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.**

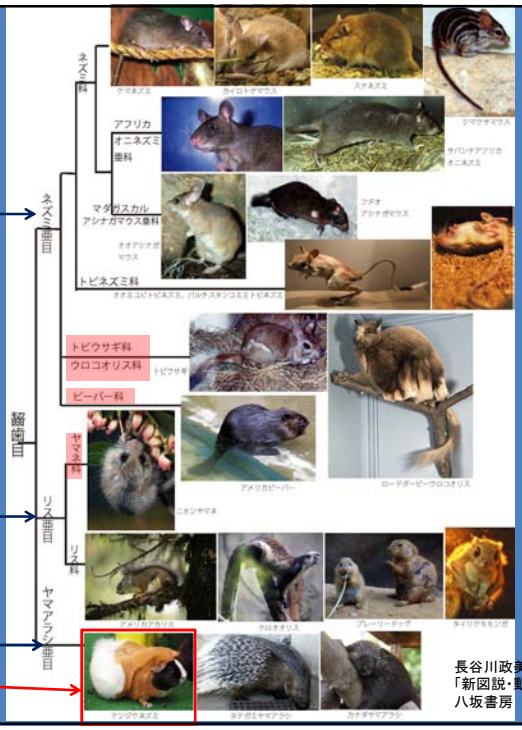
Systematic errors in phylogenetic inference

- Long-branch-attraction (LBA)

Tree of Rodentia

- Myomorpha
- Sciuromorpha
- Hystricomprpha

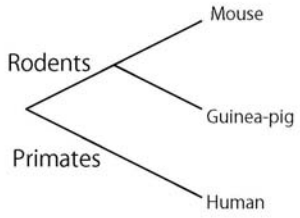
Guinea-pig



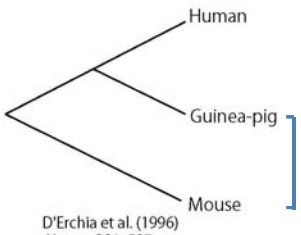
長谷川政美(2011)
「新図説・動物の起源と進化」
八坂書房

Rodents polyphyly?（齧歯類は多系統か？）

Tree-1 (Traditional)



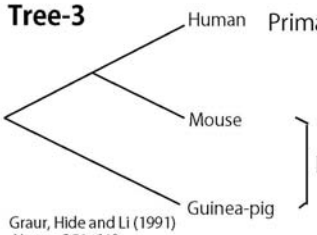
Tree-2



Rodents paraphyly
側系統

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

Tree-3



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

Parsimony（節約法）

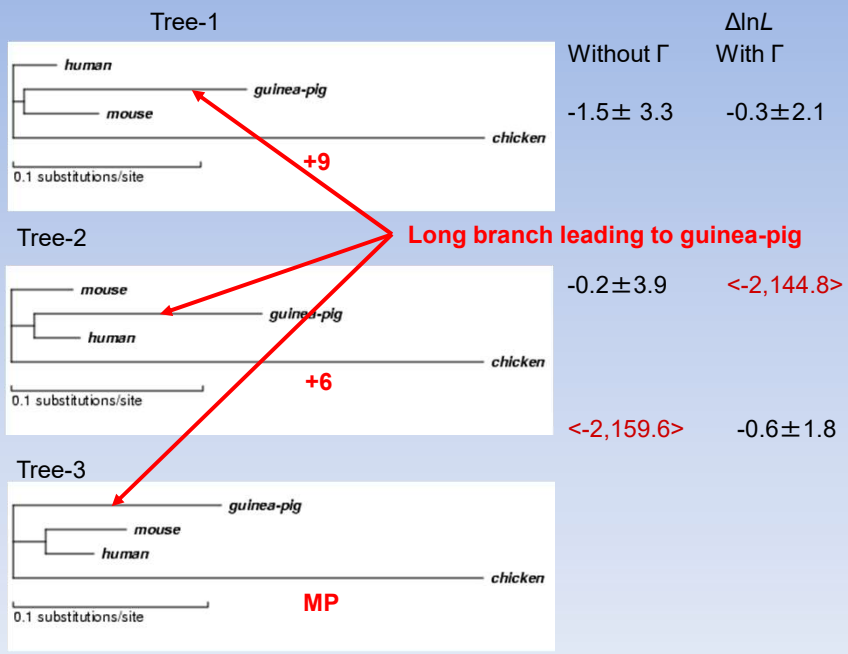
	Tree-1	Tree-2	Tree-3
Crystalli	MP	2	3
Lactalbu	5	MP	4
Hbα	4	8	MP
Hbβ	1	5	MP
NGF	1	5	MP
Factor9	1	MP	3
Ribonucl	3	MP	2
Insulin	1	1	MP
LipLipas	9	6	MP
Lipocort	4	5	MP
Total	17	20	MP
BP	0.0303	0.0108	0.9589

Reanalsis of Graur et al.'s data (Hasegawa et al. (1992) *Nature* 355, 595)

	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)



ΔlnL

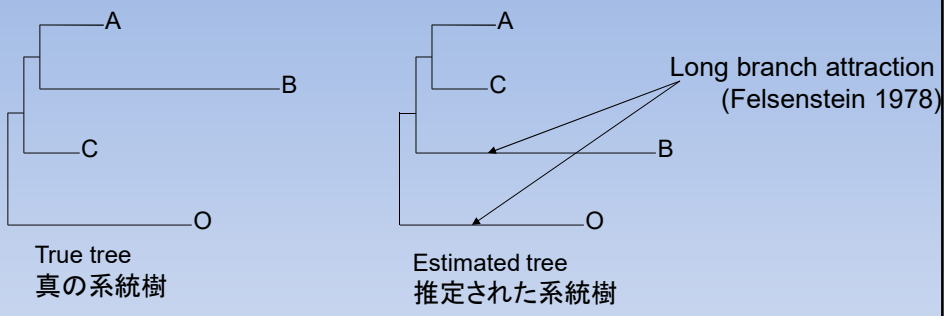
Without Γ　　With Γ

-1.5± 3.3　　-0.3±2.1

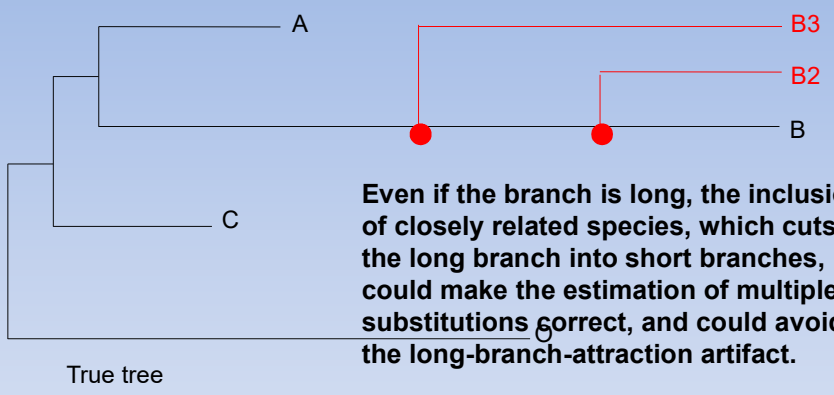
-0.2±3.9　　<-2,144.8>

<-2,159.6>　　-0.6±1.8

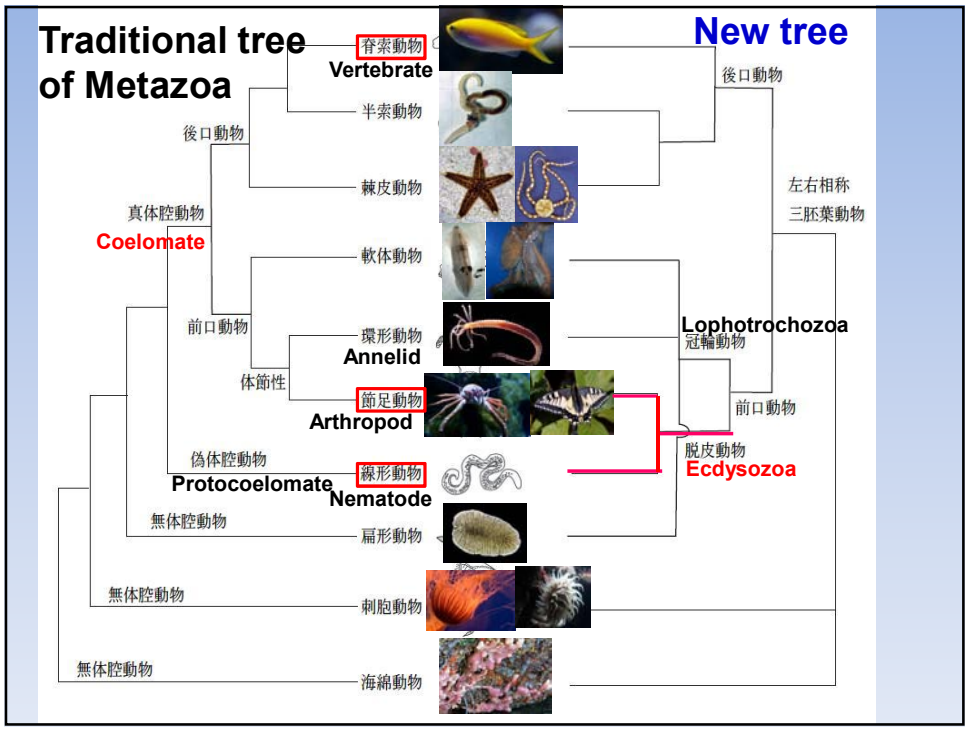
Long branch attraction artifact

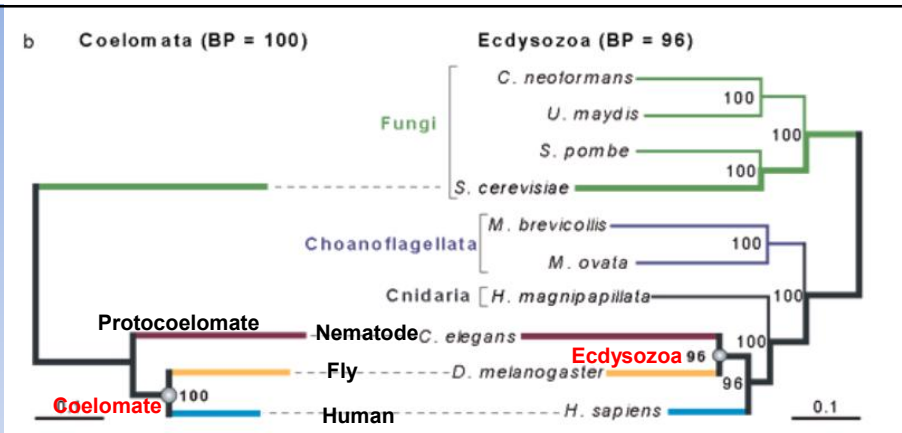


The longer the branch with many substitutions, the more extreme the underestimation of multiple-substitutions becomes, and accordingly two long branches tend to make a cluster erroneouly. The parsimony method ignores the multiple substitutions, and therefore the method is affected most seriouly. In the 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.





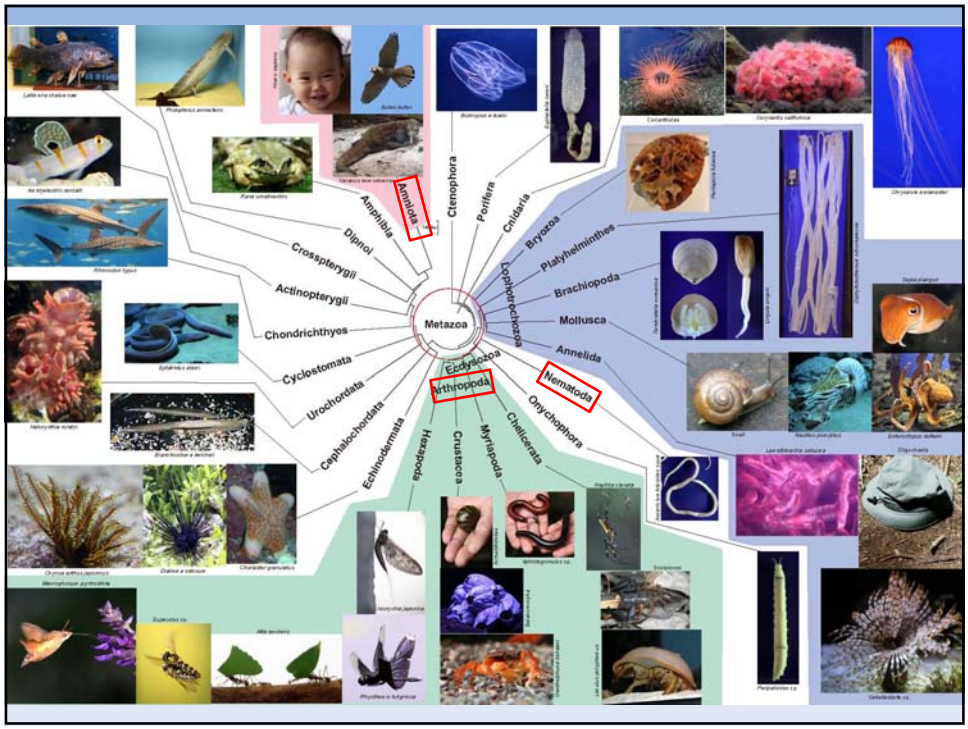
Blair et al. (2002)

Philippe et al. (2005)
Nature Reviews | Genetics

ML tree with the JTT+Γ model

Long-branch-attraction

Delsuc, Brinkmann & Philippe (2005)

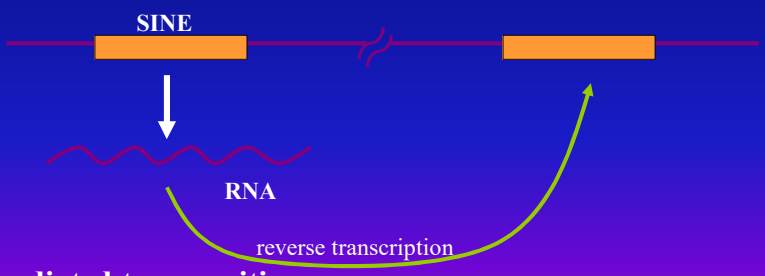


Systematic errors in phylogenetic inference

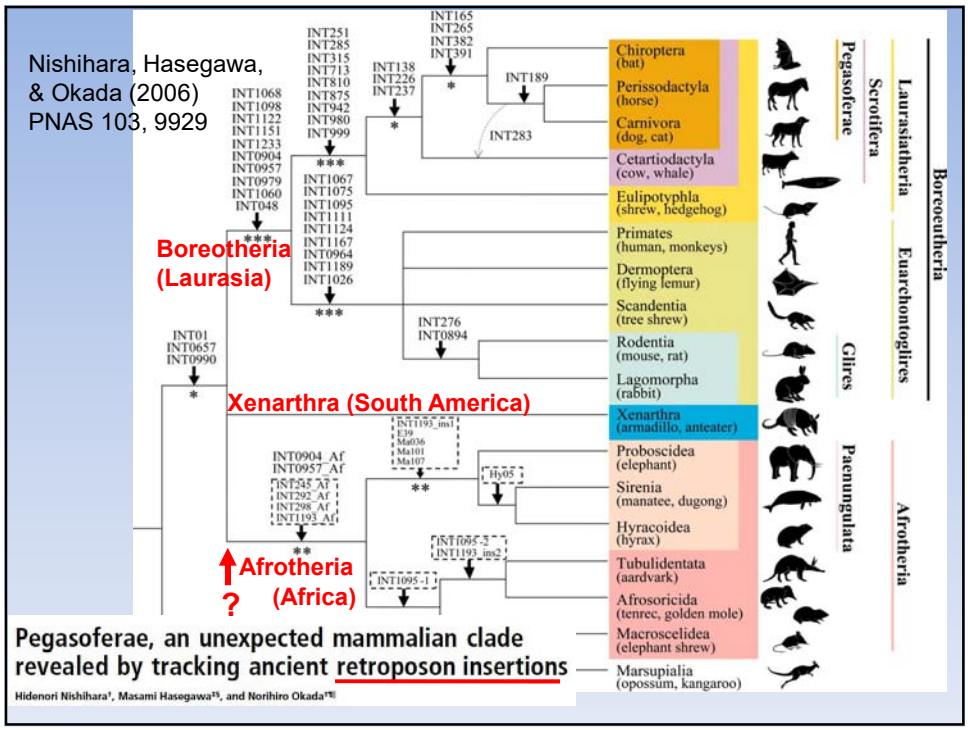
- Long-branch-attraction (LBA)
- **Heterogeneous tempo & mode of evolution among different genes**

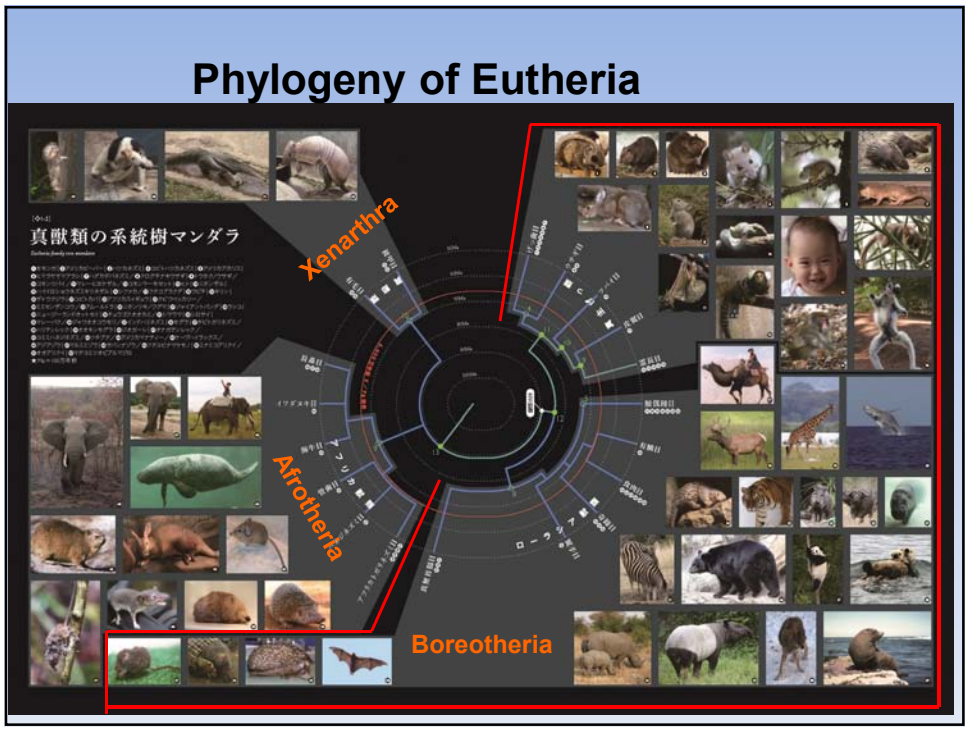
Retroposon insertion method developed by Norihiro Okada in 1990s

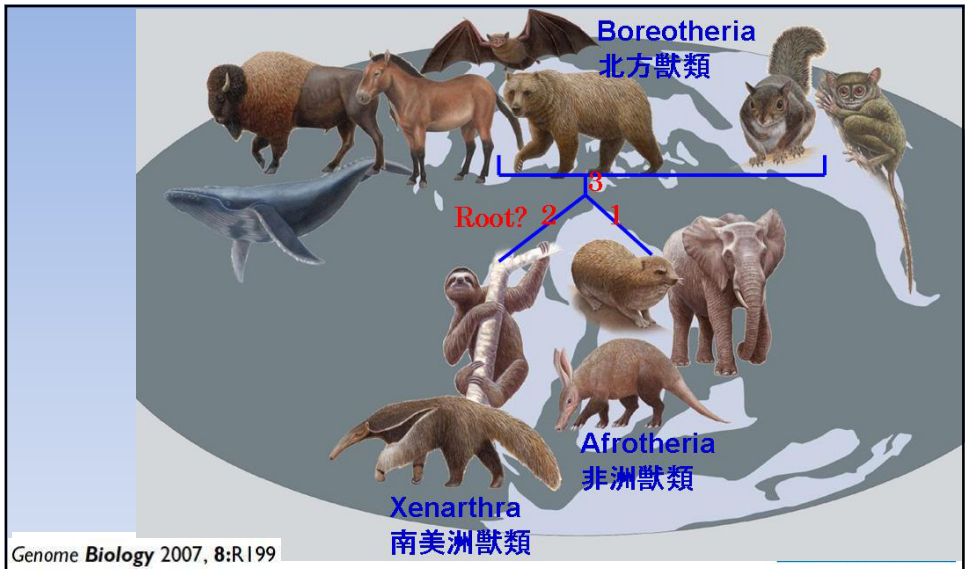
Retroposons such as SINE and LINE are inserted randomly into a genome. If the same retroposon is found in the same locus in different species, then the insertion must have occurred in the common ancestor of the two species. (Independent insertion in the same locus must be very rare.)



RNA-mediated transposition


Nishihara, Hasegawa,
& Okada (2006)
PNAS 103, 9929
Boreotheria
(Laurasia)
Xenarthra (South America)
Afrotheria
(Africa)
?
Chiroptera
(bat)
Perissodactyla
(horse)
Carnivora
(dog, cat)
Cetartiodactyla
(cow, whale)
Eulipotyphla
(shrew, hedgehog)
Primates
(human, monkeys)
Dermoptera
(flying lemur)
Scandentia
(tree shrew)
Rodentia
(mouse, rat)
Lagomorpha
(rabbit)
Xenarthra
(armadillo, anteater)
Proboscidea
(elephant)
Sirenia
(manatee, dugong)
Hyracoidea
(hyrax)
Tubulidentata
(aardvark)
Afrosoricida
(tenrec, golden mole)
Macroscelidea
(elephant shrew)
Marsupialia
(opossum, kangaroo)
Pegasoferae
Scrotifera
Laurasiatheria
Euarchontoglires
Glires
Paenungulata
Afrotheria
Boreoeutheria
Pegasoferae, an unexpected mammalian clade
revealed by tracking ancient retroposon insertions
Hidenori Nishihara, Masami Hasegawa, and Norihiro Okada


Phylogeny of Eutheria
真獣類の系統樹マンダラ
Xenarthra
Afrotheria
Boreotheria

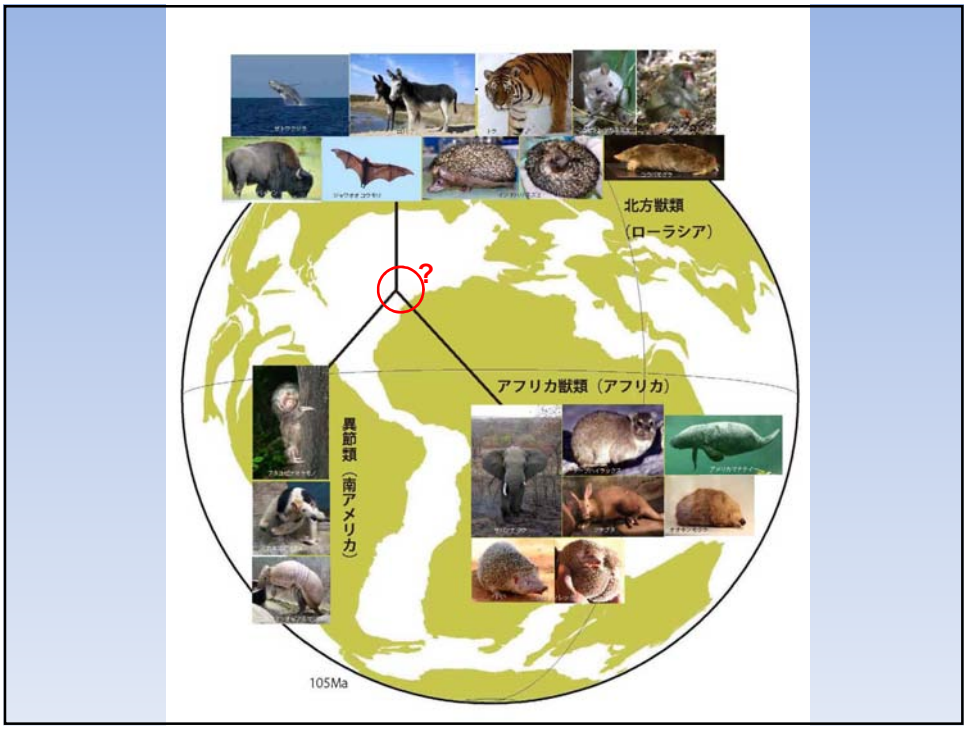


Genome Biology 2007, 8:R199

Rooting the eutherian tree: the power and pitfalls of phylogenomics

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

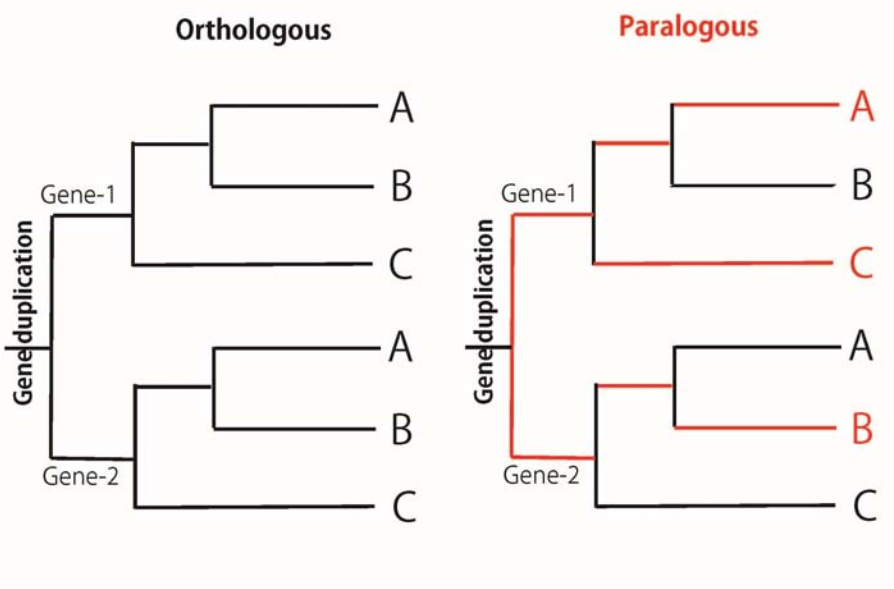
Addresses: *Graduate School of Bioscience and Biotechnology, Tokyo Institute of Technology, 4259-B-21 Nagatsuta-cho, Midori-ku, Yokohama 226-8501, Japan. †Department of Statistical Modeling, Institute of Statistical Mathematics, 4-6-7 Minami-Azabu, Minato-ku, Tokyo 106-8569, Japan. ‡School of Life Sciences,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



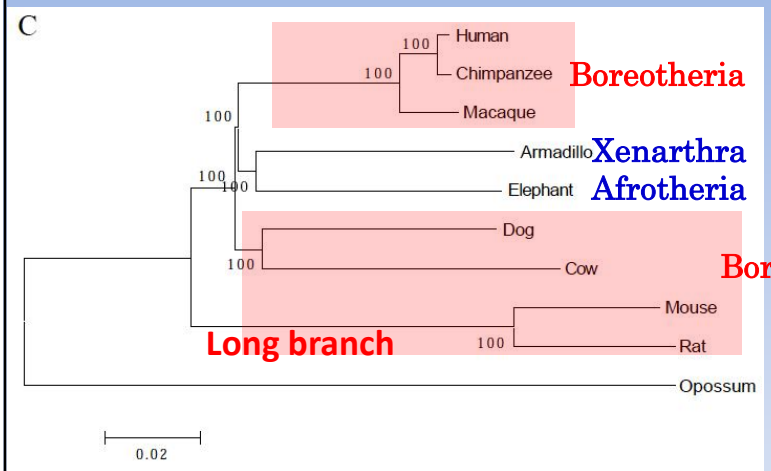
Collection of exon sequences from database

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)



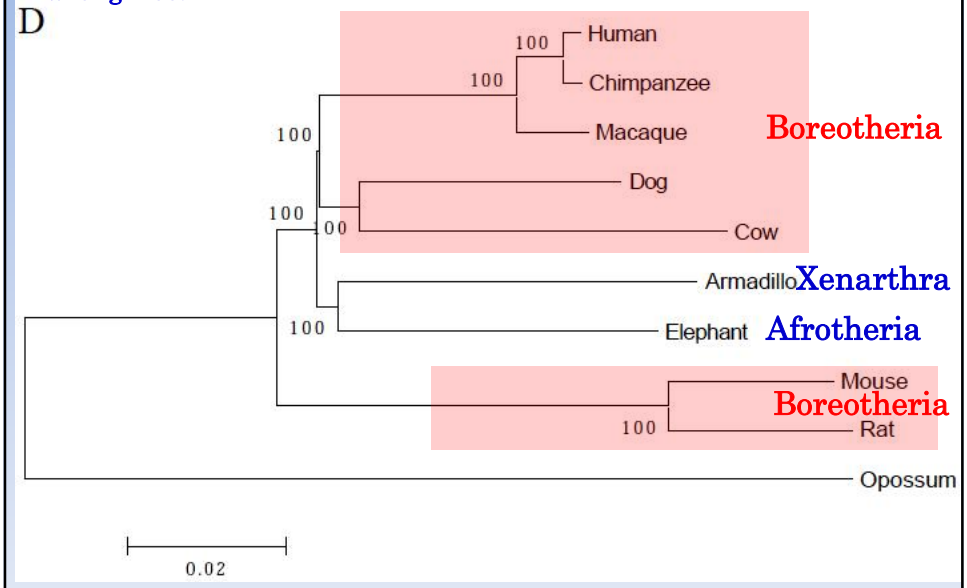
NJ tree of concatenated 1,011,870 nucleotides



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.

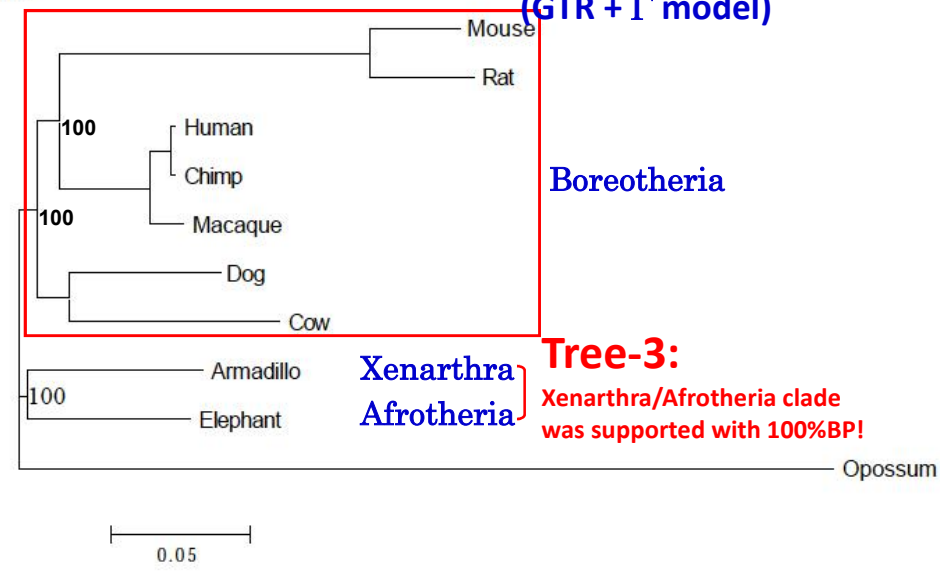
NJ tree of concatenated 337,290 amino acids

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)

A



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	$< \ln L > (\Delta \ln L \pm SE)$	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

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:

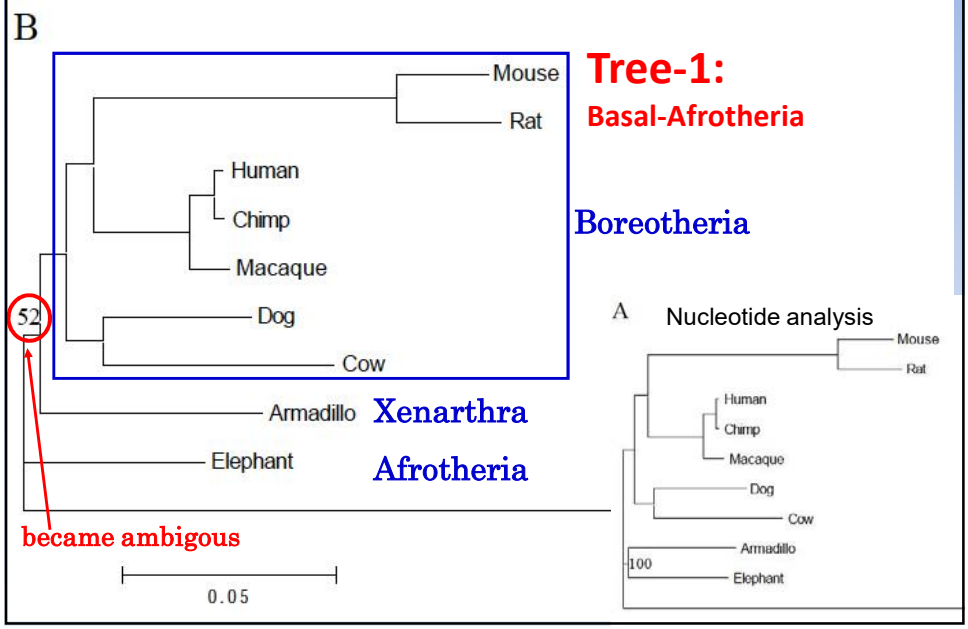
1b) Unequal rate among codon positions (GTR + Γ_8 model)ロドン座位間のrateの違い考慮

Tree	$< \ln L > (\Delta \ln L \pm SE)$	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

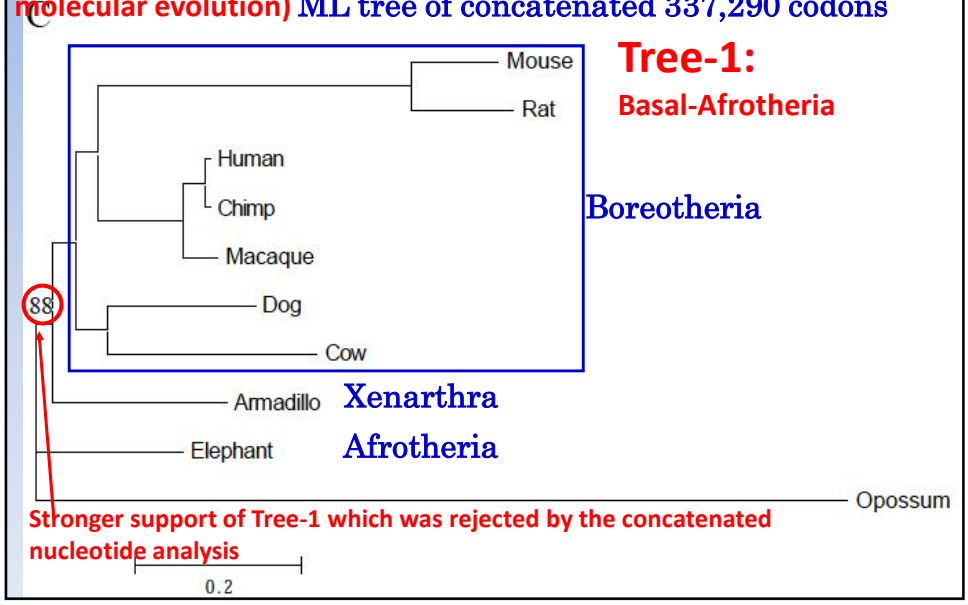
Tree	1st	2nd	3rd	$< \ln L > (\Delta \ln L \pm SE)$
1	−22.7 ± 16.6	−28.5 ± 15.1	−62.7 ± 25.1	−113.9 ± 33.6
2	−33.1 ± 15.6	−44.2 ± 13.6	−54.0 ± 25.4	−131.3 ± 32.8
3	<−1,029,462.5>	<-850,474.9>	<-2,039,573.6>	<−3,919,511.0>

But how about the amino acid analysis?

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



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



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)

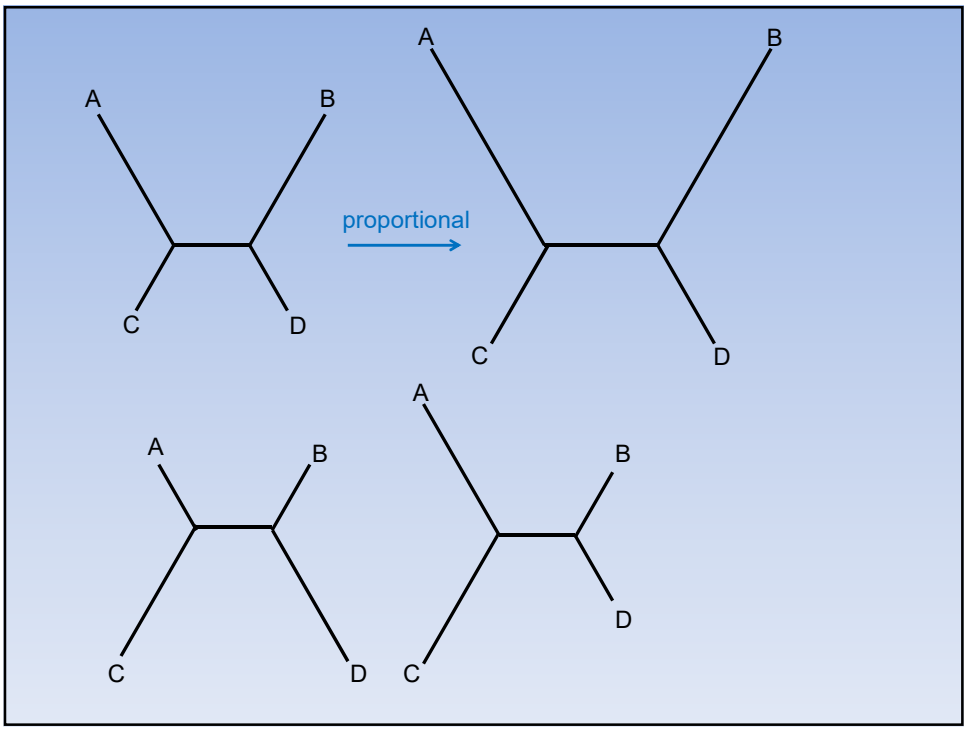
Tree	$< \ln L > (\Delta \ln L \pm SE)$	KH	wSH	BP (%)	#p	AIC
1	−117.2 ± 31.1	0.000	0.000	0.0		
2	−147.3 ± 29.7	0.000	0.000	0.0		
3	<−4,076,316.3>			100.0	26	8,152,684.6

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 + Γs 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	-			

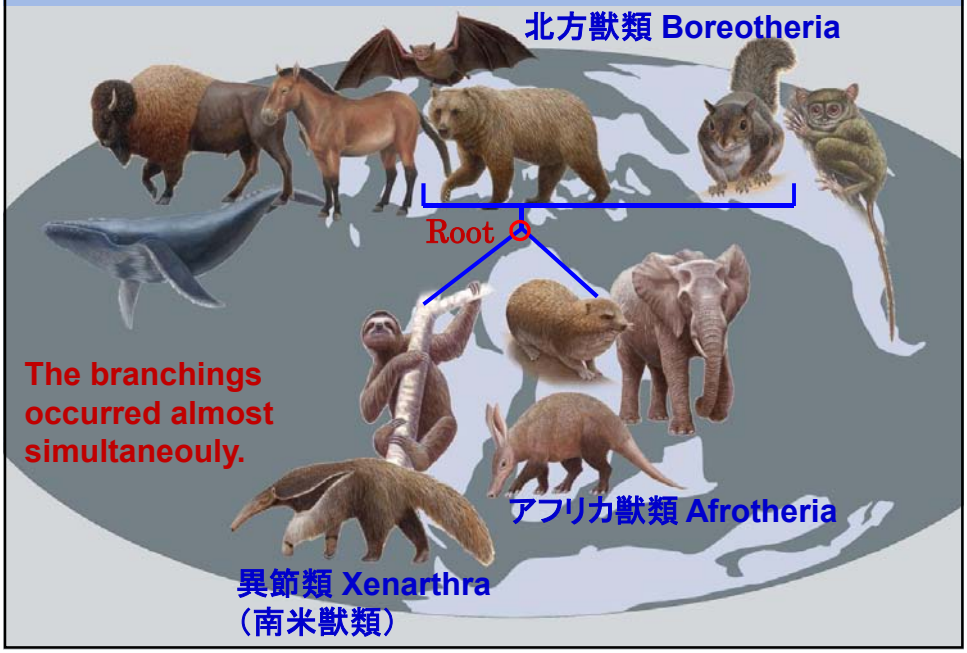
TSD, number of nucleotides in TSD. Alignments of all the loci are shown in Fig. S2.
*Two L1 loci reported by Kriegs et al. (5).
†Two L1 loci reported by Murphy et al. (17).

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

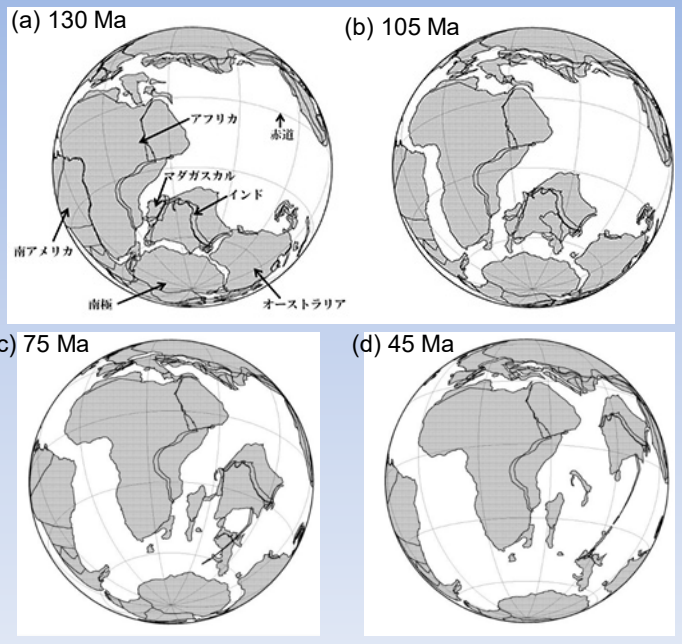
Hidenori Nishihara[a], Shigenori Maruyama[b], and Norihiro Okada[a,1]

(2009) PNAS

Configuration of continents 105Ma



Breakup of Gondwana



Longer sequences or more taxa?

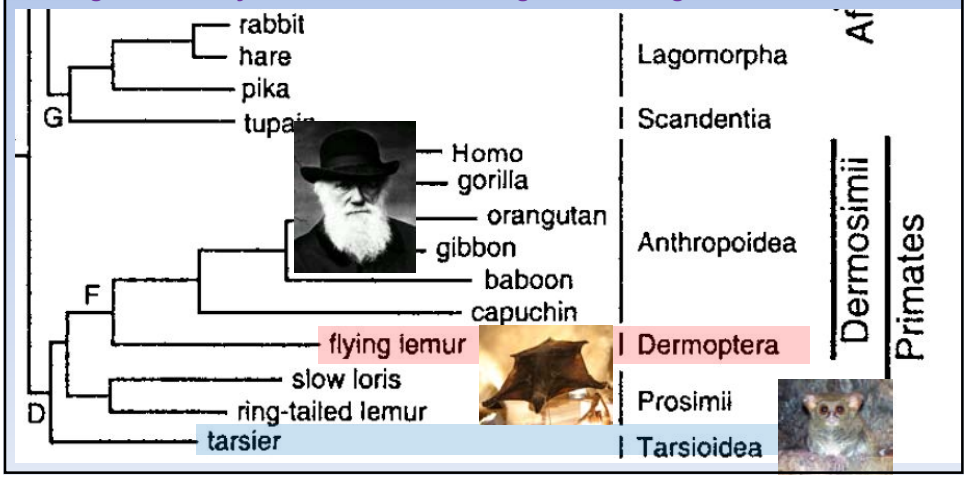
- The model misspecification gives a serious effect when taxon sampling is sparse.
- On the other hand, dense taxon sampling could relieve bised estimation by model misspecification.

Mammalian mitogenomic relationships and the root of the eutherian tree

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

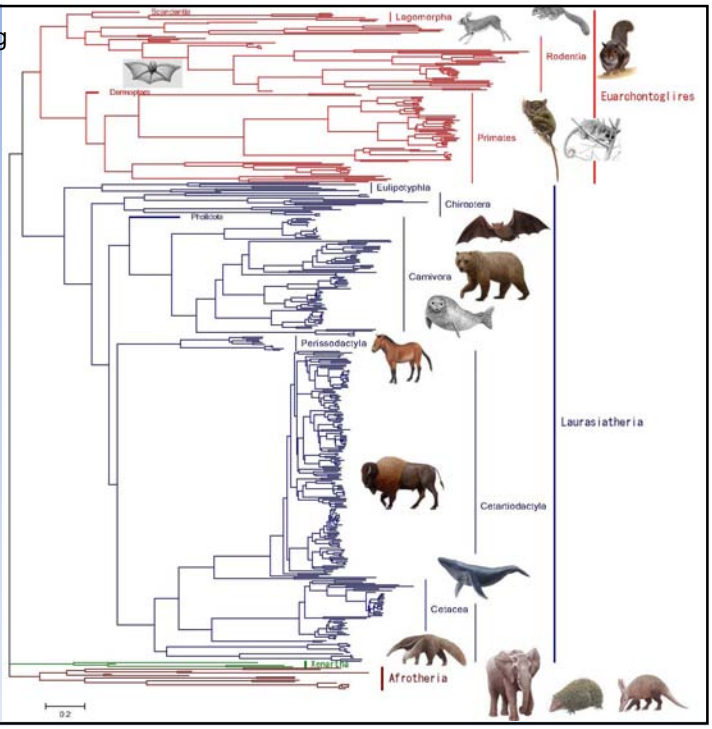
Ulfur Arnason*†, Joseph A. Adegoke*‡, Kristina Bodin*, Erik W. Born§, Yuzine B. Esa¶, Anette Gullberg*, Maria Nilsson*, Roger V. Short‖, Xiufeng Xu*,**, and Axel Janke*

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

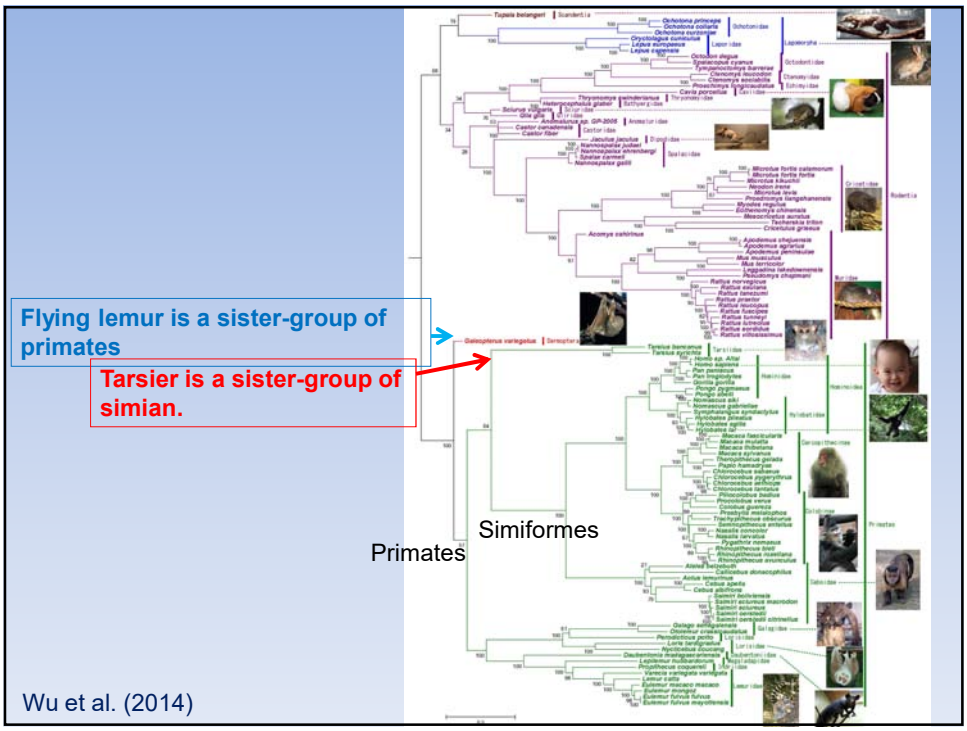


Dense taxon sampling could give a correct tree.

463 species tree of Eutheria by RAxML



Wu et al. (2014)

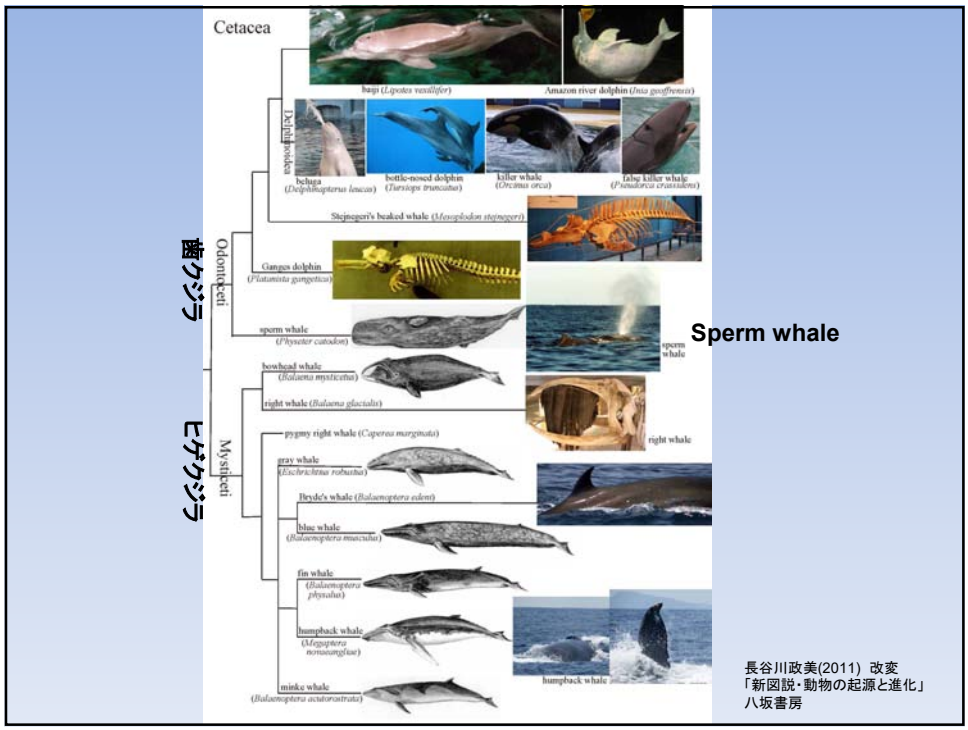


Wu et al. (2014)

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)



長谷川政美(2011) 改変
「新図説・動物の起源と進化」
八坂書房