the 3 codon positions, which takes account of the rate difference among codon positions, are done as follows:

1b) Unequal rate among codon positions (GTR + Γ_8 model)コドン座位間のrateの違い考慮

Tree	< ln *L* > (Δ ln *L* ± *SE*)	KH	wSH	BP (%)	#p	AIC
1	−113.9 ± 33.6	0.000	0.000	0.0		
2	−131.3 ± 32.8	0.000	0.000	0.0		
3	<−3,919,511.0>			100.0	78	7,839,178.0

Tree	1st	2nd	3rd	< ln *L* > (Δ ln *L* ± *SE*)
1	−22.7 ± 16.6	−28.5 ± 15.1	−62.7 ± 25.1	−113.9 ± 33.6
2	−33.1 ± 15.6	−44.2 ± 13.6	−54.0 ± 25.4	−131.3 ± 32.8
3	<−1,029,462.5>	<-850,474.9>	<-2,039,573.6>	<−3,919,511.0>



But how about the amino acid analysis?

ML tree of concatenated 337,290 amino acids (JTT + Γ model)

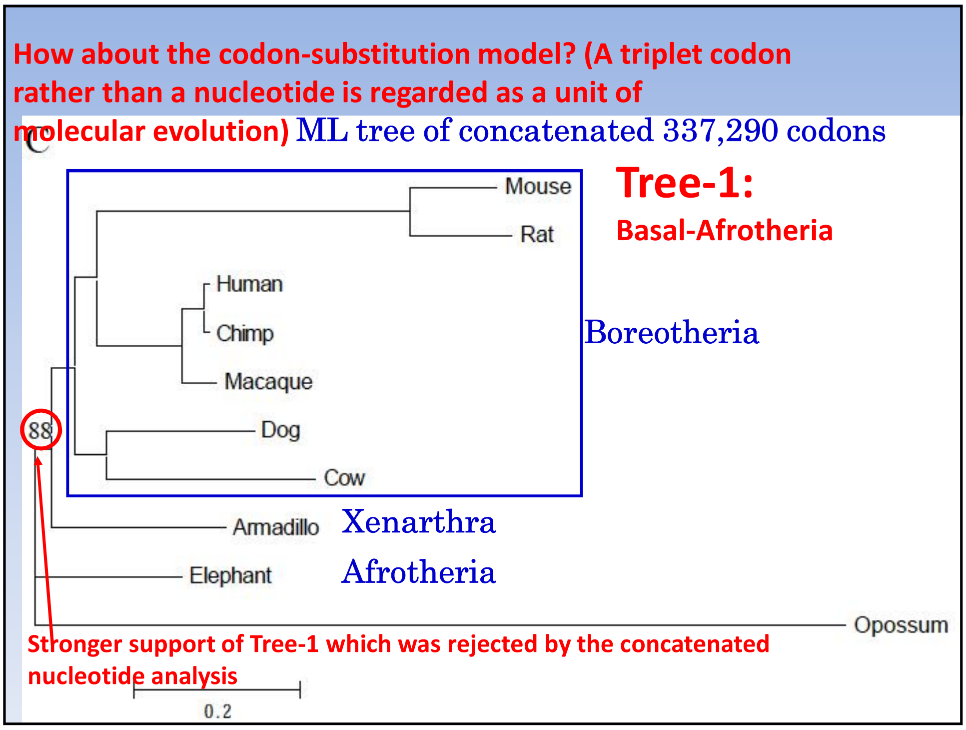
Tree-1:
Basal-Afrotheria

Boreotheria

Xenarthra

Afrotheria

became ambigous



How about the codon-substitution model? (A triplet codon rather than a nucleotide is regarded as a unit of molecular evolution) ML tree of concatenated 337,290 codons

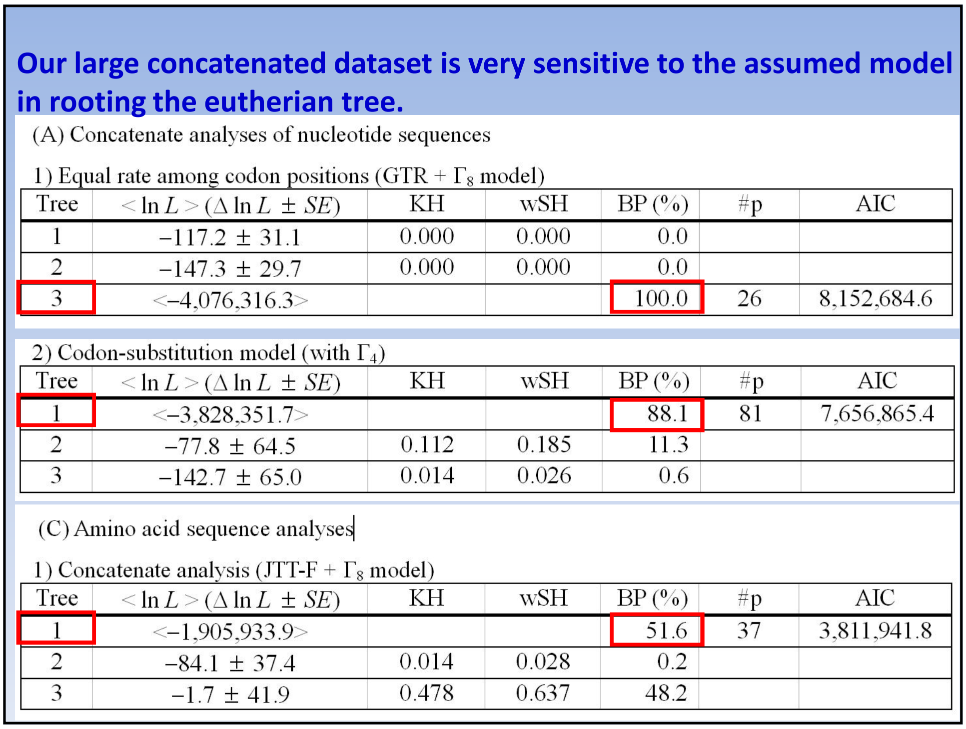
Tree-1:
Basal-Afrotheria

Boreotheria

Xenarthra

Afrotheria

Stronger support of Tree-1 which was rejected by the concatenated nucleotide analysis



Our large concatenated dataset is very sensitive to the assumed model in rooting the eutherian tree.

(A) Concatenate analyses of nucleotide sequences

1) Equal rate among codon positions (GTR + Γ_8 model)

Tree	$<\ln L>$ ($\Delta \ln L \pm SE$)	KH	wSH	BP (%)	#p	AIC
1	−117.2 ± 31.1	0.000	0.000	0.0		
2	−147.3 ± 29.7	0.000	0.000	0.0		
3	<−4,076,316.3>			100.0	26	8,152,684.6

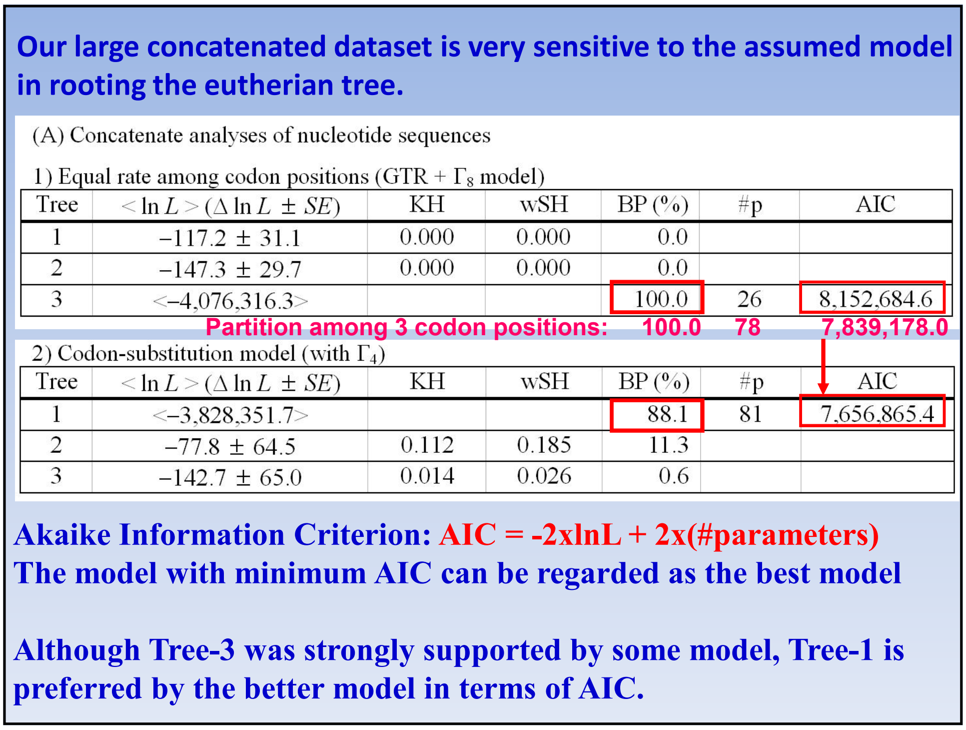
2) Codon-substitution model (with Γ_4)

Tree	$<\ln L>$ ($\Delta \ln L \pm SE$)	KH	wSH	BP (%)	#p	AIC
1	<−3,828,351.7>			88.1	81	7,656,865.4
2	−77.8 ± 64.5	0.112	0.185	11.3		
3	−142.7 ± 65.0	0.014	0.026	0.6		

(C) Amino acid sequence analyses

1) Concatenate analysis (JTT-F + Γ_8 model)

Tree	$<\ln L>$ ($\Delta \ln L \pm SE$)	KH	wSH	BP (%)	#p	AIC
1	<−1,905,933.9>			51.6	37	3,811,941.8
2	−84.1 ± 37.4	0.014	0.028	0.2		
3	−1.7 ± 41.9	0.478	0.637	48.2		



Our large concatenated dataset is very sensitive to the assumed model in rooting the eutherian tree.

(A) Concatenate analyses of nucleotide sequences

1) Equal rate among codon positions (GTR + Γ_8 model)

Tree	$<\ln L>$ ($\Delta \ln L \pm SE$)	KH	wSH	BP (%)	#p	AIC
1	−117.2 ± 31.1	0.000	0.000	0.0		
2	−147.3 ± 29.7	0.000	0.000	0.0		
3	<−4,076,316.3>			100.0	26	8,152,684.6
Partition among 3 codon positions:				100.0	78	7,839,178.0

2) Codon-substitution model (with Γ_4)

Tree	$<\ln L>$ ($\Delta \ln L \pm SE$)	KH	wSH	BP (%)	#p	AIC
1	<−3,828,351.7>			88.1	81	7,656,865.4
2	−77.8 ± 64.5	0.112	0.185	11.3		
3	−142.7 ± 65.0	0.014	0.026	0.6		

Akaike Information Criterion: AIC = -2xlnL + 2x(#parameters)
The model with minimum AIC can be regarded as the best model

Although Tree-3 was strongly supported by some model, Tree-1 is preferred by the better model in terms of AIC.

Nucleotide sequence analyses (GTR + G)

(A) Concatenate analyses of nucleotide sequences

1) Equal rate among codon positions (GTR + Γ_8 model)

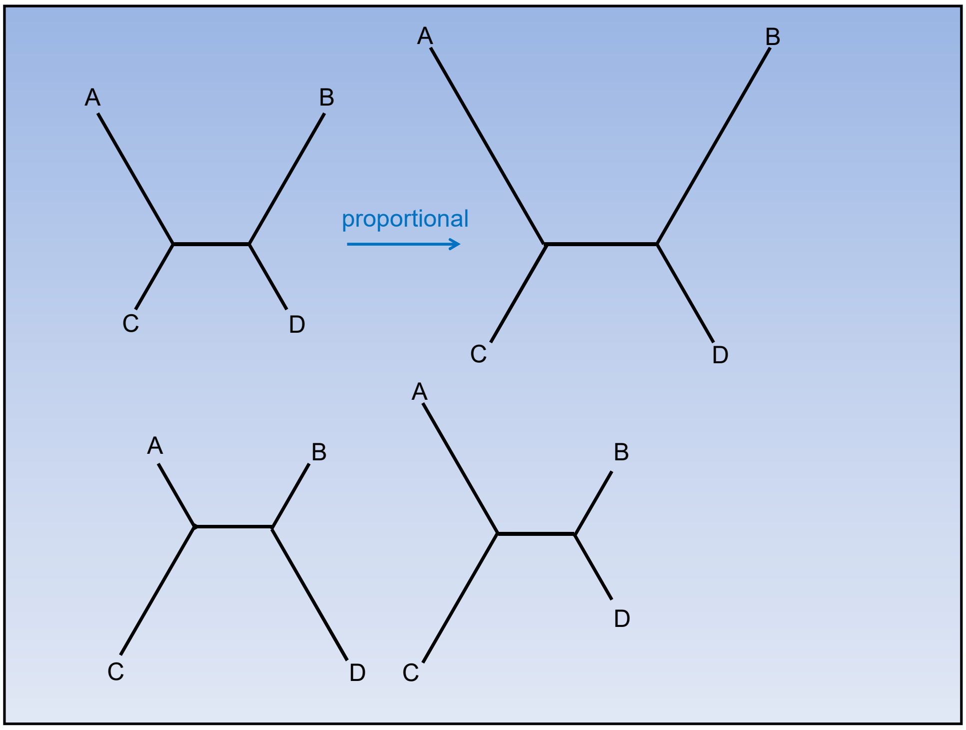
Tree	$<\ln L>$ ($\Delta \ln L \pm SE$)	KH	wSH	BP (%)	#p	AIC
1	-117.2 ± 31.1	0.000	0.000	0.0		
2	-147.3 ± 29.7	0.000	0.000	0.0		
3	$<-4{,}076{,}316.3>$			100.0	26	8,152,684.6

We next carried out ML analysis with the separate model, which takes account of the heterogeneity among different genes by assigning different parameters to different genes, then the support was changed to Tree-1, consistently with the amino acid & codon analyses.

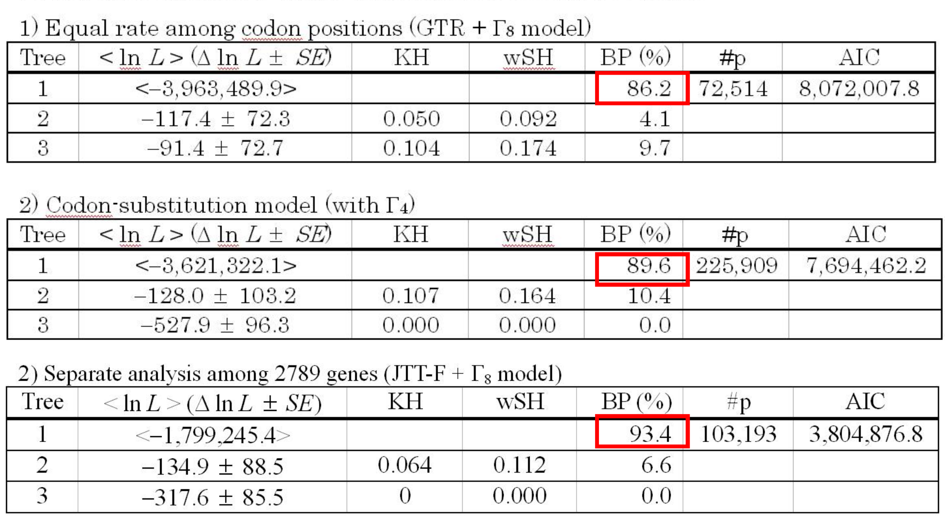
(B) Separate analyses of nucleotide sequences among 2789 genes

1) Equal rate among codon positions (GTR + Γ_8 model)

Tree	$<\ln L>$ ($\Delta \ln L \pm SE$)	KH	wSH	BP (%)	#p	AIC
1	$<-3{,}963{,}489.9>$			86.2	72,514	8,072,007.8
2	-117.4 ± 72.3	0.050	0.092	4.1		
3	-91.4 ± 72.7	0.104	0.174	9.7		



Separate analyses of 2,789 genes



(B) Separate analyses of nucleotide sequences among 2789 genes

1) Equal rate among codon positions (GTR + Γ_8 model)

Tree	< ln L > (Δ ln L ± SE)	KH	wSH	BP (%)	#p	AIC
1	<−3,963,489.9>			86.2	72,514	8,072,007.8
2	−117.4 ± 72.3	0.050	0.092	4.1		
3	−91.4 ± 72.7	0.104	0.174	9.7		

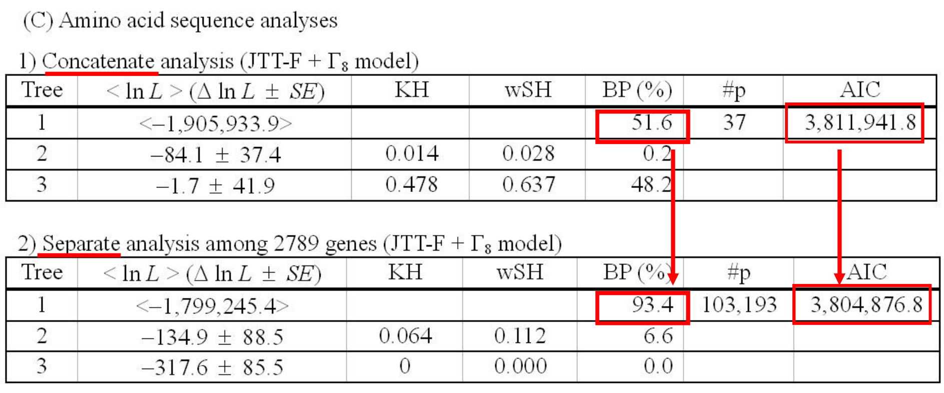
2) Codon-substitution model (with Γ_4)

Tree	< ln L > (Δ ln L ± SE)	KH	wSH	BP (%)	#p	AIC
1	<−3,621,322.1>			89.6	225,909	7,694,462.2
2	−128.0 ± 103.2	0.107	0.164	10.4		
3	−527.9 ± 96.3	0.000	0.000	0.0		

2) Separate analysis among 2789 genes (JTT-F + Γ_8 model)

Tree	< ln L > (Δ ln L ± SE)	KH	wSH	BP (%)	#p	AIC
1	<−1,799,245.4>			93.4	103,193	3,804,876.8
2	−134.9 ± 88.5	0.064	0.112	6.6		
3	−317.6 ± 85.5	0	0.000	0.0		

Tree-1 is robustly supported irrespective of the model if the difference among genes is taken into account by the separate analysis.



(C) Amino acid sequence analyses

1) Concatenate analysis (JTT-F + Γ_8 model)

Tree	< ln L > (Δ ln L ± SE)	KH	wSH	BP (%)	#p	AIC
1	<−1,905,933.9>			51.6	37	3,811,941.8
2	−84.1 ± 37.4	0.014	0.028	0.2		
3	−1.7 ± 41.9	0.478	0.637	48.2		

2) Separate analysis among 2789 genes (JTT-F + Γ_8 model)

Tree	< ln L > (Δ ln L ± SE)	KH	wSH	BP (%)	#p	AIC
1	<−1,799,245.4>			93.4	103,193	3,804,876.8
2	−134.9 ± 88.5	0.064	0.112	6.6		
3	−317.6 ± 85.5	0	0.000	0.0		

Akaike Information Criterion: AIC = -2xlnL + 2x(#parameters)
The model with minimum AIC can be regarded as the best model.

Tree-1 is preferred by both concatenate and separate analyses of amino acid sequences, but the better model in terms of AIC, that is the separate analysis, gives higher resolution.

Conclusion

- The strong support of Tree-3 by the concatenated analysis of nucleotide sequences is probably an artifact due to neglect of heterogeneity of tempo & mode among different genes.
- This may often be a problem in phylogenetic analyses of genome-scale data.
- Tree-1 is the most likely tree from the 1Mbp data, but the best available model cannot exclude an alternative tree, particularly Tree-2. Therefore, the rooting problem of eutherian mammals still remains unresolved even with the genome-scale analysis.
- Probably, two branchings among the three major groups of eutherian mammals occurred successively in a short time interval.

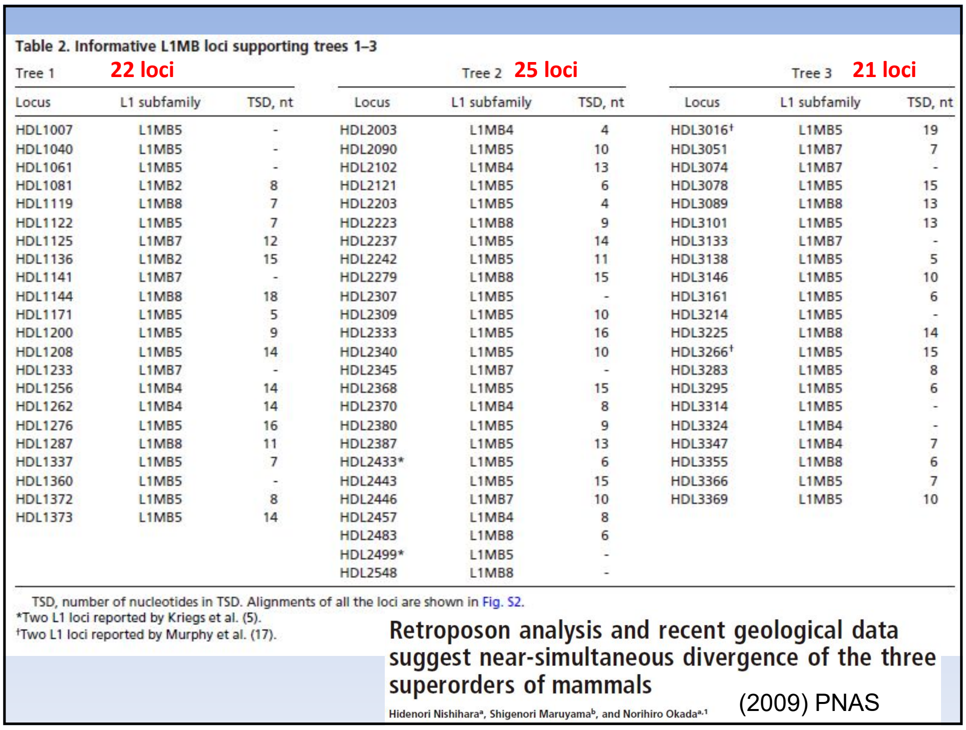


Table 2. Informative L1MB loci supporting trees 1–3

Tree 1 22 loci			Tree 2 25 loci			Tree 3 21 loci		
Locus	L1 subfamily	TSD, nt	Locus	L1 subfamily	TSD, nt	Locus	L1 subfamily	TSD, nt
HDL1007	L1MB5	-	HDL2003	L1MB4	4	HDL3016†	L1MB5	19
HDL1040	L1MB5	-	HDL2090	L1MB5	10	HDL3051	L1MB7	7
HDL1061	L1MB5	-	HDL2102	L1MB4	13	HDL3074	L1MB7	-
HDL1081	L1MB2	8	HDL2121	L1MB5	6	HDL3078	L1MB5	15
HDL1119	L1MB8	7	HDL2203	L1MB5	4	HDL3089	L1MB8	13
HDL1122	L1MB5	7	HDL2223	L1MB8	9	HDL3101	L1MB5	13
HDL1125	L1MB7	12	HDL2237	L1MB5	14	HDL3133	L1MB7	-
HDL1136	L1MB2	15	HDL2242	L1MB5	11	HDL3138	L1MB5	5
HDL1141	L1MB7	-	HDL2279	L1MB8	15	HDL3146	L1MB5	10
HDL1144	L1MB8	18	HDL2307	L1MB5	-	HDL3161	L1MB5	6
HDL1171	L1MB5	5	HDL2309	L1MB5	10	HDL3214	L1MB5	-
HDL1200	L1MB5	9	HDL2333	L1MB5	16	HDL3225	L1MB8	14
HDL1208	L1MB5	14	HDL2340	L1MB5	10	HDL3266†	L1MB5	15
HDL1233	L1MB7	-	HDL2345	L1MB7	-	HDL3283	L1MB5	8
HDL1256	L1MB4	14	HDL2368	L1MB5	15	HDL3295	L1MB5	6
HDL1262	L1MB4	14	HDL2370	L1MB4	8	HDL3314	L1MB5	-
HDL1276	L1MB5	16	HDL2380	L1MB5	9	HDL3324	L1MB4	-
HDL1287	L1MB8	11	HDL2387	L1MB5	13	HDL3347	L1MB4	7
HDL1337	L1MB5	7	HDL2433*	L1MB5	6	HDL3355	L1MB8	6
HDL1360	L1MB5	-	HDL2443	L1MB5	15	HDL3366	L1MB5	7
HDL1372	L1MB5	8	HDL2446	L1MB7	10	HDL3369	L1MB5	10
HDL1373	L1MB5	14	HDL2457	L1MB4	8			
			HDL2483	L1MB8	6			
			HDL2499*	L1MB5	-			
			HDL2548	L1MB8	-			

TSD, number of nucleotides in TSD. Alignments of all the loci are shown in Fig. S2.

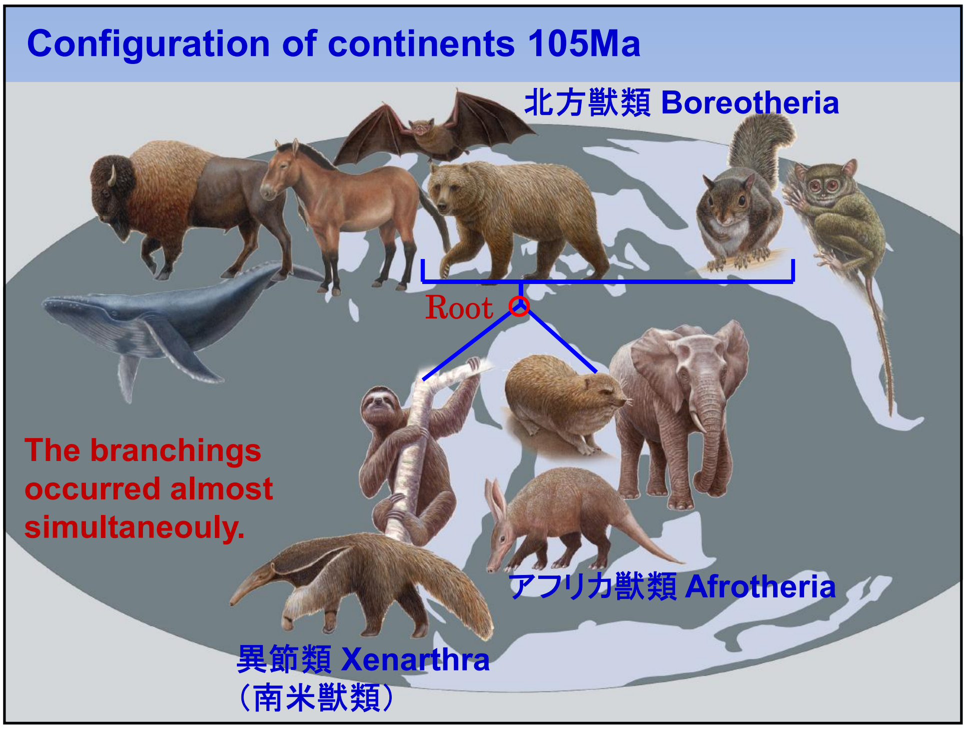
*Two L1 loci reported by Kriegs et al. (5).

†Two L1 loci reported by Murphy et al. (17).

Retroposon analysis and recent geological data suggest near-simultaneous divergence of the three superorders of mammals

Hidenori Nishihara[a], Shigenori Maruyama[b], and Norihiro Okada[a,1]

(2009) PNAS



Configuration of continents 105Ma

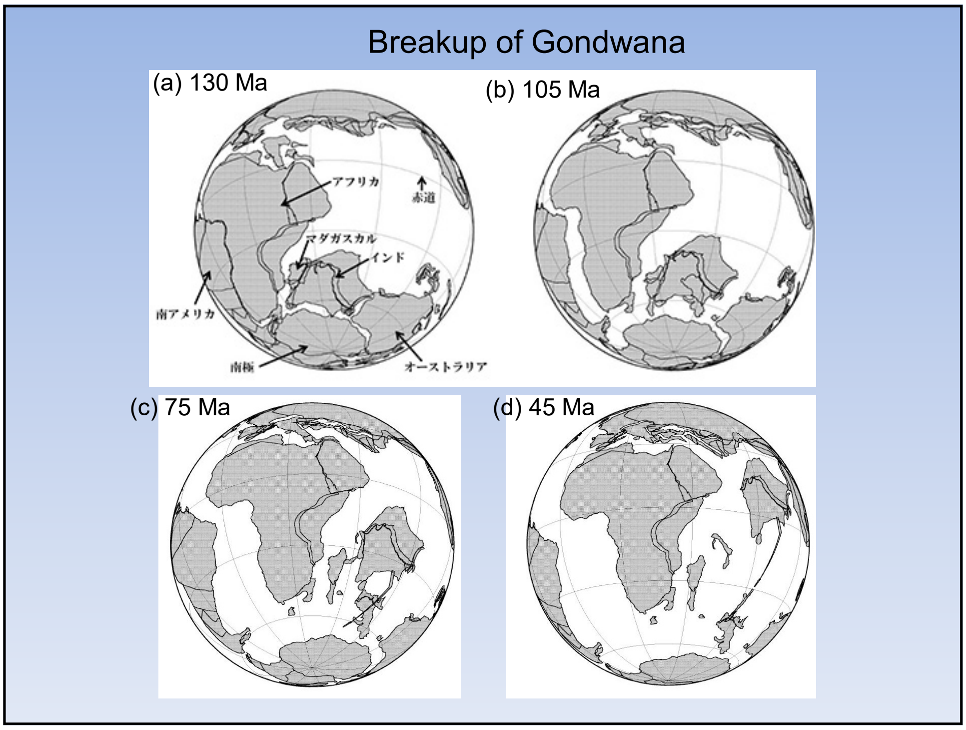
北方獣類 Boreotheria

Root

The branchings occurred almost simultaneouly.

アフリカ獣類 Afrotheria

異節類 Xenarthra
（南米獣類）



Breakup of Gondwana

(a) 130 Ma

(b) 105 Ma

(c) 75 Ma

(d) 45 Ma

Longer sequences or more taxa?

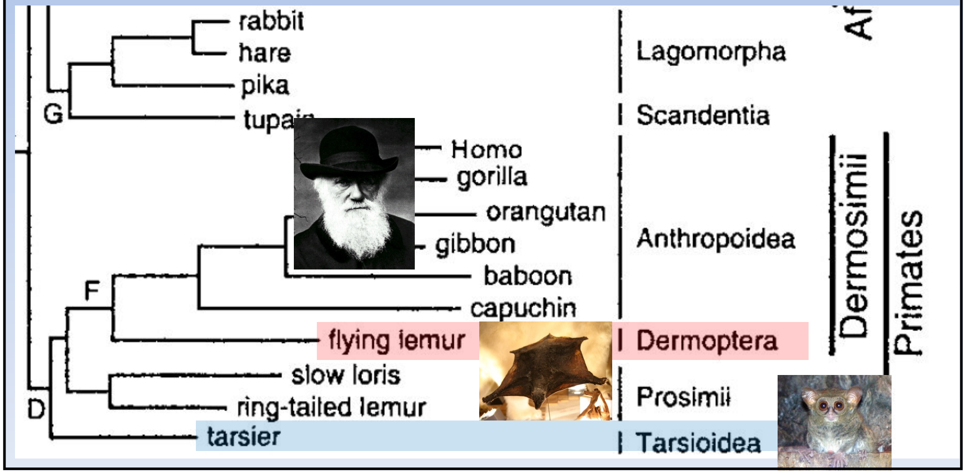
- The model misspecification gives a serious effect when taxon sampling is sparse.
- On the other hand, dense taxon sampling could relieve bised estimation by model misspecification.

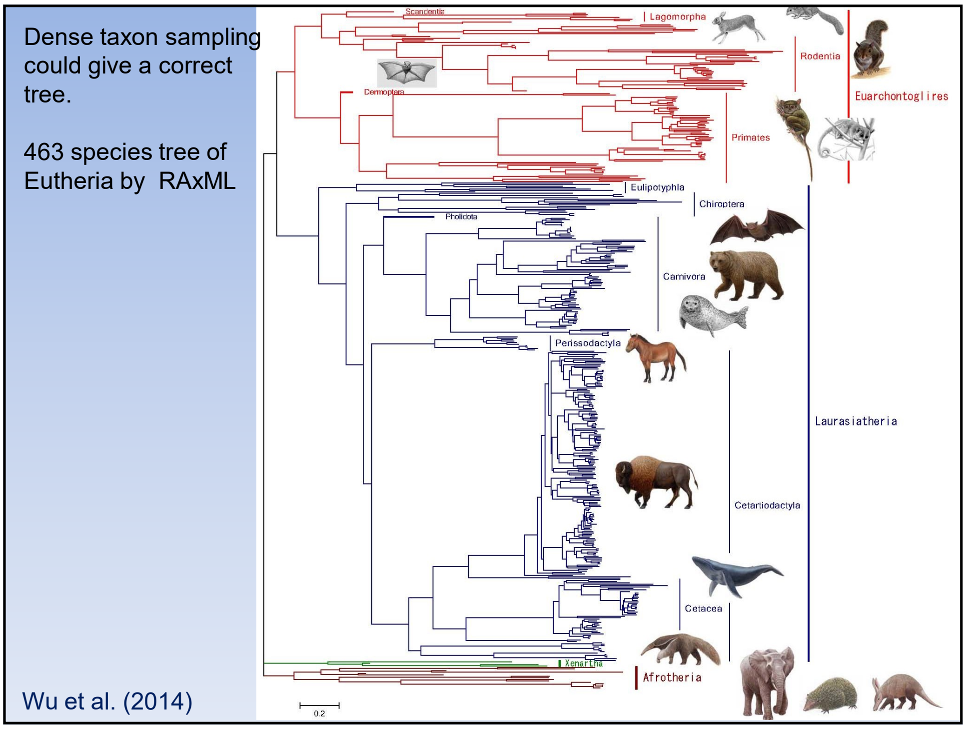
Mammalian mitogenomic relationships and the root of the eutherian tree

PNAS | June 11, 2002 | vol. 99 | no. 12 | 8151–8156

Ulfur Arnason[*,†], Joseph A. Adegoke[*,‡], Kristina Bodin[*], Erik W. Born[§], Yuzine B. Esa[¶], Anette Gullberg[*], Maria Nilsson[*], Roger V. Short[‖], Xiufeng Xu[*,**], and Axel Janke[*]

Fying lemur evolved within Primates?
Tarsier is a basal lineage in Primates?
Mitogenome analyses are often misleading in estimating a tree.

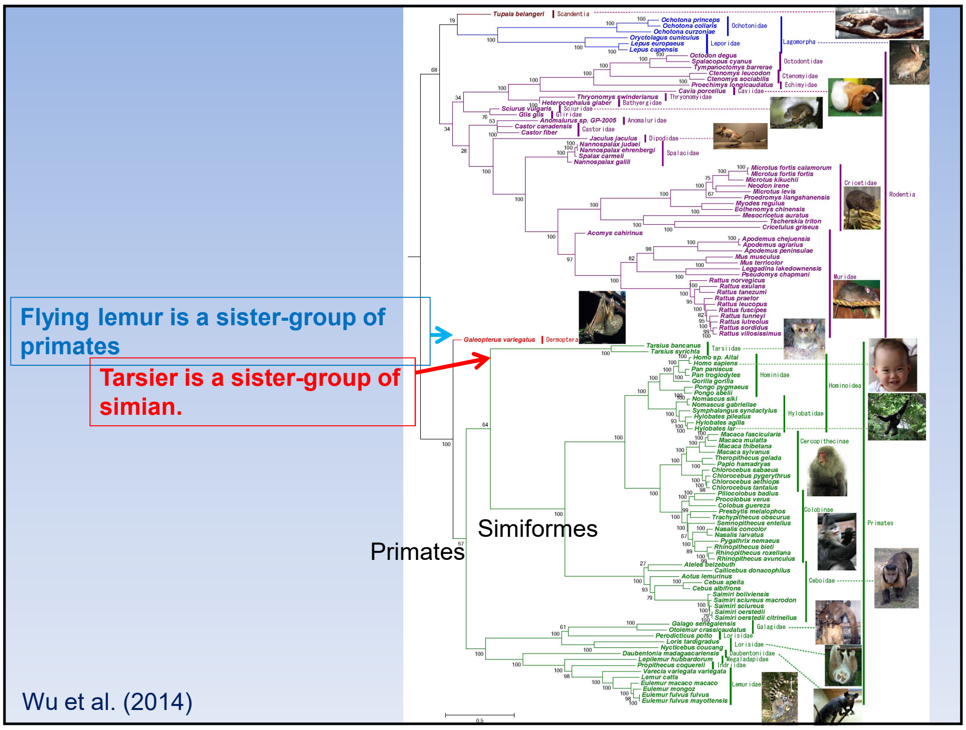




Dense taxon sampling could give a correct tree.

463 species tree of Eutheria by RAxML

Wu et al. (2014)



Flying lemur is a sister-group of primates

Tarsier is a sister-group of simian.

Simiformes

Primates

Wu et al. (2014)

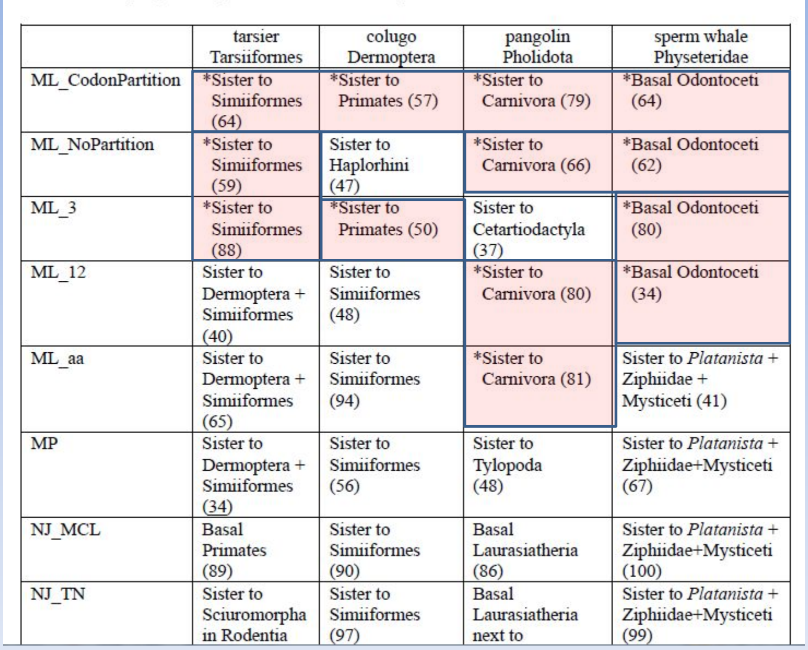
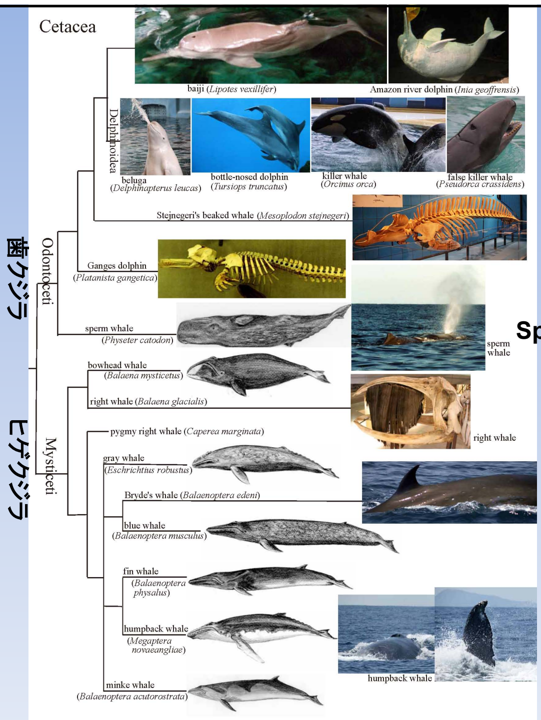


Table 1: Phylogenetic positions estimated by various methods and models

	tarsier Tarsiiformes	colugo Dermoptera	pangolin Pholidota	sperm whale Physeteridae
ML_CodonPartition	*Sister to Simiiformes (64)	*Sister to Primates (57)	*Sister to Carnivora (79)	*Basal Odontoceti (64)
ML_NoPartition	*Sister to Simiiformes (59)	Sister to Haplorhini (47)	*Sister to Carnivora (66)	*Basal Odontoceti (62)
ML_3	*Sister to Simiiformes (88)	*Sister to Primates (50)	Sister to Cetartiodactyla (37)	*Basal Odontoceti (80)
ML_12	Sister to Dermoptera + Simiiformes (40)	Sister to Simiiformes (48)	*Sister to Carnivora (80)	*Basal Odontoceti (34)
ML_aa	Sister to Dermoptera + Simiiformes (65)	Sister to Simiiformes (94)	*Sister to Carnivora (81)	Sister to *Platanista* + Ziphiidae + Mysticeti (41)
MP	Sister to Dermoptera + Simiiformes (34)	Sister to Simiiformes (56)	Sister to Tylopoda (48)	Sister to *Platanista* + Ziphiidae+Mysticeti (67)
NJ_MCL	Basal Primates (89)	Sister to Simiiformes (90)	Basal Laurasiatheria (86)	Sister to *Platanista* + Ziphiidae+Mysticeti (100)
NJ_TN	Sister to Sciuromorpha in Rodentia	Sister to Simiiformes (97)	Basal Laurasiatheria next to	Sister to *Platanista* + Ziphiidae+Mysticeti (99)

Wu et al. (2014)



Sperm whale

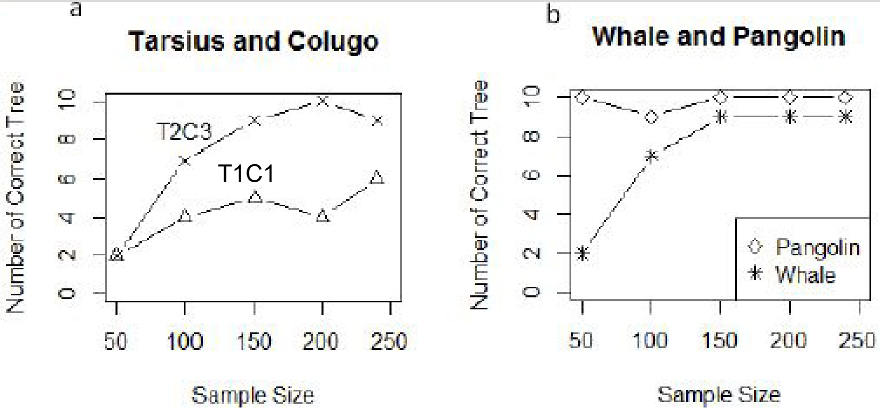
長谷川政美(2011) 改変
「新図説・動物の起源と進化」
八坂書房

食肉目

有鱗目・センザンコウ

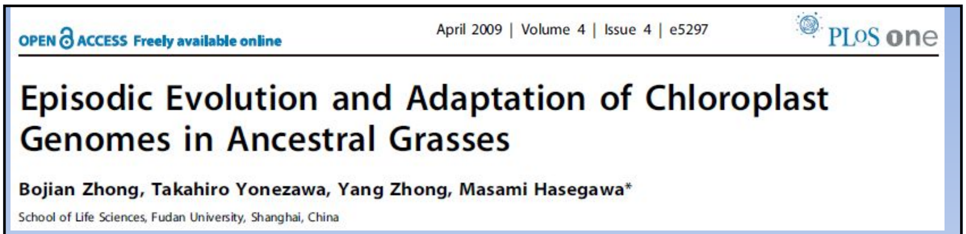
長谷川政美(2011) 改変
「新図説・動物の起源と進化」
八坂書房

Effect of the density of taxon sampling



T1C1: 1 Tupai + 1 Colugo
T2C3: 2 Tupais + 3 Colugos

Wu et al. (2014)



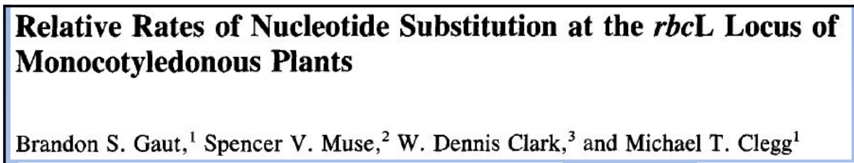
OPEN ACCESS Freely available online April 2009 | Volume 4 | Issue 4 | e5297 PLoS one

Episodic Evolution and Adaptation of Chloroplast Genomes in Ancestral Grasses

Bojian Zhong, Takahiro Yonezawa, Yang Zhong, Masami Hasegawa*

School of Life Sciences, Fudan University, Shanghai, China

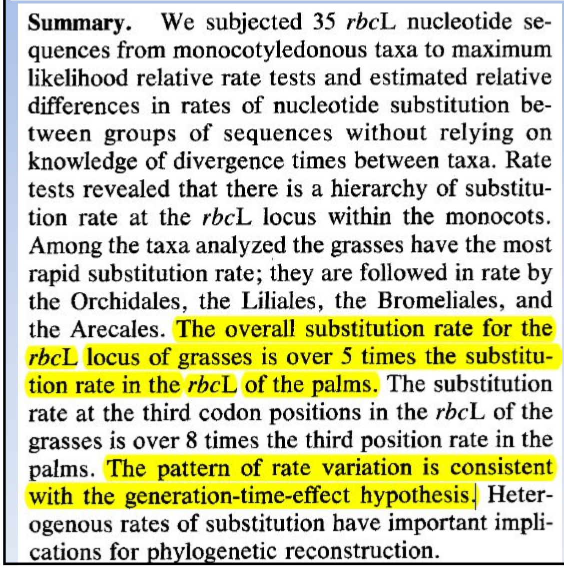
- Poaceae (イネ科)consists of about 10,000 species, and is one of the most flourishing group in Angiosperm. This group includes the most important crops such as rice, wheat and maize.
- It has been known since the early study (Gaut et al., *J. Mol. Evol*. **35**, 292(1992)) that evolutionary rate of mtDNA of Poaceae is higher than that of palms(ヤシ) which also belong monocots. This was attributed to the difference of the generation time, but the details have been unknown.



Relative Rates of Nucleotide Substitution at the *rbc*L Locus of Monocotyledonous Plants

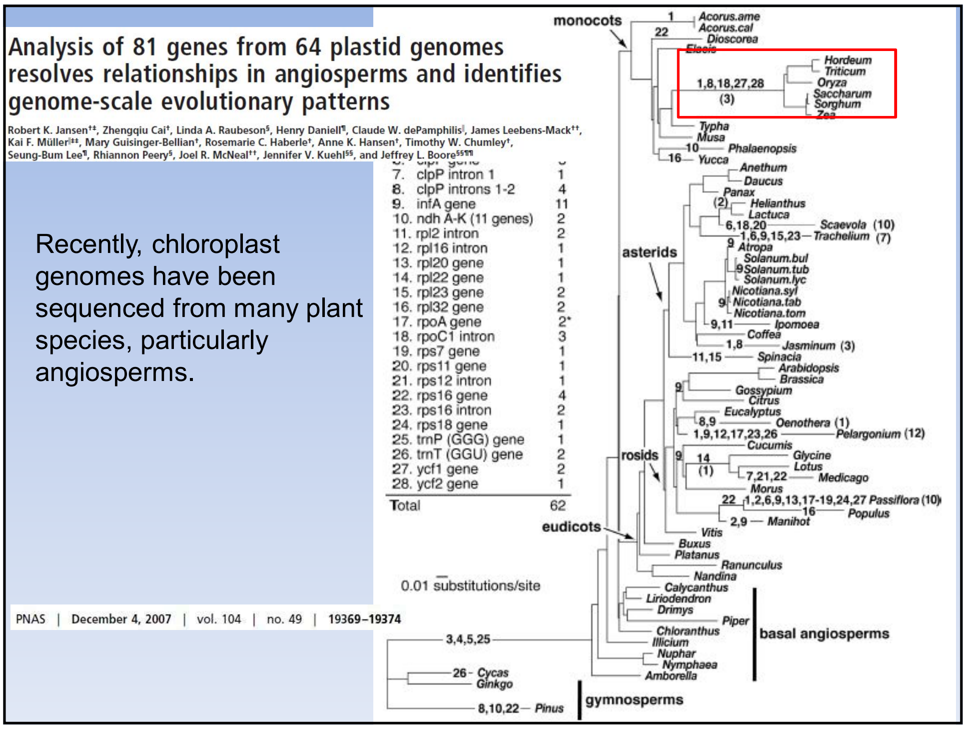
Brandon S. Gaut,[1] Spencer V. Muse,[2] W. Dennis Clark,[3] and Michael T. Clegg[1]

J Mol Evol (1992) 35:292–303



Summary. We subjected 35 *rbc*L nucleotide sequences from monocotyledonous taxa to maximum likelihood relative rate tests and estimated relative differences in rates of nucleotide substitution between groups of sequences without relying on knowledge of divergence times between taxa. Rate tests revealed that there is a hierarchy of substitution rate at the *rbc*L locus within the monocots. Among the taxa analyzed the grasses have the most rapid substitution rate; they are followed in rate by the Orchidales, the Liliales, the Bromeliales, and the Arecales. The overall substitution rate for the *rbc*L locus of grasses is over 5 times the substitution rate in the *rbc*L of the palms. The substitution rate at the third codon positions in the *rbc*L of the grasses is over 8 times the third position rate in the palms. The pattern of rate variation is consistent with the generation-time-effect hypothesis. Heterogenous rates of substitution have important implications for phylogenetic reconstruction.

Grasses (イネ科:一年生) の葉緑体遺伝子はPalms (ヤシ: 多年生)の遺伝子よりも 進化速度が速い。世代時間が関係している？



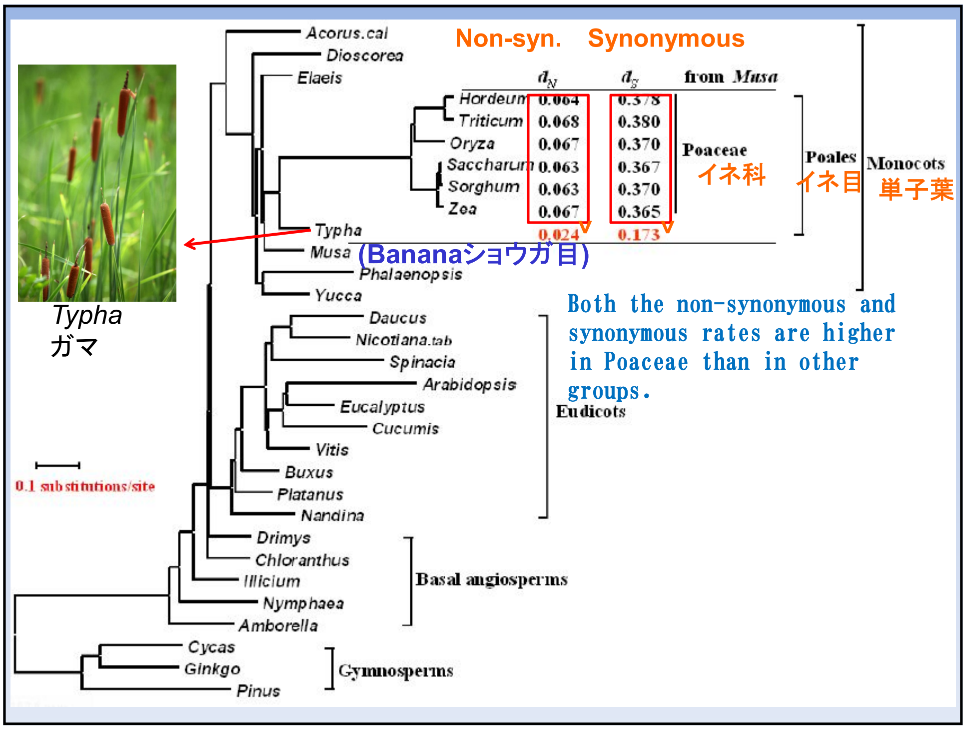
Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns

Robert K. Jansen[†‡], Zhengqiu Cai[†], Linda A. Raubeson[§], Henry Daniell[¶], Claude W. dePamphilis[‖], James Leebens-Mack[††], Kai F. Müller[‖‡‡], Mary Guisinger-Bellian[†], Rosemarie C. Haberle[†], Anne K. Hansen[†], Timothy W. Chumley[†], Seung-Bum Lee[¶], Rhiannon Peery[§], Joel R. McNeal[††], Jennifer V. Kuehl[§§], and Jeffrey L. Boore[§§¶¶]

PNAS | December 4, 2007 | vol. 104 | no. 49 | 19369–19374

Recently, chloroplast genomes have been sequenced from many plant species, particularly angiosperms.

7. clpP intron 1	1
8. clpP introns 1-2	4
9. infA gene	11
10. ndh A-K (11 genes)	2
11. rpl2 intron	2
12. rpl16 intron	1
13. rpl20 gene	1
14. rpl22 gene	1
15. rpl23 gene	2
16. rpl32 gene	2
17. rpoA gene	2*
18. rpoC1 intron	3
19. rps7 gene	1
20. rps11 gene	1
21. rps12 intron	1
22. rps16 gene	4
23. rps16 intron	2
24. rps18 gene	1
25. trnP (GGG) gene	1
26. trnT (GGU) gene	2
27. ycf1 gene	2
28. ycf2 gene	1
Total	62



Non-syn. Synonymous

	d_N	d_S	from *Musa*
Hordeum	0.064	0.378	Poaceae イネ科
Triticum	0.068	0.380	
Oryza	0.067	0.370	
Saccharum	0.063	0.367	
Sorghum	0.063	0.370	
Zea	0.067	0.365	
Typha	0.024	0.173	

Both the non-synonymous and synonymous rates are higher in Poaceae than in other groups.

ω = Non-synonymous/Synonymous rate

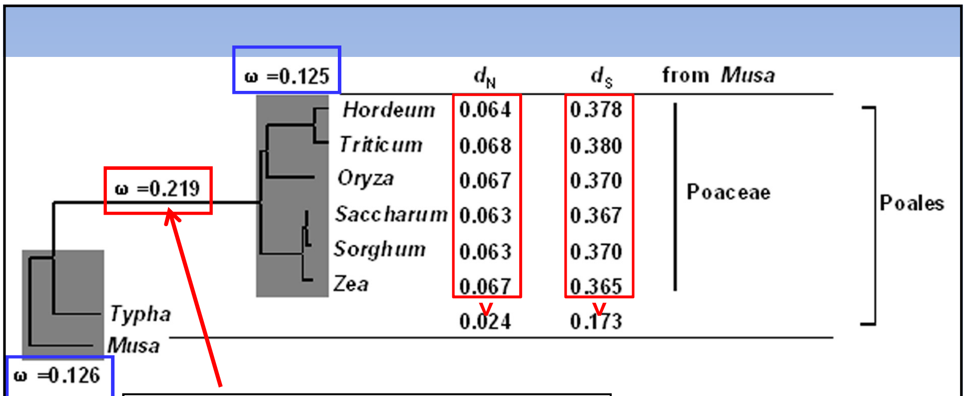
- Since synonymous substitutions do not experience sever constraints, the rate must be higher than that of non-synonymous substitutions which are under functional constraints when the neutral theory holds. → $\omega < 1$
- When positive selection is operating (adaptive evolution), $\omega > 1$ could hold.

d_S (K_S) : number of synonymous substitutions per synonymous site

d_N (K_A): number of nonsynonymous substitutions per nonsynonymous site

$\omega = d_N / d_S$: nonsynonymous/synonymous rate ratio

The ω ratio is widely used in detecting adaptive evolution. Is the rate acceleration of grasses correlated with the ω ratio change?

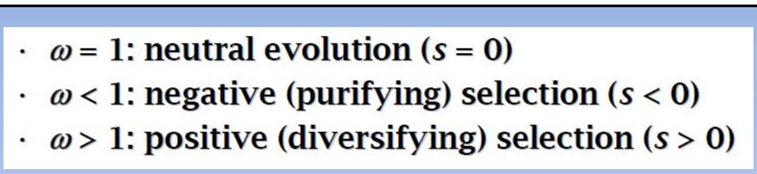


	d_N	d_S
Hordeum	0.064	0.378
Triticum	0.068	0.380
Oryza	0.067	0.370
Saccharum	0.063	0.367
Sorghum	0.063	0.370
Zea	0.067	0.365
	0.024	0.173

A pronounced increase of ω ratio occurred in the common ancestral branch of Poaceae, followed by reversion in the terminal branches.

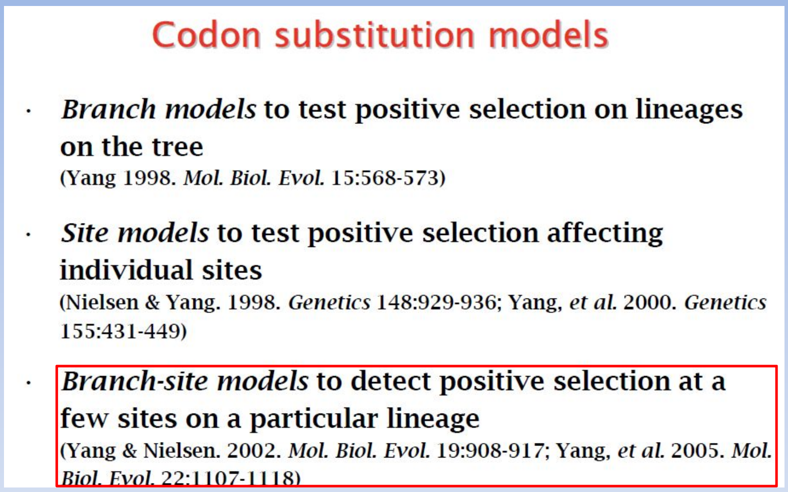
High ω-ratio can be interpreted in two different ways

- Relaxation of constraints
- Possibility of positive selection (adaptive evolution)



- $\omega = 1$: neutral evolution ($s = 0$)
- $\omega < 1$: negative (purifying) selection ($s < 0$)
- $\omega > 1$: positive (diversifying) selection ($s > 0$)

$\omega > 1$ is usually regarded as an evidence of positive selection, but since the analysis averages over the entire genomes, we would not get such a high value even if positive selection is operating in some parts of some proteins. Therefore, we apply the branch-site model to detect these parts.



Codon substitution models

- *Branch models* to test positive selection on lineages on the tree
  (Yang 1998. *Mol. Biol. Evol.* 15:568-573)
- *Site models* to test positive selection affecting individual sites
  (Nielsen & Yang. 1998. *Genetics* 148:929-936; Yang, *et al.* 2000. *Genetics* 155:431-449)
- *Branch-site models* to detect positive selection at a few sites on a particular lineage
  (Yang & Nielsen. 2002. *Mol. Biol. Evol.* 19:908-917; Yang, *et al.* 2005. *Mol. Biol. Evol.* 22:1107-1118)

Branch-site test of positive selection in the common ancestral grass branch

gene	LRT	Positively selected sites (Candidates)
*atp*E	0.0484	2T→K, 17S→C, 41A→N, 64M→W, 132V→W
*cem*A	0.0021	55N→R, 76Y→K, 161W→F, 190I→F, 204I→C
*clp*P	0.0081	26R→V, 48V→T, 86F→T, 112I→P, 134E→R, 182T→D
*rpo*B	0.0352	90R→F, 338G→K, 1026G→N
*rps*11	0.0082	54V→P, 62A→S, 82A→R, 105L→S, 115R→A, 120L→R

This list should be regarded as merely candidates of positively selected sites, a
additional evidence should be given with the structural biological or experiment
analyses in order to be identified as positively selected sites.