

Phylogenomic Data Analyses Provide Evidence that Xenarthra and Afrotheria Are Sister Groups

Björn M. Hallström, Morgan Kullberg, Maria A. Nilsson, and Axel Janke

Department of Cell and Organism Biology, Division of Evolutionary Molecular Systematics, University of Lund, Lund, Sweden

The phylogenetic positions of the 4 clades, Euarchontoglires, Laurasiatheria, Afrotheria, and Xenarthra, have been major issues in the recent discussion of basal relationships among placental mammals. However, despite considerable efforts these relationships, crucial to the understanding of eutherian evolution and biogeography, have remained essentially unresolved. Euarchontoglires and Laurasiatheria are generally joined into a common clade (Boreoeutheria), whereas the position of Afrotheria and Xenarthra relative to Boreoeutheria has been equivocal in spite of the use of comprehensive amounts of nuclear encoded sequences or the application of genome-level characters such as retroposons. The probable reason for this uncertainty is that the divergences took place long time ago and within a narrow temporal window, leaving only short common branches. With the aim of further examining basal eutherian relationships, we have collected conserved protein-coding sequences from 11 placental mammals, a marsupial and a bird, whose nuclear genomes have been largely sequenced. The length of the alignment of homologous sequences representing each individual species is 2,168,859 nt. This number of sites, representing 2840 protein-coding genes, exceeds by a considerable margin that of any previous study. The phylogenetic analysis joined Xenarthra and Afrotheria on a common branch, Atlantogenata. This topology was found to fit the data significantly better than the alternative trees.

Introduction

Ongoing genome sequencing and analyses of large expressed sequence tag (EST) libraries have aided in recent progress toward molecular resolution of the most basal parts of the tree of placental mammals, notably the relationship among the 4 clades, Afrotheria, Xenarthra, Laurasiatheria, and Euarchontoglires (as defined by Murphy, Elzirik, Johnson et al. 2001). Laurasiatheria and Euarchontoglires are generally identified as sister groups, forming the so-called Boreoeutheria (Murphy, Elzirik, O'Brien et al. 2001; Springer and de Jong 2001; Waddell et al. 2001). However, the relationship between Boreoeutheria, Afrotheria, and Xenarthra has proven to be difficult to establish. Currently 3 different hypotheses exist, each favoring one of the 3 possible alternative topologies in the rooted eutherian tree.

The traditional morphological hypothesis favors a basal divergence between Xenarthra and remaining eutherians, Epitheria (McKenna 1975; Novacek 1992; Shoshani and McKenna 1998). This hypothesis has also been supported by a recent study employing retroposons as phylogenetic markers (Kriegs et al. 2006), although only 2 relevant markers were found in support of this particular basal split. The authors concluded, however, that while the 2 retroposons provide strong support for the Epitheria hypothesis, their presence might not be sufficient to statistically reject other possible topologies. The commonly favored molecular tree has Afrotheria as a sister group to all remaining placental mammals and places Xenarthra as a sister group to Boreoeutheria (Murphy, Elzirik, O'Brien et al. 2001; Waddell et al. 2001; Amrine-Madsen et al. 2003). This hypothesis was also supported in a recent molecular study based on both protein-coding and noncoding genomic sequences (Nikolaev et al. 2007). The third hypothesis joins Xenarthra and Afrotheria on a common branch (Atlantogenata or Xenafrotheria) with a basal split between this group and

Boreoeutheria. This hypothesis received support in some studies of mitochondrial sequences (Waddell et al. 1999; Kjer and Honeycutt 2007) and in 2 recent nuclear studies, one of which was based on long interspersed nuclear element (LINE-1) sequences (Waters et al. 2007) and the other was based on the examination of insertion–deletion (indel) differences and 2 retroposon inserts (Murphy et al. 2007).

The instability of the basal portion of the eutherian molecular tree indicates that the divergences connected to this part of the tree took place within a limited period of time. It is therefore conceivable that resolution of these relationships will require large amounts of sequence data. Only this will overcome the disturbing influence of noise that has accumulated through the long separate evolution of the individual branches involved and that masks the few shared substitutions. The basal eutherian relationships were recently addressed by Nikolaev et al. (2007) in a study based on 205 kb of coding and 430 kb of conserved noncoding nuclear sequences. Despite this amount of data, the study resulted in conflicting hypotheses as the noncoding sequences found it equally likely that Afrotheria or Atlantogenata was the sister to remaining eutherians.

In order to examine basal placental mammal divergences in more detail with the aim to resolve their basal divergences, we have in the current study collected the yet largest amount of data of protein-coding genes from placental mammals that are represented by extensive amounts of genomic sequences. We maximized both the amount of homologous sequences and the number of taxa. The root of the tree was established by using the corresponding sequences of a marsupial and a bird as outgroup.

Methods

The protein-coding data for 13 species were retrieved from the Ensembl FTP site (Release 41), and the RefSeq database of human RNA sequences were downloaded from the National Center for Biotechnology Information FTP site. The species used in this study include: human (*Homo sapiens*), rat (*Rattus norvegicus*), mouse (*Mus musculus*), rabbit (*Oryctolagus cuniculus*), dog (*Canis familiaris*), cow (*Bos taurus*), chimpanzee (*Pan troglodytes*), rhesus

Key words: Atlantogenata, phylogenomics, placental mammals, phylogeny, Xenarthra, Afrotheria.

E-mail: bjorn.hallstrom@cob.lu.se.

Mol. Biol. Evol. 24(9):2059–2068. 2007

doi:10.1093/molbev/msm136

Advance Access publication July 13, 2007

monkey/macaque (*Macaca mulatta*), armadillo (*Dasypus novemcinctus*), elephant (*Loxodonta africana*), small Madagascar hedgehog/tenrec (*Echinops telfairi*), opossum (*Monodelphis domestica*), and chicken (*Gallus gallus*). The last 2 species were used as outgroups to the placental mammals. Putatively orthologous sequences from all species were identified by using the reciprocal best-hit method (Rivera et al. 1998) with the BlastN search algorithm requiring an expected value (*E* value) no larger than 1×10^{-12} , as implemented in the program EST-e-mate (Hallström B, Janke A, in preparation). The sequences were translated to the corresponding amino acids (aa) and multiple sequence alignments (MSA) were created using ClustalW as implemented in EST-e-mate. Columns with gaps in the alignment were removed. Using the aa alignments, the corresponding nucleotide (nt) alignments and their reading frames were deduced. All alignments with an observed nt distance above 30% between any 2 species were discarded in order to compensate for possible inaccuracies of the MSAs at high sequence divergence. This approach also reduces analytical problems related to multiple substitutions and further ensures to include only orthologous sequences in the analysis. All alignments were inspected manually, and obviously erroneous alignments were either corrected or removed from the analysis. The sequences that are included in the alignment were classified using the Protein Analysis THrough Evolutionary Relationships (PANTHER) classification system (Thomas et al. 2003) by submitting a list of accession numbers for the human orthologs to the “Batch search” provided in the Web interface <http://www.pantherdb.org> (Mi et al. 2007).

From these alignments, 2 main data sets were created for analyses. One maximizes the sequence data for all species and minimizes missing data. This MSA has sequence data for all 13 species and comprises 1,961 aligned sequences (1,569,141 nt). It is named “maxspe.” The other MSA allows missing data in the alignment and maximizes the amount of sequence data instead, allowing for missing sequences. These could involve the chicken and one placental species from a clade for which genomic data for 2 or more species were available. Thus, sequences of cow or dog, elephant or tenrec, and mouse or rat were allowed to be missing. This data set comprises 2,840 (2,168,859 nt) aligned sequences and is named “maxgen”. Accession numbers for human orthologues of the maxgen data set are given in Supplementary Materials online.

These 2 data sets, maxspe and maxgen, were analyzed using maximum likelihood (ML) as implemented in Treefinder (TF; Jobb et al. 2004), Tree-Puzzle 5.2 (TP; Strimmer and von Haeseler 1996), or PAML 3.15 (Yang 1997) with the general time reversible (GTR) model of substitution (Lanave et al. 1984) and rate heterogeneity for nt data (Yang 1994), as suggested by Modeltest (Posada and Crandall 1998) and the Jones-Taylor-Thomton (JTT) model with rate heterogeneity for aa data, as suggested by ProtTest (Abascal et al. 2005). Four discrete categories of gamma distribution and one invariable, $4\Gamma + I$, were used in the analyses using TP and TF. Eight discrete categories of gamma distribution, 8Γ , were used when the analysis was done in PAML because the program package lacks the option for invariable sites. Bootstrap support values were cal-

culated for 100 replicates using TF. Jackknife analysis was done by creating 100 replicates with 99% of the maxspe data set randomly removed. These replicates had lengths of approximately 15 kb and were analyzed with PHYML (Guindon and Gascuel 2003) using the same models as for TP and TF. Different tree topologies were statistically evaluated using the Shimodaira-Hasegawa test (pSH; Shimodaira and Hasegawa 1999) and 1-sided Kishino-Hasegawa test (pKH; Goldman et al. 2000), as well as a confidence value for Expected Likelihood Weight (cELW; Strimmer and Rambaut 2002) as implemented in TP.

A partitioned ML analysis was done in TF. The alignments were partitioned into 3 partitions, based on codon position, for which the GTR and rate heterogeneity parameters were estimated separately. Finally, the sequences were also analyzed using Bayesian inference with MrBayes 3.1.2 (Huelsenbeck and Ronquist 2001), running for 2,000,000 Markov chain Monte Carlo (MCMC) generations with 1 cold and 3 heated chains, discarding the first 200,000 generations as burn-in and then sampling each 100th tree.

In order to determine whether the data conformed to compositional homogeneity, a chi-square test as implemented in TP was performed on the base frequencies for each species. To further investigate if the compositional heterogeneity between different lineages had any effect on the tree reconstruction, we recoded the nt sequences to R (purine) for A and G sites and to Y (pyrimidine) for T and C sites. Although a lot of information is lost when R/Y coding is used, the differences in base composition between lineages are greatly decreased. Thus, this procedure is merely used to detect possible reconstruction artifacts that may be caused by compositional biases rather than to replace the standard analysis. Additionally, the data were analyzed using the PAML program baseml with a nonstationary model (nhomo = 3), which allows for different base composition on different branches (Yang and Roberts 1995).

The 2 data sets from Nikolaev et al. (2007), available as supporting information with the paper, were downloaded and analyzed using the same ML methods as for our own data. This was done to produce information about these data sets that were not explicitly reported in the paper, notably pairwise distances between species. A sliding-window approach was applied to study the distribution of sequence distances throughout the full maxgen alignment and the aa sequence alignment of Nikolaev et al. (2007). A window of 200 aa was used, in which the observed distance were calculated, where after the window was moved 100 positions forward and the same procedure was repeated until the end of the alignment was reached.

Relative branch lengths were calculated against the branch leading to the elephant. This branch was arbitrarily chosen because from the aa ML analysis it appears to be moderately fast evolving.

Divergence time estimates were done with r8s (Sanderson 2002) and the multidivtime (MDT; Thorne and Kishino 2002) program package (<http://abacus.gene.ucl.ac.uk/>). For optimizations of the r8s divergence time estimations, the nonparametric rate smoothing method and the Powell algorithm were used on the branch lengths of the ML tree from TP (Sanderson 2002). Also, 10 different starting conditions were used to avoid local optima. For

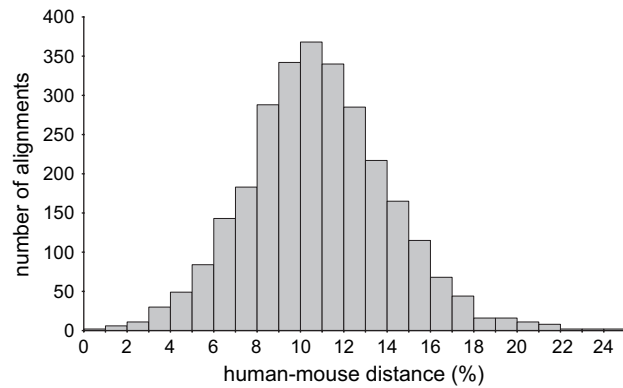


FIG. 1.—Distribution of the observed nt human–mouse pairwise distances for single-nt alignments of the maxgen MSA.

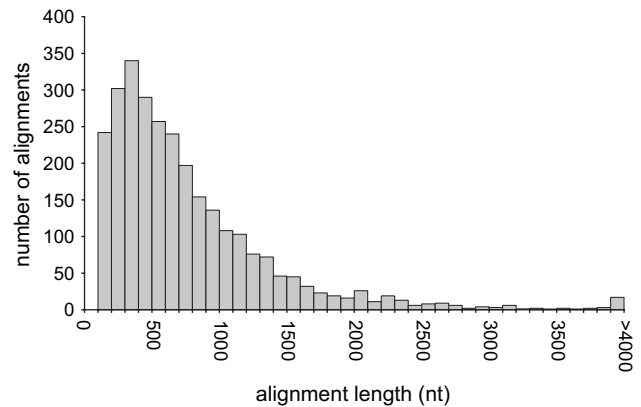


FIG. 2.—Distribution of the lengths of single alignment of nt data in the maxgen MSA.

calculating standard deviations (SDs), 100 bootstrap replicates of the sequence data were created using SEQBOOT. The branch lengths for each replicate were estimated using TP under the appropriate rate heterogeneity model and the tree given in figure 3. For each of those tree replicates, the divergence times were estimated by the r8s program and SD were calculated from the different results.

A parametric Bayesian analysis for divergence time estimates as implemented in MDT was made on nt and aa sequence data of the maxgen data, using 8 gamma categories of variable sites. A total of 1,000,000 generations of MCMC estimates for dates and rates parameters were sampled after an equally long burn-in. Besides from the calibration points below the prior expected age of the root was set to 135 and its standard deviation to 15. All analyses were performed several times, verifying convergence between the chains. For estimates of the placental mammal divergence times, several constraints have been applied for calibrating the molecular clock. These are marked in figure 6 and are taken from Benton and Donoghue (2007).

Results

Prior to the phylogenetic analyses, the general properties of the data set, such as distances, alignment lengths, and nt and aa compositions, were examined. The distribution of nt distances for the individual human–mouse alignments is shown in figure 1. These distances, which were generally low, were evenly spread around the mean. The average human–mouse nt distance was $10.8\% \pm 3.3\%$. Exclusion of sequences with high E values reduces distances and the influence of multiple substitutions, which can become problematic at larger distances. The average length of the individual alignments was 763 ± 612 nt. The distribution of the alignment lengths is shown in figure 2. The great majority of alignments are between 300 and 4000 nt in length. Classification of genes according to the PANTHER system and the number of genes in each class are shown in table 1.

The sequences were examined with respect to possible base frequency biases between lineages. The observed base frequencies are shown in table 2. None of the species passed at the 5% significance level of a chi-square test on the homogeneity of base composition, thereby suggesting that the differences in base composition were systematic. When all

3 codon positions were recoded to R and Y, the compositional bias was significantly decreased (table 2). After R/Y recoding, only 4 species (cow, rabbit, tenrec, and chicken) did not pass the chi-square test at the 5% significance level. Comparing results of the phylogenetic analyses of the R/Y recoded with the native sequence data allows examination of how the compositional bias may affect tree reconstruction. The composition of aa sequences was less heterogeneous. Only the outgroup taxa and the tenrec showed significantly deviating aa compositions.

ML aa trees were constructed under a JTT $4\Gamma + I$ model of sequence evolution. The GTR $4\Gamma + I$ model was applied to nt sequences, whereas R/Y coded sequences were analyzed using a 2-state (Felsenstein 1981) $4\Gamma + I$ model. All ML analyses resulted in the same tree topology irrespective of the use of different data sets or program packages. ML bootstrap values for internal branches were in all instances 100%. The aa ML tree is shown in figure 3. The partitioned likelihood analysis yielded the same tree as shown in figure 3 and the same relative branch lengths.

The 3 possible trees that can be constructed for Xenarthra, Afrotheria, and Boreoeutheria plus outgroup are shown in figure 4. The 3 trees represent the Atlantogenata hypotheses (Xenarthra is the sister group to Afrotheria), the Epitheria hypothesis (Xenarthra is the sister group to all remaining placental mammals), and finally the Afrotheria hypothesis (Afrotheria is the sister group to all remaining placental mammals). The 3 alternatives were compared in an extended ML analysis. Table 3 shows likelihood values, pSH, pKH, and cELW values, for the 3 different topologies, as calculated by TF using nt sequences. The aa sequences were also submitted to the corresponding analyses (table 4). All ML tests of both nt and aa sequences with both TF and TP conclusively identified the Atlantogenata hypothesis as being the most likely, whereas the 2 alternative topologies, Epitheria and Afrotheria, were significantly rejected by all ML tests. Bayesian analysis of both nt and aa sequences using MrBayes identified the Atlantogenata hypothesis with probability (P) 1.0 on all branches.

ML analysis of RY-coded nt sequences gave rise to the same tree topology as obtained in the analysis of unmodified nt sequences. In this case too, the alternative trees were significantly rejected in all analyses. Also the nonstationary

Table 1
Classification of Biological Function of the Human Orthologs, according to PANTHER Database

Function	No of Genes
Biological process unclassified	638
mRNA transcription regulation	204
Protein phosphorylation	156
Proteolysis	132
Cell proliferation and differentiation	129
Cell structure	122
Neurogenesis	105
Developmental processes	84
Cell cycle control	78
Cation transport	76
Transport	74
Cell adhesion	73
G protein-mediated signaling	72
Cell motility	68
Other metabolism	64
Protein biosynthesis	63
General vesicle transport	58
mRNA transcription	54
mRNA splicing	50
Intracellular signaling cascade	49
Signal transduction	49
Protein modification	47
Cell communication	47
Ligand-mediated signaling	45
Cell adhesion-mediated signaling	43
Apoptosis	43
Intracellular protein traffic	42
Other intracellular signaling cascade	41
DNA replication	40
Receptor protein tyrosine kinase signaling pathway	39
Nucleoside, nucleotide, and nucleic acid metabolism	37
Cell structure and motility	37
MAPKKK cascade	36
Chromatin packaging and remodeling	34
Endocytosis	34
Muscle development	33
Protein glycosylation	33
Oncogenesis	33
Muscle contraction	33
Stress response	32
Mesoderm development	32
Other neuronal activity	29
Other receptor-mediated signaling pathway	29
Embryogenesis	28
Immunity and defense	28
Mitosis	28
Cell cycle	28
Protein folding	26
Other polysaccharide metabolism	26
Chromosome segregation	26
Vision	25
Lipid metabolism	25
Protein targeting	25
Calcium-mediated signaling	24
Cytokinesis	23
DNA repair	21
Phosphate metabolism	21
Nerve-nerve synaptic transmission	21
Oncogene	21
Exocytosis	21
Phospholipid metabolism	21
Anion transport	21
Cell surface receptor-mediated signal transduction	20
Receptor-mediated endocytosis	20
Neurotransmitter release	19
Miscellaneous	19
Skeletal development	19
Protein metabolism and modification	18

Table 1
Continued

Function	No of Genes
Inhibition of apoptosis	18
Small molecule transport	17
Other developmental process	17
Neuronal activities	17
Induction of apoptosis	17
Tumor suppressor	16
Nuclear transport	15
Other intracellular protein traffic	15
Metabolism of cyclic nucleotides	15
Carbohydrate metabolism	15
Receptor protein serine/threonine kinase signaling pathway	14
Protein targeting and localization	14
Segment specification	14
Oogenesis	14
Phagocytosis	14
Electron transport	14
Protein complex assembly	13
mRNA transcription initiation	13
Synaptic transmission	12
Carbohydrate transport	12

model, as implemented in PAML (nhomo = 3), found the Atlantogenata hypothesis (tree-1) as being the most likely. The nonstationary model uses different frequency parameters for different branches and thus allows analysis of standard (ATGC) nt sequences, even in the case of compositional heterogeneity. The congruent results in both these analyses indicate that the compositional heterogeneity of the nt data has only limited effect on the tree reconstruction. Despite significant deviations in nt composition in many lineages of placental mammals, the ML topology remained the same and differences in relative branch lengths were also limited. This consistency is important in regard to the molecular estimates of divergence times.

Parallel to the ML analyses a jackknife analysis was performed with 100 replicates of 1% of the maxspe data set (~15 kb). Of the 100 samples, 84 showed the Atlantogenata topology, whereas 11 supported the Afrotheria tree and 5 the Epitheria hypothesis. Other branches remained unchanged with a support of 99–100%. Extended jackknife analysis showed that ≈30% of the nt-coding and ≈55% of the aa-coding maxgen alignments were needed for providing statistically significant support for the best hypothesis (Atlantogenata).

Analysis of nt composition showed that the tenrec deviated distinctly from the nt composition of other placental mammals studied. This bias, which also affected the aa sequence composition, remained after recoding the nt sequences to R and Y. In order to study the effect that this bias might have on the tree reconstruction, the ML analyses were repeated after excluding the tenrec from the data set. The removal of the tenrec had only a minimal influence on the result by slightly increasing the statistical support for the Atlantogenata hypothesis.

The effect of taxon sampling on tree construction was examined in additional ML analyses that included the same species as used in the genomic study of Cannarozzi et al. (2007). Thus, this part of the current study was limited to the sequences of cow, dog, mouse, rat, human, with chicken

Table 2
The Frequencies of nt Usage for the Individual Species; the Frequencies of R and Y for the R/Y-Recorded Data; and the Probability Values of the chi-square test on Compositional Heterogeneity for Nucleotides, R/Y-Coded nt, and aa

Species	$f(A)$	$f(T)$	$f(G)$	$f(C)$	$f(R)$	$f(Y)$	$P(nt)$, %	$P(R/Y)$, %	$P(aa)$, %
Chicken	28.6	24.0	24.6	22.8	53.2	46.8	0.00	0.00	0.00
Opossum	29.1	24.9	23.7	22.2	52.8	47.2	0.00	5.94 ^a	0.02
Chimpanzee	28.5	24.3	24.3	22.9	52.8	47.2	0.00	47.78 ^a	99.34 ^a
Human	28.5	24.3	24.3	22.9	52.8	47.2	0.00	43.23 ^a	99.83 ^a
Macaque	28.5	24.3	24.3	22.9	52.8	47.2	0.00	63.55 ^a	99.84 ^a
Mouse	27.6	23.4	25.1	23.9	52.7	47.3	0.00	23.14 ^a	53.85 ^a
Rat	27.5	23.3	25.1	24.1	52.7	47.3	0.00	11.42 ^a	24.53 ^a
Rabbit	27.5	23.1	25.2	24.2	52.7	47.3	0.00	3.40	17.47 ^a
Cow	27.7	23.4	24.9	23.9	52.6	47.4	0.00	0.79	45.35 ^a
Dog	28.2	24.0	24.5	23.3	52.7	47.2	0.01	68.46 ^a	99.53 ^a
Armadillo	28.3	24.1	24.3	23.2	52.7	47.3	0.00	23.13 ^a	93.73 ^a
Elephant	28.2	24.0	24.5	23.3	52.7	47.3	0.00	33.48 ^a	73.30 ^a
Tenrec	27.2	22.9	25.3	24.6	52.5	47.5	0.00	0.00	0.03

^a Species for which compositional homogeneity could not be rejected at a 5% level of significance, according to the chi-square test, and are therefore assumed to be homogeneous.

and/or a marsupial used as outgroup. All analyses performed with rate homogeneity models on the 2.3 mega-base maxgen data set identified rodents as sister to remaining placental mammals, that is, the same tree as in Cannarozzi et al. (2007). However, when a rate heterogeneity ($4\Gamma + I$) model was used, a grouping of human and rodents on the same branch became the most strongly supported alternative, although other topologies could not be rejected. The effect of adding more taxa was studied by the stepwise addition of species belonging to the orders Lagomorpha (rabbit), Proboscidea (elephant), Xenarthra (armadillo), and Tenrecidae (tenrec). When the sequences of the rabbit, the elephant, and the armadillo or the tenrec were included, the topology became consistent with the Boreoeutheria hypothesis, that is (primates, [rodents, rabbit], [cow, dog]),

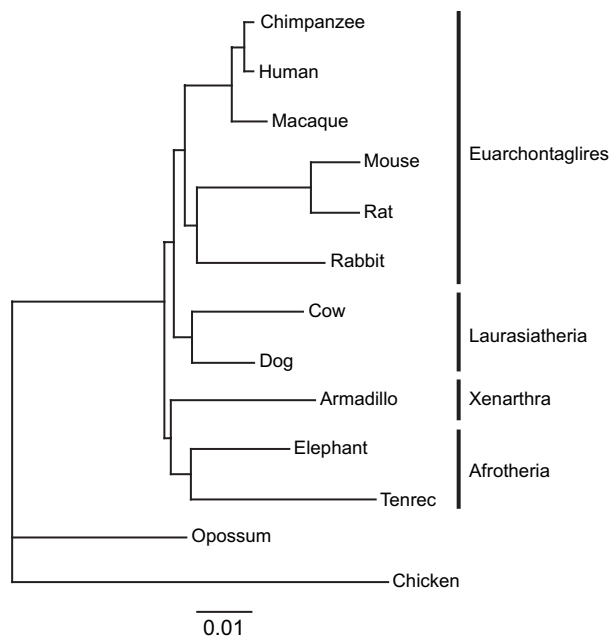


FIG. 3.—The phylogenetic tree, with scaled branch lengths, reconstructed for 722,953 aa sites (maxgen MSA) under the JTT $4\Gamma + I$ model of sequence evolution.

with the elephant and the armadillo or the tenrec as sister group to the former. This general topology did not change with the addition of the chimpanzee and macaque, which increased the taxon sampling on the primate branch.

A change of outgroup did not affect the topology of the tree of placental mammals. However, the use of the opossum as a single outgroup yielded stronger support to the Atlantogenata hypothesis than did the analyses with the chicken as the only outgroup.

Reanalysis of the data used by Nikolaev et al. (2007) reproduced topologies and statistical values that were consistent with their original results. However, examination of pairwise distances among the nt sequences of the protein-coding genes of that study showed that these were approximately twice as large as those of the maxgen and maxspe alignments, whereas the noncoding data of Nikolaev et al. (2007) had just slightly lower distance values. Translation of protein-coding sequences to aa sequences resulted in suspiciously high distance values, especially relative to the outgroups. The distances along the aa alignment were analyzed in more detail by a pairwise comparison using a sliding window. Figure 5 shows the distribution of uncorrected human–opossum distances in the maxgen alignment (black line), whereas the gray line shows the corresponding values for the data set used by Nikolaev et al. (2007). As evident in figure 5, the distances within each data set differ to some extent. Thus, the distances within the maxgen alignment have a maximum of about 25% (very few have distances exceeding 15%) as a result of the restrictions for sequence retrieval and alignments. The aa alignment used by Nikolaev et al. (2007) shows a high proportion (26%) of distance values that exceed 30% and a nonnegligible amount (5%) of extremely high distances, 50–71%. Our ML analyses also confirmed that their noncoding data set could not distinguish between the Afrotheria and Atlantogenata hypotheses.

Estimation of divergence times by MDT proved difficult because the program apparently does not allow for more than 1×10^6 characters. This limitation was overcome by splitting the data into alignments with a length of 800,000 characters. Divergence time estimates from

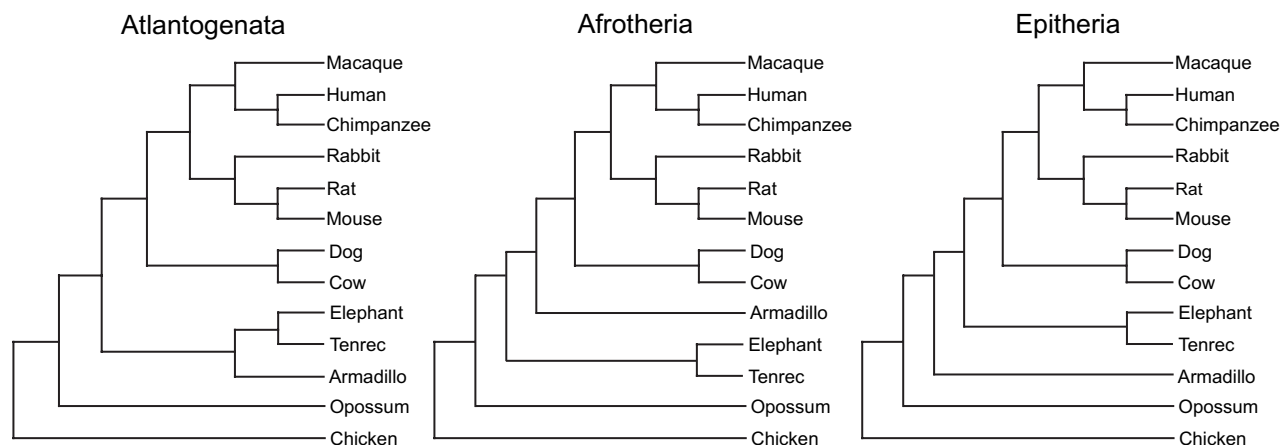


FIG. 4.—The trees representing the 3 hypotheses on the evolution of the placentals.

r8s and MDT for nt and aa are shown in figure 6 together with the constraints included. Despite the use of different data sets and approaches, the estimates do not differ markedly from each other, except for the basal divergences of Glires and Euarchontoglires. All estimates from MDT fall somewhere within the corresponding constraints. The error values for each estimate are small and are not reported here in order to avoid overconfidence in the estimates. It is probable that the statistics behind the estimates may not be optimized to sets of data corresponding to those of the current study, but this should not invalidate the fact that estimates based on large sets of data should in general be more precise than those based on smaller ones.

Discussion

One of the final questions of the relationship of placental mammal orders and higher clades concerns their most basal divergence. These basal relationships are also fundamental to all discussion of mammalian evolution. Recent phylogenetic analyses have commonly recognized 3 basal groups of placental mammals, Afrotheria, Xenarthra, and Boreoeutheria, but the relationships among the 3 groups have tended to be differently and inconclusively resolved (Waddell et al. 1999, 2001; Murphy, Elzirik, Johnson et al. 2001; Murphy, Elzirik, O'Brien et al. 2001; Murphy et al. 2007; Arnason et al. 2002; Amrine-Madsen et al. 2003; Kriegs et al. 2006; Kjer and Honeycutt 2007; Nikolaev et al. 2007). The problem associated with resolving these divergences indicates that they took place within a narrow temporal window, presumably within only a few million years. The current phylogenomic analyses of the 2.2

mega-base maxgen and the 1.6 mega-base maxspe data sets of 11 species of placental mammals unambiguously supported the Atlantogenata hypothesis, that is, a basal split between Afrotheria/Xenarthra and the remaining placentals, by significantly rejecting the 2 competing hypotheses. It appears that sequence data several times larger than that of the phylogenomic analysis of Nikolaev et al. (2007) were needed to conclusively resolve this phylogenetic question.

The effect of compositional heterogeneity of extraordinarily long sequence alignments on the phylogenetic analyses has been a vexed issue in many analyses. In the current study, the base composition was shown to be heterogeneous between several lineages. Although the percentages of A, G, C, and T appeared to be similar at first sight, a chi-square test of the long sequence alignments rejected homogeneity of nt composition for all species. Recoding the nt sequences to R and Y removed the bias for many placental species except for the cow, rabbit, and tenrec. The phylogenetic analysis of the R/Y recoded sequences or analyses using a nonstationary model on the unaltered data resulted in the same and strongly supported topology. Also, excluding species with a compositional bias did not alter the phylogenetic relationships among the basal placental mammals. This indicates that compositional heterogeneity did not affect the tree reconstruction based on this set of data. It could be argued that the chi-square test for rate heterogeneity is unnecessarily stringent for data sets of this size because a difference in base frequency of only a few per mille will cause the test to reject compositional homogeneity. In any case, the results suggested that differences at this level were not likely to adversely affect tree reconstruction in the current study.

Table 3
ML Support and Probability Values for the 3 Hypotheses, nt Data Sets

Topology	Maxgen				Maxspe			
	$\Delta\log L/\text{Standard Error (SE)}$	pSH	pKH	cELW	$\Delta\log L/\text{SE}$	pSH	pKH	cELW
1	[−9375031.47]	1.0000	1.0000	1.0000	[−6813126.01]	1.0000	1.0000	1.000
2	5.35	0.0000	0.0000	0.0000	5.20	0.0000	0.0000	1.000
3	8.18	0.0000	0.0000	0.0000	7.77	0.0000	0.0000	1.000

Table 4
ML Support and Probability Values for the 3 Hypotheses, aa Data Sets

Topology	Maxgen				Maxspe			
	$\Delta\log L/SE$	pSH	pKH	cELW	$\Delta\log L/SE$	pSH	pKH	cELW
1	[−3448629.01]	1.0000	1.0000	1.0000	[−2424209.14]	1.0000	1.0000	1.0000
2	3.43	0.0000	0.0000	0.0000	4.29	0.0000	0.0000	0.0000
3	0.49	0.0000	0.0000	0.0000	7.14	0.0000	0.0000	0.0000

Prior to the availability of genome data, the phylogeny of placental mammals was investigated with alignments that were 10–15 kbp long (Murphy, Elzirik, Johnson et al. 2001; Murphy, Elzirik, O’Brien et al. 2001; Arnason et al. 2002). However, jackknife analyses with 15 kbp of the data showed that while the presumably correct tree could be resolved in 84% of the replicates, the support for the deepest divergences was not significant. High Bayesian posterior probabilities are often reported for these short sequence data sets (e.g., Murphy, Elzirik, O’Brien et al. 2001), but they may overestimate the factual support (Simmons et al. 2004). Our analyses showed that far larger amounts of sequences were required to significantly resolve these deep splits.

Somewhat unexpectedly, the phylogenetic results did not agree with those of Nikolaev et al. (2007), despite their use of a total of 635,000 nt of protein-coding sequences and conserved noncoding sequences. The noncoding data set of Nikolaev et al. (2007) yielded equal likelihood values for the Atlantogenata and Afrotheria hypotheses. However, the Atlantogenata hypothesis was rejected in favor of the Afrotheria hypothesis in analyses of the aa data. The data set used by Nikolaev et al. (2007) showed a basal position of Xenartha and Afrotheria relative to the remaining placental mammals, but the support for the relative relationship between Xenartha, Afrotheria, and remaining placentals remained ambiguous. Their alignment contains to some extent missing data, and a considerable portion of the aa sequence shows distances to the outgroup that are almost as high as in random sequences. This may suggest that at least some of these sequences could have been erroneously aligned or may be nonhomologous. In regions show-

ing large sequence distances, the outgroup will function as a random outgroup sequence and make the tree reconstruction prone to artifacts (Sullivan and Swofford, 1997).

Hitherto, analyses of rare genomic events such as insertion of retroposons and indel differences have not provided unambiguous support for any of the 3 hypotheses related to basal relationships among placental mammals. Retroposons are a robust and well-understood tool for studying mammalian evolution and have clarified several phylogenetic issues in mammalian evolution (Schmitz et al. 2005; Kriegs et al. 2006). Yet, only a few retroposons have been identified for examining basal mammalian divergences. The statistics behind retrotransposition has not been fully developed, but it has been suggested that 3 such events are needed to provide “significant” support for a particular phylogenetic hypothesis (Waddell et al. 2001). Other genomic events such as “rare” indel events are still far from being understood for their usefulness for phylogenetic reconstruction. Four such events have been reported in favor of the Atlantogenata hypothesis (Murphy et al. 2007), but 2 of these, ZNF367 exon 5 and LAMC2 exon 13, show signs of homoplasy.

Analyses of large data sets of protein-coding sequences (Jorgensen et al. 2005; Kullberg et al. 2006; Cannarozzi et al. 2007) have suggested that rodents may not group with primates as postulated by the Euarchontoglires hypothesis. Changing the taxon sampling in the 2.2 mega-base “maxseq” alignment led to the same result when rate homogeneity models were used, whereas application of rate heterogeneity models placed rodents and primates on a common branch. The latter topology became further stabilized with the inclusion of additional taxa, for example, rabbit and elephant. In this case, the findings were in contrast to the study of Rokas and Carroll (2005) on yeast phylogenomics, which showed that the inclusion of additional taxa could decrease the support for what was taken as being the correct topology. The results suggest that large amounts of sequence data in conjunction with appropriate taxon sampling are crucial to the recovery of a strongly supported placental mammalian tree. Recently, Hedtke et al. (2006) and Gatesy et al. (2007) addressed the topic of taxon sampling and the amounts of sequence data required in phylogenomics. As discussed by Gatesy et al. (2007), exclusion of “problematic” taxa, could promote the recovery of congruent trees, even with small amounts of data. However, problematic taxa can usually only be identified in hindsight. In the current study, inclusion of the rabbit stabilized the boreoeutherian branch and similarly the elephant neutralized the adverse effect of the fast evolving tenrec, thereby stabilizing Afrotheria and resulting in a more strongly supported tree.

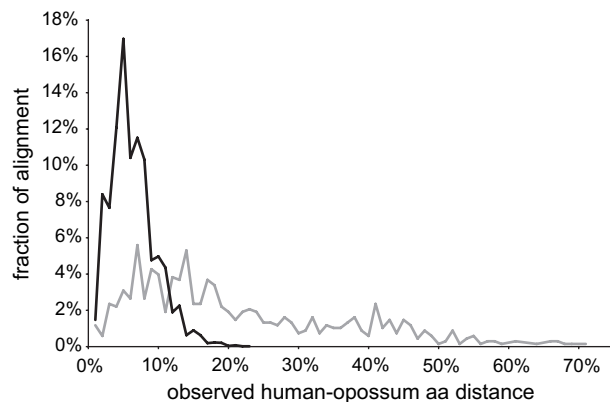


FIG. 5.—Distribution of aa distances between human and opossum calculated with a sliding window over the full alignments. The black line represents the maxgen MSA and the gray line represents the aa alignment of Nikolaev et al. (2007).

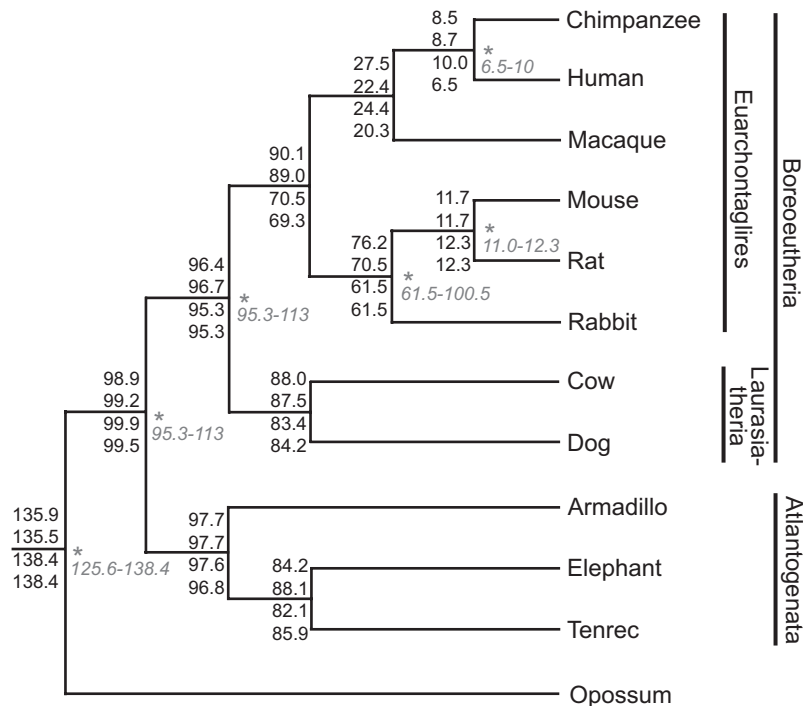


FIG. 6.—Tree with divergence time estimates. The 4 dates correspond to: MDT nt, MDT aa, r8s nt, and r8s aa (top to bottom). Constrained nodes are marked with stars and the upper and lower constrain dates are shown in gray italics.

The phylogenetic nomenclature related to the superorders Boreoeutheria, Laurasiatheria, Atlantogenata, and Afrotheria has been derived from the biogeography of extant mammalian orders (Waddell et al. 1999; Murphy, Elzirik, O'Brien et al. 2001). This is problematic, however, partly because the superorders overlap spatially and partly because the names, implicitly and explicitly, imply that some groups (Xenarthra and Afrotheria) originated on the supercontinent Gondwana in the southern hemisphere (Murphy, Elzirik, O'Brien et al. 2001). Assumptions about the origin of living groups based on their current biogeography may not be warranted, however, because they ignore the mammalian fossil record. Two recent studies have pointed out that the most parsimonious assumption for basal divergences among placental mammals is that they took place in Laurasia (Archibald 2003; Hunter and Janis 2006).

The molecular dating of the origin of Xenarthra at ≈ 97 MYA is consistent with other studies and the fossil record, which have indicated that certain orders were already present in the Cretaceous (Asher et al. 2005; Benton and Donoghue 2007). The stem group Xenarthra has probably not originated in South America but rather migrated during later stages in its evolution. The primary argument for this reasoning is that before the Cretaceous–Tertiary (K/T) boundary at 65 MYA, South America was only the home of archaic mammalian groups, such as multituberculates, gondwanatheres, and sudamericids (Flynn and Wyss 1998; Pascual and Ortiz-Jaureguizar 2007). At the K/T boundary, there was a major turnover in the South American mammalian community. The archaic mammals became extinct and were replaced by therian mammals (Pascual and Ortiz-Jaureguizar 2007). This faunal change correlates well

with molecular estimates of the origin and diversification of modern lineages of the 2 major South American mammalian groups, marsupials at 68.5 MYA and xenarthrans at 65 MYA (Delsuc et al. 2004; Nilsson et al. 2004). There is fossil evidence suggesting that stem group marsupials radiated in North America and some migrated to South America using a land bridge (the Aves ridge route) that connected the 2 continents across the Caribbean in the late Cretaceous (Case et al. 2005).

The strongly supported sister group relationship between Afrotheria and Xenarthra may serve to reduce the discussion of the biogeography of early placental mammals to 2 particular topics, Atlantogenata and Boreoeutheria. The name Atlantogenata is connected to the opening up of the Atlantic Ocean, implying that this event led to the divergence of Xenarthra and Afrotheria by vicariance (Waddell et al. 1999). The depauperate mammalian African fossil record covering the period 95–70 MYA does not suggest an existence of Xenarthra in Africa during this period, but does not exclude it either. Similarly, and more strikingly, there are no South American fossils of this age that can be related to Xenarthra or Afrotheria, despite the occurrence of other mammalian fossils on this continent. This circumstance strongly suggests that xenarthrans and other therian mammals colonized South America from the north (Laurasia) in the late Cretaceous. Likewise, the fossil record does not support an African origin of the Afrotheria because the oldest members of crown group Afrotheria, are also of Laurasian origin (Asher et al. 2003; Zack et al. 2005; Tabuce et al. 2007).

The current analyses of the relatively conserved 2.2 mega-base data set of protein-coding sequences have

yielded strong support to a basal divergence between Atlantogenata and remaining placental mammals. It would be advantageous to confirm these results by sequence independent data such as retroposon insertion as in Kriegs et al. (2006) or other rare genomic events (Boore 2006). However, the temporal narrowness of the intervals may affect the usability of those characters because incomplete lineage sorting within narrow temporal windows may interfere with these approaches (Nishihara et al. 2006).

Provided the phylogenetic results reflect the true tree and the molecular estimates are reasonably accurate, the findings place the divergence between Atlantogenata and remaining placentals (Boreoeutheria) at ≈ 99 –100 MYA. The split between Xenarthra and Afrotheria is estimated to have taken place ≈ 2 Myr later (i.e., 97–98 MYA), whereas the basal divergence within Boreoeutheria was estimated to have taken place ≈ 95 –96 MYA and those within Laurasiatheria at 84–88 MYA. These divergence dates are somewhat younger than those estimated earlier on 10–15 kbp of nuclear and mitochondrial data (Arnason and Janke 2002; Murphy et al. 2007). Irrespective of these differences, all these estimates pinpoint to the limited temporal interval within which the basal diversification of placental mammals took place. The time intervals of 2–3 Myr between major branching events correspond to or are only slightly larger than the typical duration of mammalian speciation (Curnoe et al. 2006; van Dam et al. 2006). This provides an explanation to the fact that data sets of lesser sizes may have tended, for stochastic reasons, to resolve these divergences in different ways and illustrates the need of very large sequence data or very specific rare genomic events to resolve such divergences that occurred about 100 MYA.

Supplementary Material

Supplementary Materials are available at *Molecular Biology and Evolution* online (<http://www.mbe.oxfordjournals.org/>). GenBank Accession number are included.

Acknowledgments

The Swedish Science Council, the Jörgen Lindström Foundation, and the Royal Physiographic Society (Nilsson-Ehle) supported the study. We thank Torbjörn Säll for discussion and advice on the statistical properties of large sequence data and Ulfur Arnason for comments on the manuscript.

Literature Cited

- Abascal F, Zardoya R, Posada D. 2005. ProtTest: selection of best-fit models of protein evolution. *Bioinformatics*. 21:2104–2105.
- Amrine-Madson H, Koepfli KP, Wayne RK, Springer MS. 2003. A new phylogenetic marker, apolipoprotein B, provides compelling evidence for eutherian relationships. *Mol Phylogenet Evol*. 28:225–240.
- Archibald JD. 2003. Timing and biogeography of the eutherian radiation: fossils and molecules compared. *Mol Phylogenet Evol*. 28:350–359.
- Arnason U, Adegoke JA, Bodin K, et al. (11 co-authors) 2002. Mammalian mitogenomic relationships and the root of the eutherian tree. *Proc Natl Acad Sci USA*. 99:8151–8156.
- Arnason U, Janke A. 2002. Mitogenomic analyses of eutherian relationships. *Cytogenet Genome Res*. 96:20–32.
- Asher J, Meng J, Wible JR, McKenna MC, Rougier GW, Dashzeveg D, Novacek MJ. 2005. The antiquity of Glires. *Science*. 307:1091–1094.
- Asher RJ, Novacek MJ, Geisler JH. 2003. Relationships of endemic African mammals and their fossil relatives based on morphological and molecular evidence. *J Mamm Evol*. 10:131–162.
- Benton MJ, Donoghue CJ. 2007. Paleontological evidence to date the tree of life. *Mol Biol Evol*. 24:26–53.
- Boore JL. 2006. The use of genome-level characters for phylogenetic reconstruction. *Trends Ecol Evol*. 21:439–446.
- Cannarozzi G, Schneider A, Gonnert G. 2007. A phylogenomic study of human, dog and mouse. *PLoS Comput Biol*. 3:e2.doi:10.1371/journal.pcbi.0030002.
- Case JA, Goin FJ, Woodburne MO. 2005. “South American” marsupials from the Late Cretaceous of North America and the origins of marsupial cohorts. *J Mamm Evol*. 12:461–494.
- Curnoe D, Thorne A, Coate JA. 2006. Timing and tempo of primate speciation. *J Evol Biol*. 19:59–65.
- Delsuc F, Vizcaino SF, Douzery EJ. 2004. Influence of Tertiary paleoenvironmental changes on the diversification of South American mammals: a relaxed molecular clock study within xenarthrans. *BMC Evol Biol*. 4:11.
- Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol*. 17:368–376.
- Flynn JJ, Wyss AR. 1998. Recent advances in South American mammalian paleontology. *Trends Ecol Evol*. 13:449–454.
- Gatesy J, DeSalle R, Wahlberg N. 2007. How many genes should a systematist sample? Conflicting insights from a phylogenomic matrix characterized by replicated incongruence. *Syst Biol*. 56:355–363.
- Goldman N, Anderson JP, Rodrigo AG. 2000. Likelihood-based tests of topologies in phylogenetics. *Syst Biol*. 49:652–670.
- Guindon S, Gascuel S. 2003. A simple, fast and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol*. 52:696–704.
- Hedtke S, Townsend T, Hillis D. 2006. Resolution of phylogenetic conflict in large data sets by increased taxon sampling. *Syst Biol*. 55:522–529.
- Huelsenbeck JP, Ronquist F. 2001. MrBayes: Bayesian inference of phylogeny. *Bioinformatics*. 17:754–755.
- Hunter PJ, Janis CM. 2006. Spiny Norman in the garden of eden? *J Mamm Evol*. 13:89–123.
- Jobb G, von Haeseler A, Strimmer K. 2004. TREEFINDER: a powerful graphical analysis environment for molecular phylogenetics. *BMC Evol Biol*. 4:18.
- Jorgensen FG, Hobolth A, Hornshøj H, Bendixen C, Fredholm M, Schierup MH. 2005. Comparative analysis of protein coding sequences from human, mouse and the domesticated pig. *BMC Biol*. 3:2.
- Kjer KM, Honeycutt RL. 2007. Site specific rates of mitochondrial genomes and the phylogeny of eutheria. *BMC Evol Biol*. 7:8.
- Kriegs JO, Churakov G, Kieffmann M, Jordan U, Brosius J, Schmitz J. 2006. Retroposed elements as archives for the evolutionary history of placental mammals. *PLoS Biol*. 4:e91.
- Kullberg M, Nilsson MA, Arnason U, Harley EH, Janke A. 2006. Housekeeping genes for phylogenetic analysis of eutherian relationships. *Mol Biol Evol*. 23:1493–1503.
- Lanave C, Preparata G, Saccone C, Serio G. 1984. A new method for calculating evolutionary substitution rates. *J Mol Evol*. 20:86–93.

- McKenna MC. 1975. Toward a phylogenetic classification of the mammalia. In: Lockett WP, Szalay FS, editors. *Phylogeny of the primates*. New York: Plenum Press. p. 21–46.
- Mi H, Guo N, Kejariwal A, Thomas PD. 2007. PANTHER version 6: protein sequence and function evolution data with expanded representation of biological pathways. *Nucleic Acids Res.* 35:D247–D252.
- Murphy WJ, Eizirik E, Johnson WE, Zhang YP, Ryder OA, O'Brien SJ. 2001. Molecular phylogenetics and the origins of placental mammals. *Nature.* 409:614–618.
- Murphy WJ, Eizirik E, O'Brien SJ, et al. (11 co-authors). 2001. Resolution of the early placental mammal radiation using Bayesian phylogenetics. *Science.* 294:2348–3235.
- Murphy WJ, Pringle TH, Crider TA, Springer MS, Miller W. 2007. Using genomic data to unravel the root of the placental mammal phylogeny. *Genome Res.* doi:10.1101/gr.5918807.
- Nikolaev S, Montoya-Burgos JI, Margulies EH, Program NCS, Rougemont J, Nyffeler B, Antonarakis S. 2007. Early history of mammals is elucidated with the ENCODE multiple species sequencing data. *PLoS Genet.* 3:1.
- Nilsson MA, Arnason U, Spencer PB, Janke A. 2004. Marsupial relationships and a timeline for marsupial radiation in South Gondwana. *Gene.* 340:189–196.
- Nishihara H, Hasegawa M, Okada N. 2006. Pegasoferae, an unexpected mammalian clade revealed by tracking ancient retroposon insertions. *Proc Natl Acad Sci USA.* 103:9929–9934.
- Novacek MJ. 1992. Mammalian phylogeny: shaking the tree. *Nature.* 356:121–125.
- Pascual R, Ortiz-Jaureguizar E. 2007. The Gondwanan and South American episodes: two major and unrelated moments in the history of the South American mammals. *J Mamm Evol.* doi:10.1007/s10914-007-9039-5.
- Posada D, Crandall KA. 1998. Modeltest: testing the model of DNA substitution. *Bioinformatics.* 14:817–818.
- Rivera MC, Jain R, Moore JE, Lake JA. 1998. Genomic evidence for two functionally distinct gene classes. *Proc Natl Acad Sci USA.* 95:6239–6244.
- Rokas A, Carroll SB. 2005. More genes or more taxa? The relative contribution of gene number and taxon number to phylogenetic accuracy. *Mol Biol Evol.* 22:1337–1344.
- Sanderson MJ. 2002. R8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics.* 19:301–302.
- Schmitz J, Roos C, Zischler H. 2005. Primate phylogeny: molecular evidence from retroposons. *Cytogenet Genome Res.* 108:26–37.
- Shimodaira H, Hasegawa M. 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Mol Biol Evol.* 16:1114–1116.
- Shoshani J, McKenna MC. 1998. Higher taxonomic relationships among extant mammals based on morphology, with selected comparisons of results from molecular data. *Mol Phylogenet Evol.* 9:572–584.
- Simmons MP, Pickett KM, Miya M. 2004. How meaningful are Bayesian support values? *Mol Biol Evol.* 21:188–199.
- Springer MS, de Jong WW. 2001. Which mammalian supertree to bark up? *Science.* 291:1709–1711.
- Strimmer K, Rambaut A. 2002. Inferring confidence sets of possibly misspecified gene trees. *Proc Biol Sci.* 269:137–142.
- Strimmer K, von Haeseler A. 1996. Quartet puzzling: a quartet maximum likelihood method for reconstructing tree topologies. *Mol Biol Evol.* 13:964–969.
- Sullivan J, Swofford DL. 1997. Are guinea pigs rodents? The importance of adequate models in molecular phylogenetics. *J Mamm Evol.* 4:77–86.
- Tabuce R, Marivaux L, Adaci M, Bensalah M, Hartenberger JL, Mahboubi M, Mebrouk F, Tafforeau P, Jaeger JJ. 2007. Early tertiary mammals from North Africa reinforce the molecular Afrotheria clade. *Proc Biol Sci.* 274:1159–1166.
- Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, Daverman R, Diemer K, Muruganujan A, Narechania A. 2003. PANTHER: a library of protein families and sub-families indexed by function. *Genome Res.* 13:2129–2141.
- Thorne JL, Kishino H. 2002. Divergence time and evolutionary rate estimation with multilocus data. *Syst Biol.* 51:689–702.
- van Dam JA, Abdul Aziz H, Alvarez Sierra MA, Hilgen FJ, van den Hoek Ostende LW, Lourens LJ, Mein P, van der Meulen AJ, Pelaez-Campomanes P. 2006. Long-period astronomical forcing of mammal turnover. *Nature.* 443:687–691.
- Waddell PJ, Cao Y, Hasegawa M, Mindell DP. 1999. Assessing the cretaceous superordinal divergence times within birds and placental mammals by using whole mitochondrial protein sequences and an extended statistical framework. *Syst Biol.* 48:119–137.
- Waddell PJ, Kishino H, Ota R. 2001. A phylogenetic foundation for comparative mammalian genomics. *Genome Inform.* 12:141–154.
- Waters PD, Dobigny G, Waddell PJ, Robinson TJ. 2007. Evolutionary history of LINE-1 in the major clades of placental mammals. *PLoS ONE.* 2:e158. doi:10.1371/journal.pone.0000158.
- Yang Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J Mol Evol.* 39:306–314.
- Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci.* 13:555–556.
- Yang Z, Roberts D. 1995. On the use of nucleic acid sequences to infer early branchings in the tree of life. *Mol Biol Evol.* 12:451–458.
- Zack SP, Penkrot TA, Bloch JI, Rose KD. 2005. Affinities of 'hyopsodontids' to elephant shrews and a Holarctic origin of Afrotheria. *Nature.* 434:497–501.

Arndt von Haeseler, Associate Editor

Accepted June 25, 2007