RESEARCH ARTICLES

Confirming the Phylogeny of Mammals by Use of Large Comparative Sequence Data Sets

Arjun B. Prasad,‡ Marc W. Allard,§ NISC Comparative Sequencing Program*† and Eric D. Green*†*

*Genome Technology Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD; †NIH Intramural Sequencing Center, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD; ‡Integrated Biosciences Program, George Washington University; and §Department of Biological Sciences, George Washington University

The ongoing generation of prodigious amounts of genomic sequence data from myriad vertebrates is providing unparalleled opportunities for establishing definitive phylogenetic relationships among species. The size and complexities of such comparative sequence data sets not only allow smaller and more difficult branches to be resolved but also present unique challenges, including large computational requirements and the negative consequences of systematic biases. To explore these issues and to clarify the phylogenetic relationships among mammals, we have analyzed a large data set of over 60 megabase pairs (Mb) of high-quality genomic sequence, which we generated from 41 mammals and 3 other vertebrates. All sequences are orthologous to a 1.9-Mb region of the human genome that encompasses the cystic fibrosis transmembrane conductance regulator gene (*CFTR*). To understand the characteristics and challenges associated with phylogenetic analyses of such a large data set, we partitioned the sequence data in several ways and utilized maximum likelihood, maximum parsimony, and Neighbor-Joining algorithms, implemented in parallel on Linux clusters. These studies yielded well-supported phylogenetic trees, largely confirming other recent molecular phylogenetic analyses. Our results provide support for rooting the placental mammal tree between Atlantogenata (Xenarthra and Afrotheria) and Boreoeutheria (Euarchontoglires and Laurasiatheria), illustrate the difficulty in resolving some branches even with large amounts of data (e.g., in the case of Laurasiatheria), and demonstrate the valuable role that very large comparative sequence data sets can play in refining our understanding of the evolutionary relationships of vertebrates.

Introduction

Advances in large-scale DNA sequencing are creating new opportunities for molecular phylogeneticists to examine ever-larger amounts of genomic sequence from increasing numbers of taxa. These data have the potential to greatly enhance our ability to answer difficult phylogenetic questions; however, the size and inherent imperfections of such data sets present some unique challenges for accurate tree inference. To begin with, the large numbers of characters that serve as input demand a robust computational infrastructure. Further, the fast-evolving nature of most eukaryotic genomes has yielded large amounts of nonprotein-coding sequences that are not conserved across species, making it difficult to generate complete and accurate multisequence alignments (Margulies et al. 2006).

These and other challenges in dealing with nucleotide sequence–based characters have prompted some to make phylogenetic inferences based on genomic characters that change less frequently than individual nucleotides, such as inversions, transposon insertions, and coding insertions/deletions (indels) (Shimamura et al. 1997; Murphy et al. 2004, 2007; Bashir et al. 2005; Chaisson et al. 2006; Kriegs et al. 2006). However, these genomic characters present their own challenges. First, they are less common, so few may be found to help differentiate short branches (Nishihara et al. 2005). This leads to particular difficulties when assessing support using traditional methods (e.g., bootstrapping). Second, assigning the actual character state (e.g., the presence of the same insertion at a given position or a given rearrangement shared between 2 species) can be difficult because overlapping rearrangements, changing boundaries, and/or sequence divergence can obscure the historical relationships (Murphy et al. 2004). Finally, although methods for modeling such rare genomic characters have been developed (Waddell et al. 2001; Chaisson et al. 2006), biases leading to the potential for homoplastic evolution are not well understood (Boissinot et al. 2004; Chen et al. 2005). For example, it is probable that indel events are less likely to occur independently in multiple lineages than single nucleotide changes; however, the extent of biases in indel location appears to vary among lineages and types of indel events. Thus, although rare genomic changes can be used as informative phylogenetic characters, there are still reasons why sequence-based characters are helpful as independent sources of phylogenetic information.

Meanwhile, traditional phylogenetic analyses based on nucleotide mutations present a different set of challenges. As the costs of procuring and operating large clusters of commodity computers have decreased, it has become increasingly practical to harness significant amounts of processing power to analyze very large sequence-based data sets. This provides the ability to exploit single nucleotide mutations more extensively, yielding more robust phylogenetic inferences. Additionally, there is extensive theory and experience relevant to both modeling the evolution of these characters and using the algorithms to infer phylogenetic trees. However, care must be taken to rule out sources of systematic (or nonstochastic) error, such as long-branch attraction, alignment guide trees, and base-composition biases that can hinder the use of such data sets (Kluge and Wolf 1993; Hillis et al. 2003; Philippe et al. 2005; Rokas and Carroll 2005).

Key words: Placentalia, Eutheria, Mammalia, mammalian phylogeny, phylogenomics, Atlantogenata, molecular systematics.

E-mail: egreen@nhgri.nih.gov.

Mol. Biol. Evol. 25(9):1795–1808. 2008
doi:10.1093/molbev/msn104
Advance Access publication May 2, 2008

Indeed, there remain a number of uncertainties about the mammalian phylogeny that are beginning to be clarified using both substitution-based and rare genomic character–based methods. For example, although recent molecular studies have broken the placental (eutherian) mammals into 4 groups or superorders—Afrotheria (Stanhope et al. 1998), Euarchontoglires (also called Supraprimates by Waddell et al. 2001), Laurasiatheria (Waddell, Okada, and Hasegawa 1999), and Xenarthra (Cope 1889)—the relative arrangement of these groups is uncertain (Waddell, Okada, and Hasegawa 1999; Madsen et al. 2001; Murphy, Eizirik, Johnson, et al. 2001; Waddell et al. 2001; Scally et al. 2002; Kriegs et al. 2006; Nishihara et al. 2006). Studies investigating the relationships among these 4 mammalian superorders agree that Euarchontoglires and Laurasiatheria should be grouped together to form Boreoeutheria (Murphy, Eizirik, O'Brien, et al. 2001; Springer and de Jong 2001; Waddell et al. 2001; Kriegs et al. 2006; Nishihara et al. 2006), although the exact root among the 3 remaining groups (Boreoeutheria, Afrotheria, and Xenarthra) remains unsettled. Most molecular analyses have weakly supported a basal Afrotherian root (Murphy, Eizirik, O'Brien, et al. 2001; Waddell et al. 2001; Scally et al. 2002; Amrine-Madsen et al. 2003; Waddell and Shelley 2003; Kriegs et al. 2006; Nikolaev et al. 2007; Nishihara et al. 2007), whereas a few have weakly supported the association of Afrotheria and Xenarthra to form Atlantogenata, thereby placing the root between Atlantogenata and Boreoeutheria (Waddell, Cao, et al. 1999; Madsen et al. 2001; Delsuc et al. 2002; Lin et al. 2002; Waddell and Shelley 2003; Hallstrom et al. 2007; Murphy et al. 2007; Waters et al. 2007). On the other hand, the traditional placement of Xenarthra at the root of the placental tree, with Boreoeutheria and Afrotheria together forming Epitheria, has been supported by others (Shoshani and McKenna 1998).

Several recent studies have further investigated these issues, leading to different conclusions. Kriegs et al. (2006) identified 2 shared transposon insertions in Afrotheria and Boreoeutheria that could not be found in Xenarthra or in an opossum outgroup, although they did not consider these results statistically significant. Nikolaev et al. (2007) used comparative sequence data generated for 1% of the human genome (as part of the ENCODE project) to examine the root of Placentalia; they reported significant support for a root between Afrotheria and Exafroplacentalia (Boreoeutheria + Xenarthra), though they found it necessary to perform separate analyses on conserved noncoding sequences and amino-acid sequences to exclude both other possible roots. Murphy et al. (2007) searched for informative coding indels within whole-genome sequence data, finding 4 examples supporting Atlantogenata as the root and none supporting the 2 alternative roots; they also identified 2 retroelement insertions with well-conserved flanking sequence that also support Atlantogenata as the root. In addition, Waters et al. (2007) analyzed a phylogeny of L1 sequences and found further support for Atlantogenata as the root. Hallstrom et al. (2007) and Wildman et al. (2007) used coding sequence extracted from whole-genome shotgun sequencing data to find support for an Atlantogenatan root; however, Nishihara et al. (2007) used a similar data set and found that the use of more complex models of evolution that partitioned individual genes suggested an Afrotherian root. Another unresolved issue in mammalian phylogenetics relates to the relationships among orders within Laurasiatheria. Though the monophyly of Laurasiatheria is fairly well established, the relative arrangements within this taxon have been difficult to establish (aside from placing Eulipotyphla at the base of Laurasiatheria) (Murphy, Eizirik, O'Brien, et al. 2001; Waddell et al. 2001; Arnason and Janke 2002; Chaisson et al. 2006; Nishihara et al. 2006).

To investigate some of the above issues, we sought to confirm and refine the mammalian phylogeny by examining the performance of nucleotide-based phylogenetic analyses using very large genomic sequence data sets. Specifically, we mapped, sequenced, assembled, and analyzed a large genomic region (~1.9 Mb in humans) in 41 mammals, 1 bird, and 2 fishes. Here, we report the experience and results of performing phylogenetic analyses of this large (>69 Mb) comparative sequence data set.

Methods

The comparative sequence data set analyzed here (available at http://www.nisc.nih.gov/data) is an expanded version of that reported by Thomas et al. (2003), with all sequences orthologous to a 1.9-Mb region of human chromosome 7 (build hg18, chr7:115,597,757–117,475,182) that includes 10 known genes (e.g., *CFTR*, *ST7*, and *CAV1*). All species' sequences were generated by first isolating bacterial artificial chromosome (BAC) clones from the orthologous genomic region using overgo-based hybridization methods (Thomas et al. 2002) and then generating high-quality sequence of each selected BAC (Blakesley et al. 2004). For each species, sets of overlapping BAC sequences were compiled into a single ordered and oriented sequence. The assembled BAC sequences are provided as supplementary data online (available at http://www.nisc.nih.gov/data). All the analyzed sequences were generated in this fashion as part of the NIH Intramural Sequencing Center Comparative Sequencing Program (Thomas et al. 2003), except for the human sequence (generated by the International Human Genome Sequencing Consortium [2001]) and the *Fugu* sequence (generated by the *Fugu* Genome Consortium [Aparicio et al. 2002]). Table 1 provides a list of the species whose sequences were analyzed.

The assembled sequences were aligned using the threaded blockset aligner (TBA), a local alignment program designed to generate multisequence alignments of large data sets (Blanchette et al. 2004). The final alignment size of all alignable sequence in the data set was 44 taxa by 6,270,442 characters. The initial alignment guide tree was based on the results of Murphy, Eizirik, Johnson, et al. (2001); this was then modified to test alternate hypotheses and to verify that the results were not dependent on the alignment guide tree (see Results and Discussion). The alignment was divided into partitions (i.e., corresponding subportions of the genomic region, as described below) using custom perl scripts (available on request).

Coding sequences were identified based on data from the Consensus CDS Project (http://www.ncbi.nlm.nih.gov/

Table 1
Multispecies Comparative Sequence Data Set

Clade	Scientific Name	Common Name	Total Sequence[a]	Coding[b]	Conserved Noncoding[c]
	Homo sapiens	Human	1,877,426	20,647	102,884
	Pan troglodytes	Chimpanzee	1,573,483	17,962	86,513
	Gorilla gorilla gorilla	Lowland gorilla	1,761,981	20,489	93,962
	Pongo pygmaeus abelii	Sumatran orangutan	1,478,010	18,344	80,548
	Hylobates gabriellae	Red-cheeked gibbon	2,154,624	20,122	97,708
	Colobus guereza	Black and white colobus	2,023,939	20,575	99,065
	Cercopithecus aethiops vervet	Vervet monkey	1,555,031	18,638	87,051
	Macaca mulatta	Rhesus macaque	1,678,549	20,569	92,538
Catarrhini	*Papio cynocephalus anubis*	Olive baboon	1,680,295	20,575	89,897
	Callithrix jacchus	White-tufted-ear marmoset	1,869,361	19,783	88,306
	Callicebus moloch	Dusky titi	1,810,674	18,263	84,974
	Aotus nancymai	Owl monkey	2,059,585	20,581	96,483
Platyrrhini	*Saimiri boliviensis boliviensis*	Bolivian squirrel monkey	1,695,311	16,692	67,548
	Otolemur garnettii	Small-eared galago	1,732,353	20,373	86,512
	Lemur catta	Ring-tailed lemur	1,399,362	20,545	84,060
Strepsirrhini	*Microcebus murinus*	Gray mouse lemur	1,541,029	19,103	86,239
	Rattus norvegicus	Norway rat	1,883,088	20,344	78,983
	Mus musculus	Mouse	1,486,509	19,079	73,094
	Cavia porcellus	Guinea pig	1,815,594	20,504	85,548
Rodentia	*Spermophilus tridecemlineatus*	13-lined ground squirrel	1,757,846	20,505	89,020
Lagomorpha	*Oryctolagus cuniculus*	New Zealand white rabbit	1,889,755	20,453	81,226
	Bos taurus	Cow	2,022,671	20,357	85,135
	Ovis aries	Sheep	1,816,302	20,149	76,197
	Muntiacus muntjak vaginalis	Indian muntjac	1,450,172	15,340	67,216
Cetartiodactyla	*Sus scrofa domestica*	Domestic pig	1,198,526	17,006	60,133
Perissodactyla	*Equus caballus*	Horse	1,423,288	17,580	75,633
	Felis catus	Cat	1,737,938	20,374	81,560
	Neofelis nebulosa	Clouded leopard	1,691,656	16,001	74,568
	Canis familiaris	Dog	1,317,853	16,374	69,142
Carnivora	*Mustela putorius furo*	Domestic ferret	1,494,791	20,456	75,743
	Carollia perspicillata	Seba's short-tailed bat	1,069,438	14,424	38,369
Chiroptera	*Rhinolophus ferrumequinum*	Greater horseshoe bat	1,684,815	20,495	85,118
	Atelerix albiventris	Middle-African hedgehog	1,985,767	20,081	72,111
Eulipotyphla	*Sorex araneus*	European common shrew	1,734,562	18,845	63,737
Xenarthra	*Dasypus novemcinctus*	Armadillo	1,454,970	16,850	59,554
	Loxodonta africana	African elephant	2,040,789	20593	87,812
Afrotheria	*Echinops telfairi*	Lesser hedgehog tenrec	1,765,269	18,087	74,734
	Didelphis virginiana	North American opossum	1,627,985	15,114	45,484
	Monodelphis domestica	Gray short-tailed opossum	1,174,555	12,480	33,565
Marsupialia	*Macropus eugenii*	Tammar wallaby	1,846,640	18,545	61,489
Monotremata	*Ornithorhynchus anatinus*	Duck-billed platypus	1,268,713	18,543	49,457
Aves	*Gallus gallus*	Chicken	744,025	19,934	32,648
	Tetraodon nigroviridis	*Tetraodon*	257,833	16,938	7,760
Actinopterygii	*Fugu rubripes*	*Fugu*	273,621	17,033	7,779
Total			69,805,984	825,745	3,217,103

[a] The total amount of assembled sequence (in bases) following removal of low-quality sequence and overlaps between BAC sequences (Thomas et al. 2003).
[b] The number of bases in the coding partition (see text for details).
[c] The number of bases in the conserved noncoding partition (see text for details).

CCDS), as provided on the UCSC Genome Browser (hg17 build; http://www.genome.ucsc.edu). This approach was used for all genes except *MET*, which was derived from the longest GENCODE annotation (http://genome.imim.es/gencode/) because it was not present in the Consensus CDS annotation. Coding regions within the multisequence alignment were manually edited, and areas of uncertain alignment were removed, with gap columns added where necessary to maintain phase using jalView 2.2.1 (Clamp et al. 2004). Codon position partitions were generated using every third base of the alignment.

Evolutionarily conserved sequences were identified using annotations represented on the "17-way Most Conserved" track of the UCSC Genome Browser (hg18 build) (Kent et al. 2002; Karolchik et al. 2004; Siepel et al. 2005). These annotations reflect conserved elements that were detected using phastCons (Siepel et al. 2005), which applies a 2-state conserved versus nonconserved phylogenetic hidden Markov model to a 17-species multisequence alignment. PhastCons also uses the 5-parameter general time reversible (GTR) model of sequence evolution with a scaling parameter for conserved sequence. With the parameters used for generating the 17-way Most Conserved track, 90% of the human coding bases in our analyzed genomic region reside within conserved regions. For the studies described here, we extracted all coding bases from the annotated conserved regions, leaving a conserved noncoding sequence partition of 104,918 human bases and a total alignment (including gaps) of 132,422 bases. A character state matrix (coding plus conserved noncoding) was created by adding

all manually edited protein-coding sequence alignments to the above conserved noncoding sequence alignment. We also generated another conserved sequence matrix for comparison purposes using Gblocks 0.91b, which uses a phylogenetically naive approach to identify sequence conservation (parameters: minimum 23 sequences for a conserved position, minimum 37 sequences for a flanking position, maximum 8 contiguous nonconserved positions, minimum initial block length of 10, minimum block length of 10, and half allowed gap positions) (Castresana 2000). The resulting matrix of conserved sequences contained 77,961 bases.

The extraction of specific bases and conversion of data files to FASTA, NEXUS, or PHYLIP formats for subsequent analyses were performed using custom perl scripts (available on request). Maximum parsimony tree searching was performed using PAUP* 4.0b10 (Swofford 2003). All trees were rooted using the fishes (*Fugu* and *Tetraodon*) as outgroup taxa, and a constraint for the monophyly of mammals was used. Maximum parsimony trees were generated using random addition replicates as well as bootstrapped with 1,000 replicates of 10 random addition subtree pruning regrafting runs. Neighbor-Joining trees were generated with 1,000 bootstrap replicates using maximum likelihood (ML) HKY85 distances. Incongruence length difference (ILD) or partition homogeneity tests were performed with 1,000 replicates of 10 Tree Bisection-Reconnection random addition tree searches with PAUP* (Bull et al. 1993; Cunningham 1997).

The monophyly of mammals was constrained for all tree searching, except when performing ILD tests, Shimodaira–Hasegawa tests (SH tests), or otherwise noted (Shimodaira and Hasegawa 1999). Bayesian phylogenetic analysis was performed with both partitioned likelihood models and single partition models using the MPI version of MrBayes v3.1.2 and GTR + I + Γ models, as suggested by MrModeltest or Modeltest (Posada and Crandall 1998; Nylander 2004). For the MrBayes analysis, model parameters were estimated from the data, and default priors were used. Metropolis-coupled Markov chain Monte Carlo chains were run for 500,000 or 1,000,000 generations, sampling every 100 generations and using 6 heated chains per each of 2 independent runs. Stationarity was confirmed by manual inspection for convergence of independent runs as well as topological and likelihood value stability. Majority rules consensus trees were generated from the final two-thirds of sampled trees using PAUP. Bootstrapped Bayesian runs were performed using seqboot from PHYLIP to create 100 bootstrap data sets, which were then independently analyzed with MrBayes using the settings described above (500,000 generations, sampling every 100 generations, with the final 6,668 sampled trees used for the consensus) (Felsenstein 2007). The RY-coded coding plus conserved non-coding sequence matrix was run to 5,000,000 generations, the average likelihood values increased until around 500,000 generations, but a single tree became completely resolved before 200,000 generations (data not shown).

ML tree searches were performed with the GTR + Γ model using RAxML-VI-HPC v2.1.3 (Stamatakis 2006). A proportion of invariable sites was not used because that parameter (I) is not implemented in RAxML-VI-HPC v2.1.3 (Stamatakis 2006). The best ML trees were obtained by performing greater than 20 independent tree searches from both completely random and random addition parsimony-based starting trees (default for RAxML) using the "−f d" high-performance hill-climbing algorithm. Highest likelihood trees from multiple runs of RAxML were the same as the trees obtained from multiple runs of PHYML using GTR + Γ + I models for several data sets tested (Guindon and Gascuel 2003). Even though its hill-climbing algorithm was slightly more likely to end at local maxima, we used RAxML because a single run with the conserved sequence partition took roughly one-third the time (~1.5 h) and one-tenth the RAM (~400 Mb) compared with PHYML; this allowed for more efficient use of available cluster resources. Bootstrap replicates were performed using the parallelized MPI-enabled version of RAxML-VI-HPC v2.1.3 with default settings. SH tests were performed with the best ML trees found using 15 random addition runs of RAxML using constraints for taxa in question. The best trees were used for SH tests with PAUP* and 10,000 RELL replicates with the GTR + Γ + I model (Shimodaira and Hasegawa 1999). All tree searching was performed using Linux clusters. Analyses with PAUP and RAxML were scripted and split using custom perl and shell scripts. Trees were visualized with the assistance of TreeGraph (Muller and Muller 2004).

Results and Discussion

Overview

We generated and compiled a high-quality comparative sequence data set consisting of sequences from 44 vertebrate species (table 1), all of which are orthologous to a 1.9-Mb region on human chromosome 7 (Thomas et al. 2003). This entire genomic region is syntenic in all mammals, reptiles, and fishes that we have examined to date (including some whose sequence was not analyzed in this study). Together, the consistent long-range synteny, Blast-based sequence comparisons, and nature of the cross-species BAC isolation and mapping process (see Thomas et al. [2002]) confirm the orthologous relationship of the sequences within the analyzed data set. The amount of assembled, annotated, and quality-trimmed sequence in the data set varies from 257 kb from *Fugu* to 2 Mb from elephant, with this variance reflecting both intrinsic differences in the size of the genomic region among species as well as incomplete sequence coverage for some species (see table 1).

Sequences were aligned using TBA (Blanchette et al. 2004). Because TBA produces local multisequence alignments, it handles small inversions or other local rearrangements well and avoids incorporating regions where the alignment uncertainty is high (Blanchette et al. 2004; Pollard et al. 2006). Protein-coding sequences were excised and manually edited to constrain coding indels to multiples of 3 bases, unless there was significant evidence for other indel sizes. We also removed any portions of the alignment where gap positions could not be easily determined. From this alignment, we made a coding sequence matrix, which contained 20,647 human-coding bases and 21,129 total characters; this matrix was then analyzed with ML, Bayesian, maximum parsimony, and Neighbor-Joining

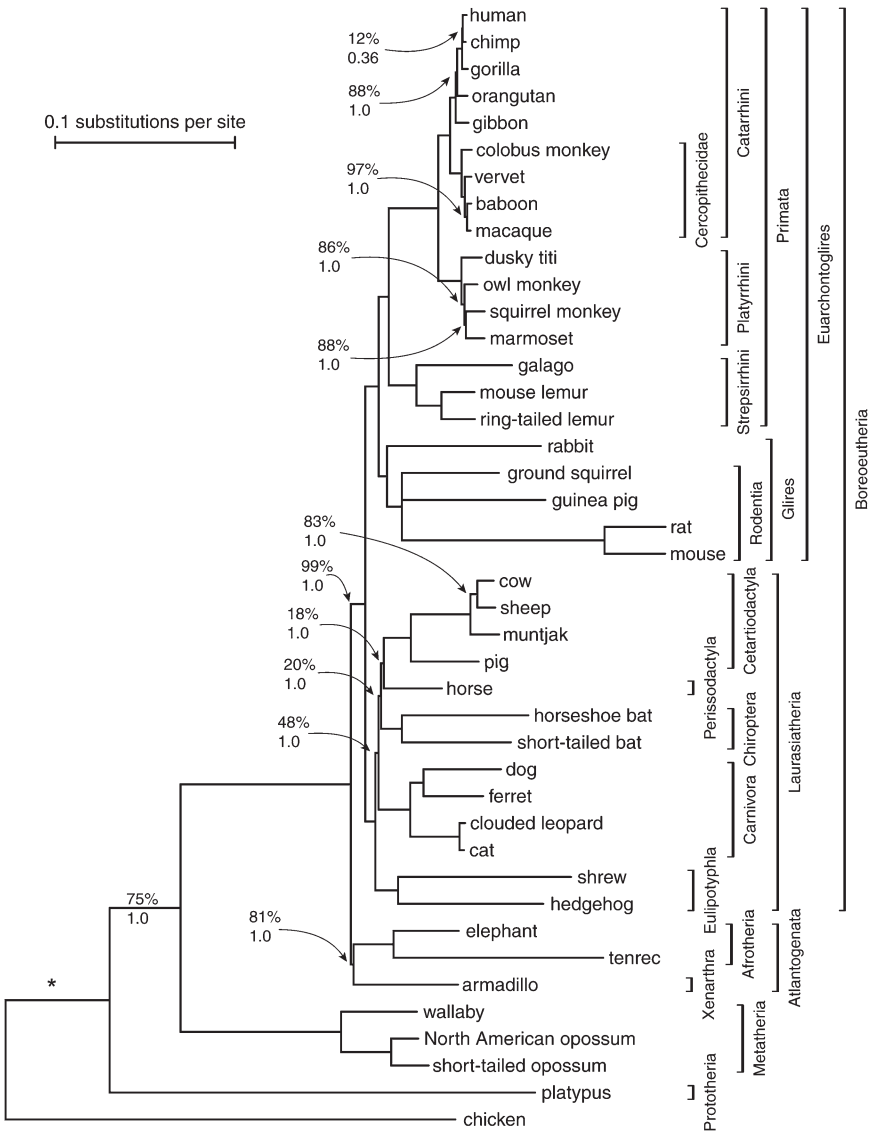

FIG. 1.—ML tree derived from the analysis of the coding sequence partition using RY-coded bases and a codon position partitioned CF + Γ model. Branch lengths indicate likelihood-inferred substitutions per site with a GTR + Γ model. ML bootstrap proportions are listed above Bayesian posterior probabilities for all branches at less than 100% bootstrap proportion and 1.0 Bayesian posterior probability support. Platypus was constrained to the mammals (its branch is marked with an asterisk to reflect this). The fishes (*Tetraodon* and *Fugu*) were used to root, but their branches are not shown. Branch lengths were optimized using ML from nucleotide-coded data with a GTR + Γ model.

approaches. Bayesian analysis yielded a highly resolved tree with posterior probabilities of 1.0 for all nodes (supplementary fig. 1, Supplementary Material online). This tree is fully congruent with the ML tree, and both trees are largely in agreement with other recent phylogenetic studies (Waddell, Cao, et al. 1999; Murphy, Eizirik, Johnson, et al. 2001; Murphy, Eizirik, O'Brien, et al. 2001; Waddell et al. 2001; Delsuc et al. 2002; Lin et al. 2002; Amrine-Madsen et al. 2003; Phillips and Penny 2003; Springer et al. 2003, 2004; Waddell and Shelley 2003; Kriegs et al. 2006; Nishihara et al. 2006; Hallstrom et al. 2007; Murphy et al. 2007), though a few nodes are different than those reported in some of these studies (see below and fig. 1).

The ML bootstrap proportions are significantly lower than Bayesian posterior probabilities. To investigate the relationship between the bootstrap proportions and the Bayesian posterior probabilities, we developed a "Bayesian bootstrap score" based on the majority rules consensus tree at stationarity for each bootstrapped data set. The resulting score was only slightly higher than the ML bootstrap proportions (see supplementary fig. 2, Supplementary Material online); these results support the notion that Bayesian posterior probabilities may be misleadingly high in some cases and are not necessarily comparable with traditional bootstrap measures of support (Waddell et al. 2001, 2002; Douady et al. 2003; Yang 2007). It is possible that more complex models and/or Bayesian techniques for combining data could lead to posterior probabilities that better reflect the uncertainties in the tree (Edwards et al. 2007; Liu and Pearl 2007). Some bootstrap replicates resulted in unlikely rearrangements of chicken and platypus (e.g., platypus outside the chicken and other mammals). We also saw this

when the coding plus conserved noncoding sequence matrix was nucleotide coded and analyzed with maximum parsimony. We believe that these findings are a consequence of the long branches leading to the monotreme and reptile species represented in this study, as well as the low taxon sampling of those groups and the distant fish outgroup. Because these unlikely arrangements affected bootstrap support values, we constrained the monophyly of mammals unless otherwise noted.

Although Bayesian posterior probabilities are high for all branches except the *Homo–Pan–Gorilla* group (human, chimpanzee, and gorilla), ML support is not sufficient to resolve some branches. In an attempt to rectify this and to take advantage of our notably large sequence data set, we investigated using conserved noncoding sequence for tree construction. Conserved bases were identified using phastCons (Siepel et al. 2005), which utilizes a phylogenetic hidden Markov model to distinguish conserved versus nonconserved sequence and a GTR model of substitution rates to identify conserved segments (other details are provided in Methods). With the parameters used, 90% of the known coding sequence and 5.5% of the presumed noncoding sequence within the region were identified as conserved. We generated a matrix containing 153,552 bases by combining the coding sequence alignment and the conserved noncoding sequence alignment to form a coding plus conserved noncoding sequence matrix; this matrix yielded highly resolved trees with strong branch support (fig. 2).

Although we found significant differences between the trees generated with the coding versus conserved noncoding sequence partitions based on ILD tests ($P = 0.001$), this was entirely due to the less-conserved third codon positions (Bull et al. 1993). When third codon positions were excluded (i.e., using partitions consisting of codon positions 1 and 2 vs. conserved noncoding sequences), there was no significant difference between the results obtained with each partition (ILD test $P = 0.432$; see table 2). RY-based coding of the data (discussed below) eliminated the significant differences between partitions, even when the third codon positions were included (ILD test $P = 0.328$; see table 2). Indeed, we only encountered differences between the trees generated with the 2 partitions on branches that were weakly supported by the coding sequence data set (figs. 1 and 2). Finally, we performed likelihood and Bayesian analysis with models partitioned by codon position (i.e., with the same model of evolution but allowing parameters of those models to vary between partitions); no significant differences in tree topology were seen, although minor differences in branch support scores were noted. Because phastCons bases its identification of conserved sequences on a phylogenetic tree, we also used Gblocks, a phylogenetically naive program for finding conserved regions. We found no differences in the ML tree generated from conserved regions identified by phastCons and Gblocks, with the associated bootstrap proportions similar in both cases (data not shown).

We also examined coding sequence for high-confidence indels, finding 24 phylogenetically informative indels (supplementary fig. 6, Supplementary Material online). A large number of the coding indels were shared by closely related species, such as rat and mouse (7 indels supporting), and separated the fish as our outgroup (5 indels supporting). Notably, we found 3 indels that are homoplastic on any of our trees, 2 of which (labeled "l" and "p" in supplementary fig. 6, Supplementary Material online) likely reflect multiple independent deletion events as they were detected in marsupials and only 1–3 species in Euarchontoglires. The third indel that appears homoplastic on our trees joins dusky titi to the apes (human, chimpanzee, gorilla, orangutan, and gibbon); this may be the result of lineage sorting or independent events. Although multiple deletion events may be relatively rare, the observed homoplasy suggests that caution should be used in interpreting support for taxa based on small numbers of such events.

Nonphylogenetic Signals

Large sequence data sets, such as the one analyzed here, offer the potential to resolve weakly supported branches; however, they can also be prone to detecting nonphylogenetic signals that confound the results (Philippe et al. 2005). We examined several potential sources of "systematic error" or "nonphylogenetic signals," attempting to exclude them or to control for their influence (Philippe et al. 2005). Specifically, we considered base-composition bias, incongruence across the genomic region, missing data, influence of the alignment guide tree, and long-branch attraction as possible sources of nonphylogenetic signal. We further examined long-branch attraction during the analysis of various individual taxa.

Base Composition

There are significant differences in base composition among species, ranging from 45% to 58% G + C in the coding sequence partition and 32% to 45% G + C in the conserved noncoding sequence partition. The chi-square test for homogeneity of base frequencies across taxa was highly significant ($P < 0.000001$ for all partitions examined, except with codon second positions $P = 0.00926835$), though the validity of this test is questionable because it does not take phylogenetic structure into account. To reduce the effects of nonphylogenetic signals due to base-composition differences among species, we coded the nucleotides as purines or pyrimidines (RY coding) (Phillips et al. 2001, 2004; Philippe et al. 2005). This approach also has the benefit of removing signals deriving from the more common transitions that may be associated with higher rates of saturation due to reversals. Indeed, we found that RY-coding eliminated significant differences in trees supported by the less-conserved codon third positions (table 2). Because of the large data set size, we maintained sufficient signal with RY-coded data to make robust phylogenetic inferences (figs. 1 and 2). We further found that RY-coding eliminates almost all base-composition differences among species. The coefficient of variation between purines and pyrimidines was 84% lower than the coefficient of variation between G + C and A + T for the coding sequence partition and 87% lower for the conserved noncoding sequence partition. RY-coding also eliminated significant differences in trees supported by codon third positions (ILD $P = 0.948$).

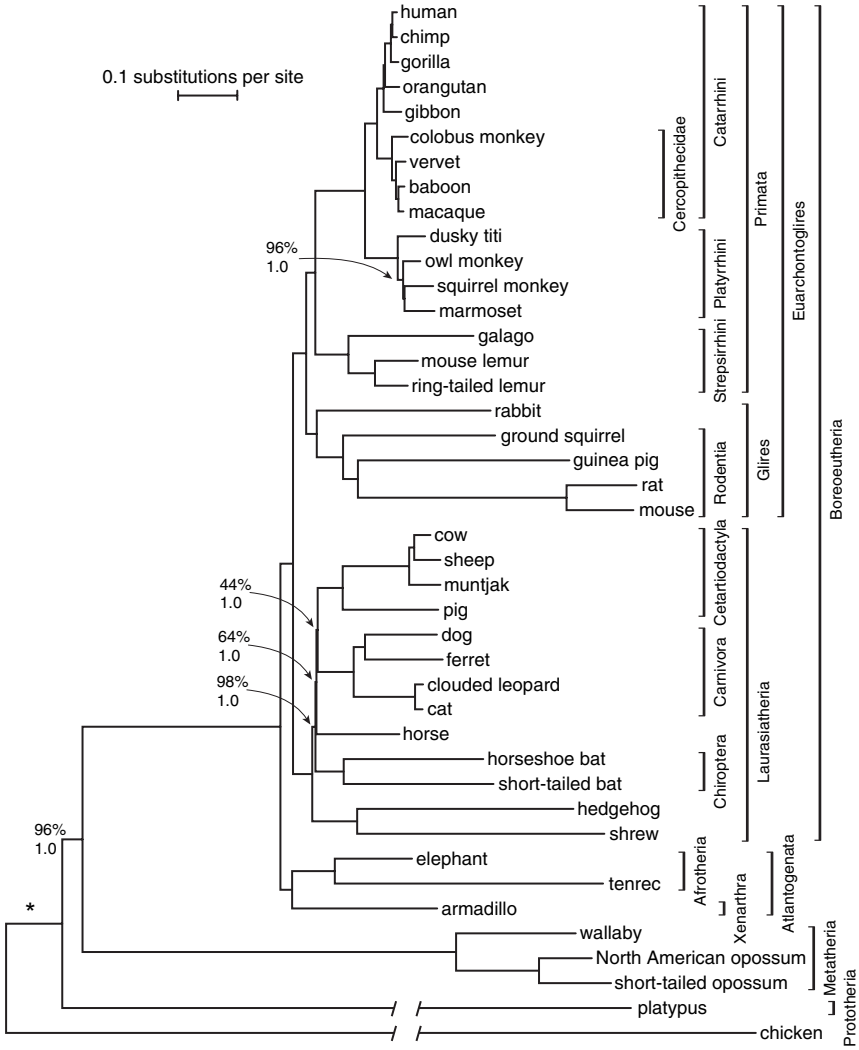

FIG. 2.—ML tree derived from the analysis of coding plus conserved noncoding sequence matrix using RY-coded bases. A CF + Γ model was used, with 4 partitions: 3 for codon positions and 1 for conserved noncoding sequence. Long branches leading to platypus and chicken were abbreviated for clarity. Other features are the same as indicated in figure 1.

Incongruence Across the Genomic Region

Combining phylogenetic data is thought to potentially be problematic (Bull et al. 1993). For example, recent genome-wide studies in yeast and bacteria encountered problems with combining large numbers of protein-coding sequence alignments (Comas et al. 2007; Edwards et al. 2007). To check for heterogeneity of support across the alignment, we split the coding plus conserved noncoding sequence matrix into 10 equal-sized segments and analyzed

Table 2
Pairwise ILD *P* Values[a]

Partition	Coding[b]	Coding Position 1[c]	Coding Position 2[c]	Coding Position 3[c]	Coding Positions 1 and 2[c]	Conserved Noncode[d]
Coding						<0.001
Coding pos1[c]			0.323	<0.001		0.460
Coding pos2[c]		0.648		0.983		0.324
Coding pos3[c]		0.569	0.960		0.002	<0.001
Coding pos1 and pos2				0.599		0.432
Conserved noncoding[d]	0.328	0.361	0.696	0.951	0.163	

[a] Values above diagonal are for NT-coded data, below are for RY-coded data.
[b] All protein-coding sequences.
[c] Codon position 1, 2, or 3 (as indicated) within coding sequence.
[d] Conserved noncoding sequence.

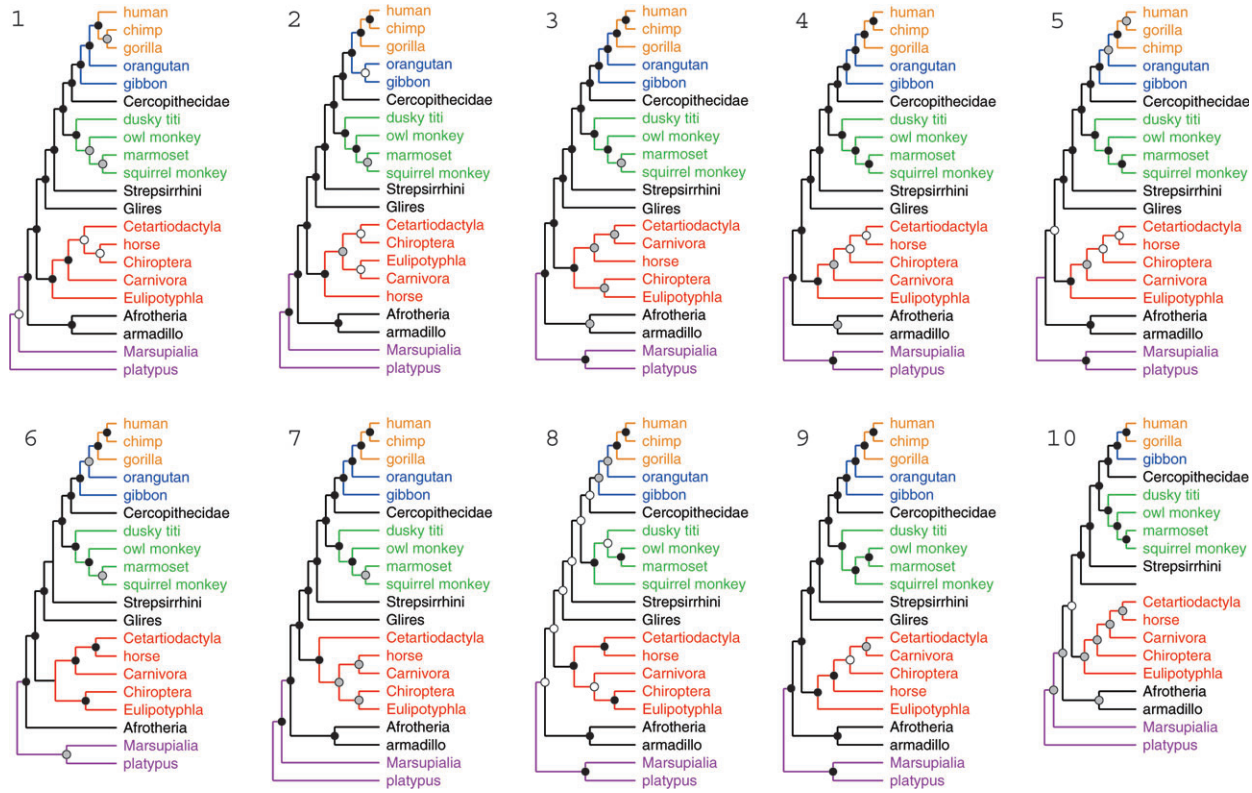

FIG. 3.—ML trees for each of 10 sequential, equal-sized partitions from the coding plus conserved noncoding sequence matrix. Numbers (1–10) reflect the specific partition used. The arrangement of taxa and branches indicated in colors other than black vary among partitions. Nodes annotated with hollow circles have less than 50% bootstrap proportions, those with shaded circles have 50% to 75% bootstrap proportions, and those with solid circles have 75% bootstrap proportions or greater. Branches that are the same in all trees are indicated in black, with some collapsed to higher level taxa for simplicity.

each with ML (fig. 3 and table 3). Although we found support for the Atlantogenata hypothesis with all 10 segments, there was considerable variation in the strength of that support, with some segments providing much larger shares of the overall support. One segment (6; see table 3 and fig. 3) did not contain any armadillo sequence (due to a gap in the BAC map) and thus could not provide support for the placental root. There was considerable heterogeneity of support among the segments for some other clades as well. Laurasiatheria was generally not well resolved in any of our analyses, and orders within Laurasiatheria did not have a consistent relationship among the different segments. The relationship between the marsupials and platypus varied as well, perhaps because of the long branches and poor taxon sampling of only one monotreme. Additionally, the relationships among owl monkey, marmoset, and squirrel monkey differed among partitions, though with weak support (discussed further below). A majority of segments supported Marsupionta, though overall support was strongly in favor of Theria (figs. 2 and 3).

Missing Data

Missing data have been considered a source of systematic error in phylogenetic analyses (Huelsenbeck 1991; Kearney 2002), although some simulation studies suggest

Table 3
Relative Likelihood Support for Placental Root across Coding Plus Conserved Noncoding Sequence Matrix

	Partition												
	1	2	3	4	5	6[a]	7	8	9	10	Total	Combined[b]	SH Test *P*
Atlantogenata	0	0	0	0	0	n/a	0	0	0	0	0	0	Best
Epitheria	4.5	1.7	2.9	9.1	3.9	n/a	1.6	16.7	7.5	8.0	56.0	69.4	<0.0001
Exafroplacentaila	4.5	1.8	0.5	7.9	3.8	n/a	1.6	16.9	6.6	6.7	50.3	62.9	<0.0001

NOTE.—n/a, not applicable.

[a] Partition 6 contains a region where there is incomplete armadillo sequence, so no placental root can be inferred.

[b] Likelihood score for entire coding plus conserved noncoding sequence matrix with 4 partitions for model parameters (3 for codon position and 1 for conserved noncoding).

that missing data should not be a problem for data sets with sufficient characters to provide robust signal (Rosenberg and Kumar 2003; Philippe et al. 2004, 2005; Wiens 2005). The coding sequence alignment had 11% missing data (including indels and sequencing gaps). To test if the missing data were biasing our analyses, we performed ILD tests on the RY-coded coding sequence partition, comparing alignment columns with gaps in 3 or fewer species and those with gaps in more than 3 species; no significant differences were seen ($P = 0.591$, supplementary table 1, Supplementary Material online). Similar results were obtained when the same analysis was performed with the RY-coded conserved noncoding sequence partition ($P = 0.627$). Finally, to see if additional missing data would strongly affect the analysis, we randomly deleted 25% of nongap bases from the coding plus conserved noncoding sequence matrix and observed no effect on the resulting ML tree, other than slightly changing the bootstrap support for a few branches (data not shown).

Alignment Guide Tree

We also analyzed a matrix consisting of all TBA-aligned sequence (containing 1,798,347 human bases). Because of computational constraints, we analyzed this data set only by maximum parsimony and ML methods. The trees derived from these analyses were almost completely resolved; however, by permuting the alignment guide tree, we were able to change the relative arrangement of those branches showing 100% bootstrap support in this analysis. Using only the conserved and protein-coding portions of the alignment yielded a tree with fewer well-resolved branches; however, branches with >70% bootstrap support were resistant to permutations of the alignment guide tree. Notably, the only branches with bootstrap proportions <70% were the interordinal branches within Laurasiatheria. These results confirm that difficult-to-resolve branches are more susceptible to biases introduced by aligning less-conserved sequences and that biases due to alignment guide trees can be largely controlled by only considering conserved sequences and strongly supported branches. To control for any possible effect of the alignment guide tree on our phylogenetic analysis, we permuted the alignment guide tree and reanalyzed the data for all controversial branches to confirm that the alignment guide tree was not biasing the results.

Individual Taxa

Euarchontoglires

The primate portion of the tree was strongly supported regardless of the partition or alignment guide tree used with the exception of the *Homo–Pan–Gorilla* group, which had insufficient informative changes in the RY-coded coding sequence partition to resolve (fig 1). However, when transitions are included, there is sufficient signal and 100% support for the sister relationship between chimpanzees and humans (supplementary fig. 1, Supplementary Material online). The major primate clades (including Catarrhini, Platyrrhini, and Strepsirrhini) were all supported at 100% by ML, Bayesian, Neighbor-Joining, and maximum parsimony approaches. These results largely agree with other recently reported molecular phylogenies for the primates (Barroso et al. 1997; Goodman et al. 1998; Poux and Douzery 2004; Ray et al. 2005; Opazo et al. 2006), with the exception of the relationships within Cebidae (marmoset, squirrel monkey, and owl monkey) (Barroso et al. 1997). Molecular systematic studies have disagreed on the arrangement of these 3 taxa. We find support for the association of squirrel monkey with marmoset. Although we see consistent support for this association regardless of alignment guide tree, RY- or nucleotide-coding, or tree inference algorithm, bootstrap proportions are relatively weak, and the support varies across the genomic region under study (fig. 3). The monophyly of both Glires and Euarchontoglires was strongly supported by Bayesian, ML, and maximum parsimony approaches, as in other recent studies (Thomas et al. 2003; Douzery and Huchon 2004); note that this is in sharp contrast to the results of Misawa and Janke (2003). Notably, Neighbor-Joining support for Glires and Euarchontoglires was also strong (bootstrap proportion 91%, supplementary fig. 5, Supplementary Material online) in contrast to the Neighbor-Joining results of Wildman et al. (2007), who used whole-genome shotgun sequence data; this is potentially explained by the much greater taxon sampling afforded by our data set. The placement of guinea pig securely within Rodentia is also strongly supported by our data (SH test $P < 0.0001$ excluding guinea pig as an outgroup to the other rodents), in agreement with others (Sullivan and Swofford 2004); this result holds when the alignment guide tree is permuted to place guinea pig outside the rodents.

Laurasiatheria

Within Laurasiatheria, there has been considerable disagreement about the arrangement and composition of the historical order Insectivora, although most recent molecular studies divide up this group among several orders, with tenrec falling in Afrotheria (Stanhope et al. 1998; Lin et al. 2002; Nishihara et al. 2006) and shrew and Erinaceous hedgehogs falling in Eulipotyphla (within Laurasiatheria) (Madsen et al. 2001; Murphy, Eizirik, Johnson, et al. 2001; Lin et al. 2002; Malia et al. 2002; Amrine-Madsen et al. 2003). Many studies of mitochondrial DNA have placed the Eulipotyphlans at the root of Placentalia, probably because of the unusually high AT content of their mitochondrial genomes; however, recent studies with greater taxon sampling and more complex models have also placed them in Laurasiatheria (Waddell, Cao, et al. 1999; Arnason et al. 2002; Lin et al. 2002; Gibson et al. 2005; Kjer and Honeycutt 2007). Our Neighbor-Joining tree has Eulipotyphla at the root of Placentalia, but this may be a consequence of the long-branch lengths of the 2 Eulipotyphlans (hedgehog and shrew) represented in our data set (fig. 2 and supplementary fig. 5, Supplementary Material online). Using ML, Bayesian, and maximum parsimony approaches, our analyses consistently place Eulipotyphla at the root of Laurasiatheria, as do most other recent studies (fig. 2 and supplementary figs. 3 and 5; Supplementary Material online) (Murphy, Eizirik, O'Brien, et al. 2001; Waddell et al.

2001; Arnason et al. 2002; Scally et al. 2002; Amrine-Madsen et al. 2003; Waddell and Shelley 2003; Nishihara et al. 2006; Nikolaev et al. 2007).

The placement of Perissodactyla, represented here by the horse, has been another source of controversy. Usually, this group is placed either sister to Cetartiodactyla (Murphy, Eizirik, Johnson, et al. 2001; Lin et al. 2002) or sister to Carnivora (Murphy, Eizirik, O'Brien, et al. 2001; Arnason and Janke 2002; Amrine-Madsen et al. 2003), in most cases with weak support (Waddell, Cao, et al. 1999). Schwartz et al. (2003) found a single transposon insertion supporting the Perissodactyla–Carnivora association; meanwhile, Nishihara et al. (2006) found a single transposon insertion supporting a Perissodactyla–Carnivora association and 5 insertions supporting a Perissodactyla–Chiroptera–Carnivora (Pegasoferae, Nishihara et al. 2006) association (i.e., excluding the traditional Perissodactyla–Cetartiodactyla association of hoofed mammals that we see with this analyses). Of note, Nishihara et al. (2006) did also find one transposon insertion that conflicted with the Pegasoferae hypothesis. Even with the large number of characters analyzed here, we only found weak bootstrap support for the placement of Perissodactyla (fig. 2), although it tended to associate closest with the Cetartiodactylans and secondarily with the Carnivores. Across the region, support for the arrangement of orders varied significantly, with only segments 4, 5, and 10 agreeing and none of the segments agreeing with the ML tree for the entire matrix (fig. 3). Thus, the arrangement of orders within Laurasiatheria appears to be difficult to resolve even with large amounts of sequence data and reasonably large numbers of species represented. We further found that the relative arrangement of Laurasiatherian orders was highly sensitive to alignment guide tree artifacts, though not in a predictable way. Using the coding plus conserved noncoding sequence matrix, we performed SH tests with the 5 most supported Laurasiatherian trees from the literature; none could be excluded with high confidence, and this likely is due, in part, to the short branches separating Laurasiatherian orders. Perhaps with increased taxon sampling, this problem will be more tractable. It may be that a strong nonphylogenetic signal or incomplete lineage sorting is obscuring the interordinal relationships within Laurasiatheria. Methods that treat gene trees and species trees simultaneously, such as that described by Liu and Pearl (2007), might also be able to better resolve such regions.

Theria

Although considered a mammal, the phylogenetic placement of monotremes has long been controversial. The hierarchical placement of monotremes as an outgroup of the other mammals has been challenged by molecular and morphological studies that placed Monotremata as a sister group to the marsupials in a clade called Marsupionta (Janke et al. 1996, 1997). Our results, however, agree with recent molecular studies that yielded significant evidence (including coding indels) in support of the monophyly of Theria (placental mammals and marsupials), with the monotremes as the first branch of the mammalian tree (Killian et al. 2001; Phillips and Penny 2003; van Rheede et al. 2006). We find strong bootstrap support (≥94%, see figs. 1 and 2) for Theria by Neighbor-Joining, maximum parsimony, Bayesian, and ML approaches, and SH tests with nucleotide-coded data just reach significance ($P = 0.0251$) in excluding Marsupionta with the coding plus conserved noncoding sequence matrix. However, there is significant heterogeneity of support for Theria across the region, and a majority of the segments (6/10) support Marsupionta (fig. 3). These findings are consistent with other recent molecular and morphological analyses that supported the monophyly of Theria (Hu et al. 1997; Phillips and Penny 2003; van Rheede et al. 2006) but illustrate the difficulty of determining the relationships between these clades.

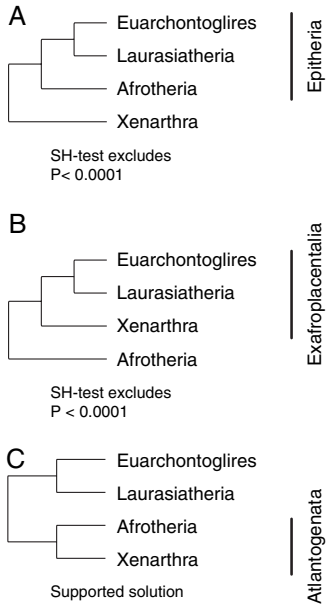

Fig. 4.—Three possible roots for Placentalia. SH test results from the coding plus conserved noncoding sequence matrix for both nucleotide- and RY-coded matrices. (*A*) Hypothesis rooting Placentalia between Xenarthra and Epitheria (Boreoeutheria + Afrotheria). (*B*) Hypothesis rooting Placentalia between Afrotheria and Exafroplacentalia (Boreoeutheria + Afrotheria). (*C*) Hypothesis rooting Placentalia between Boreoeutheria and Atlantogenata (Afrotheria + Xenarthra).

Placentalia

Although the nucleotide-coded protein-coding sequence partition failed to resolve the root of Placentalia (ML bootstrap <60%, supplementary fig. 1, Supplementary Material online), the RY-coded coding sequence partition supports an Atlantogenatan root (fig. 1). Adding the conserved noncoding partition provides high statistical support for Atlantogenata, both with nucleotide- and RY-coded data (figs. 2 and 4). Bootstrap support of 100% is seen with all model-based approaches used (including ML, Bayesian, Neighbor-Joining, and minimum evolution supplementary figs. 4 and 5). SH tests using the coding plus conserved noncoding sequence matrix exclude Epitheria and Exafroplacentalia, with the results significant past the limits of PAUP and CONSEL ($P < 0.0001$) (fig. 4). Because of the limited number of Afrotherian and Atlantogenatan species in this study, some caution is warranted in

interpreting these results. Maximum parsimony analysis of the nucleotide-coded coding sequence partition supports an Epitherian root (e.g., fig. 4*A*), but when codon third position sites are removed or the sequence is RY-coded, bootstrap support is reduced to <50%. Maximum parsimony analysis of the coding plus conserved noncoding sequence partition also weakly supports Epitheria (supplementary fig. 4).

To exclude the influence of biases introduced by the alignment guide tree, we realigned the sequences using a guide tree based on the highest likelihood tree constrained to each possible root, then repeated the likelihood analysis. In all cases, the ML tree derived from the coding plus conserved noncoding sequence matrix was rooted between Atlantogenata and Boreoeutheria with 100% bootstrap support and highly significant SH test results ($P < 0.001$). Because tenrec has a significantly longer branch length with these data than elephant or armadillo, we tried individually removing tenrec and elephant. With either tenrec or elephant missing, we still saw ≥99% ML bootstrap support for Atlantogenata. We also separated the coding plus conserved noncoding sequence matrix into 10 equally sized partitions and analyzed each separately (fig. 3). Although the likelihood values for Atlantogenata varied significantly among the partitions, all partitions supported an Atlantogenatan root (table 3).

These results agree with some other recent large-scale analyses on mostly independent data sets (e.g., Hallstrom et al. [2007]; Murphy et al. [2007]; and Wildman et al. [2007]) but conflict with the findings of Nikolaev et al. (2007), Nishihara et al. (2007), and Kriegs et al. (2006). Kriegs et al. (2006) found 2 transposon insertions in Afrotheria and Boreoeutheria that were not found in armadillo or sloth. These transposon sequences are quite old, and the flanking sequence is not well conserved. Thus, the transposon-associated sequences may have mutated out of recognition in the 30 Myr from the placental root to the divergence of armadillo and sloth; alternatively, multiple transposon insertions may have occurred in Afrotheria and Boreoeutheria (Springer et al. 2003; Kriegs et al. 2006; Murphy et al. 2007). Homoplasy for transposon insertions due to targeted insertions or lineage sorting on short branches, though presumably rare, has been reported (Pecon-Slattery et al. 2004; van de Lagemaat et al. 2005; Yu and Zhang 2005; Nishihara et al. 2006) and could also explain these results.

Nikolaev et al. (2007) analyzed amino acid and conserved noncoding genomic sequences from 14 species to examine the root of Placentalia. Using ML analyses of conserved noncoding sequence from the ENCODE pilot project regions (http://www.genome.gov/10005107), they exclude the Epitheria hypothesis and, separately, use amino acid sequences derived from the same regions to exclude the Atlantogenata hypothesis. Notably, analyses using their largest data set (conserved noncoding sequence) failed to differentiate between rooting Placentalia at Atlantogenata or Exafroplacentalia. Additionally, their limited taxon and outgroup sampling argues for caution in interpreting the final results (Delsuc et al. 2002); for example, when we only analyzed data from the 14 species studied by Nikolaev et al. (2007), we still found support for Atlantogenata as the root, although the bootstrap support was weak. The data used in our analysis contain significantly more taxa, both ingroup and outgroup, than other recent large-scale nucleotide-based analyses, and this may affect the results significantly.

Summary

In summary, we used a comparative sequence data set that contains a remarkably large number of conserved bases to derive a phylogeny that provides additional evidence to resolve some of the controversial branches in the mammalian lineage. We find significant support for an Atlantogenatan root of Placentalia, as well as additional evidence for the monophyly of Theria. Our studies highlight the difficulties in resolving some very short mammalian branches (e.g., interordinal relationships within Laurasiatheria), even with large amounts of data. Our work further illustrates the value of large genomic sequence data sets for improving the resolution of phylogenetic trees, in this case, to clarify some of the remaining ambiguities within the mammalian tree. Sequences from an increasing number of mammalian taxa should help to resolve the remaining ambiguities associated with the short branches within and between the placental orders.

Supplementary Material

Supplementary figures 1–6 and table 1 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank Adam Siepel and Webb Miller for helpful comments about this manuscript. We also thank members of the NISC Comparative Sequencing Program (particularly B. Blakesley, G. Bouffard, J. McDowell, B. Maskeri, M. Park, N. Hanson, and P. Thomas) for providing leadership in the generation of the comparative sequence data analyzed here. We also thank Associate Editor Scott Edwards and 2 anonymous reviewers for unusually thorough and helpful comments that ultimately resulted in a significantly improved manuscript. This research was supported in part by the Intramural Research Program of the National Human Genome Research Institute, National Institutes of Health.

Literature Cited

Amrine-Madsen 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.

Aparicio S, Chapman J, Stupka E, et al. (41 co-authors). 2002. Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. Science. 297:1301–1310.

Arnason U, Adegoke JA, Bodin K, Born EW, Esa YB, Gullberg A, Nilsson M, Short RV, Xu X, Janke A. 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.

Barroso CML, Schneider H, Schneider MPC, Sampaio I, Harada ML, Czelusniak J, Goodman M. 1997. Update on the phylogenetic systematics of new world monkeys: further DNA evidence for placing the pygmy marmoset (*Cebuella*) within the genus Callithrix. Int J Primatol. 18:651–674.

Bashir A, Ye C, Price AL, Bafna V. 2005. Orthologous repeats and mammalian phylogenetic inference. Genome Res. 15:998–1006.

Blakesley RW, Hansen NF, Mullikin JC, et al. (22 co-authors). 2004. An intermediate grade of finished genomic sequence suitable for comparative analyses. Genome Res. 14:2235–2244.

Blanchette M, Kent WJ, Riemer C, et al. (12 co-authors). 2004. Aligning multiple genomic sequences with the threaded blockset aligner. Genome Res. 14:708–715.

Boissinot S, Entezam A, Young L, Munson PJ, Furano AV. 2004. The insertional history of an active family of L1 retrotransposons in humans. Genome Res. 14:1221–1231.

Bull JJ, Huelsenbeck JP, Cunningham CW, Swofford DL, Waddell PJ. 1993. Partitioning and combining data in phylogenetic analysis. Syst Biol. 42:384–397.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol. 17:540–552.

Chaisson MJ, Raphael BJ, Pevzner PA. 2006. Microinversions in mammalian evolution. Proc Natl Acad Sci USA. 103: 19824–19829.

Chen JM, Stenson PD, Cooper DN, Ferec C. 2005. A systematic analysis of LINE-1 endonuclease-dependent retrotranspositional events causing human genetic disease. Hum Genet. 117:411–427.

Clamp M, Cuff J, Searle SM, Barton GJ. 2004. The Jalview Java alignment editor. Bioinformatics. 20:426–427.

Comas I, Moya A, Gonzalez-Candelas F. 2007. From phylogenetics to phylogenomics: the evolutionary relationships of insect endosymbiotic gamma-proteobacteria as a test case. Syst Biol. 56:1–16.

Cope ED. 1889. The Edentata of North America. Am Nat. 23:657–664.

Cunningham CW. 1997. Can three incongruence tests predict when data should be combined? Mol Biol Evol. 14:733–740.

Delsuc F, Scally M, Madsen O, Stanhope MJ, de Jong WW, Catzeflis FM, Springer MS, Douzery EJ. 2002. Molecular phylogeny of living xenarthrans and the impact of character and taxon sampling on the placental tree rooting. Mol Biol Evol. 19:1656–1671.

Douady CJ, Delsuc F, Boucher Y, Doolittle WF, Douzery EJ. 2003. Comparison of Bayesian and maximum likelihood bootstrap measures of phylogenetic reliability. Mol Biol Evol. 20:248–254.

Douzery EJ, Huchon D. 2004. Rabbits, if anything, are likely Glires. Mol Phylogenet Evol. 33:922–935.

Edwards SV, Liu L, Pearl DK. 2007. High-resolution species trees without concatenation. Proc Natl Acad Sci USA. 104:5936–5941.

Felsenstein J. 2007. PHYLIP (phylogeny inference package). Distributed by the author. Seattle (WA): Department of Genome Sciences, University of Washington.

Gibson A, Gowri-Shankar V, Higgs PG, Rattray M. 2005. A comprehensive analysis of mammalian mitochondrial genome base composition and improved phylogenetic methods. Mol Biol Evol. 22:251–264.

Goodman M, Porter CA, Czelusniak J, Page SL, Schneider H, Shoshani J, Gunnell G, Groves CP. 1998. Toward a phylogenetic classification of primates based on DNA evidence complemented by fossil evidence. Mol Phylogenet Evol. 9:585–598.

Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol. 52:696–704.

Hallstrom BM, Kullberg M, Nilsson MA, Janke A. 2007. Phylogenomic data analyses provide evidence that Xenarthra and Afrotheria are sister groups. Mol Biol Evol. 24: 2059–2068.

Hillis DM, Pollock DD, McGuire JA, Zwickl DJ. 2003. Is sparse taxon sampling a problem for phylogenetic inference? Syst Biol. 52:124–126.

Hu Y, Wang Y, Luo Z, Li C. 1997. A new symmetrodont mammal from China and its implications for mammalian evolution. Nature. 390:137–142.

Huelsenbeck JP. 1991. When are fossils better than extant taxa in phylogenetic analysis? Syst Zool. 40:458–469.

International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. Nature. 409:860–921.

Janke A, Gemmell NJ, Feldmaier-Fuchs G, von Haeseler A, Paabo S. 1996. The mitochondrial genome of a monotreme–the platypus (*Ornithorhynchus anatinus*). J Mol Evol. 42: 153–159.

Janke A, Xu X, Arnason U. 1997. The complete mitochondrial genome of the wallaroo (*Macropus robustus*) and the phylogenetic relationship among Monotremata, Marsupialia and Eutheria. Proc Natl Acad Sci USA. 94:1276–1281.

Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, Kent WJ. 2004. The UCSC table browser data retrieval tool. Nucleic Acids Res. 32:D493–D496.

Kearney M. 2002. Fragmentary taxa, missing data, and ambiguity: mistaken assumptions and conclusions. Syst Biol. 51:369–381.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. 2002. The human genome browser at UCSC. Genome Res. 12:996–1006.

Killian JK, Buckley TR, Stewart N, Munday BL, Jirtle RL. 2001. Marsupials and Eutherians reunited: genetic evidence for the Theria hypothesis of mammalian evolution. Mamm Genome. 12:513–517.

Kjer KM, Honeycutt RL. 2007. Site specific rates of mitochondrial genomes and the phylogeny of eutheria. BMC Evol Biol. 7:8.

Kluge AG, Wolf AJ. 1993. Cladistics: what's in a Name. Cladistics. 9:183–199.

Kriegs JO, Churakov G, Kiefmann M, Jordan U, Brosius J, Schmitz J. 2006. Retroposed elements as archives for the evolutionary history of placental mammals. PLoS Biol. 4:e91.

Lin YH, McLenachan PA, Gore AR, Phillips MJ, Ota R, Hendy MD, Penny D. 2002. Four new mitochondrial genomes and the increased stability of evolutionary trees of mammals from improved taxon sampling. Mol Biol Evol. 19:2060–2070.

Liu L, Pearl DK. 2007. Species trees from gene trees: reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. Syst Biol. 56:504–514.

Madsen O, Scally M, Douady CJ, Kao DJ, DeBry RW, Adkins R, Amrine HM, Stanhope MJ, de Jong WW, Springer MS. 2001. Parallel adaptive radiations in two major clades of placental mammals. Nature. 409:610–614.

Malia MJ Jr, Adkins RM, Allard MW. 2002. Molecular support for Afrotheria and the polyphyly of Lipotyphla based on analyses of the growth hormone receptor gene. Mol Phylogenet Evol. 24:91–101.

Margulies EH, Chen CW, Green ED. 2006. Differences between pair-wise and multi-sequence alignment methods affect vertebrate genome comparisons. Trends Genet. 22:187–193.
Misawa K, Janke A. 2003. Revisiting the Glires concept—phylogenetic analysis of nuclear sequences. Mol Phylogenet Evol. 28:320–327.
Muller J, Muller K. 2004. TREEGRAPH: automated drawing of complex tree figures using an extensible tree description format. Molecular Ecology Notes. 4:786–788.
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. 294:2348–2351.
Murphy WJ, Pevzner PA, O'Brien SJ. 2004. Mammalian phylogenomics comes of age. Trends Genet. 20:631–639.
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. 17:413–421.
Nikolaev S, Montoya-Burgos JI, Margulies EH, Rougemont J, Nyffeler B, Antonarakis SE. 2007. Early history of mammals is elucidated with the ENCODE multiple species sequencing data. PLoS Genet. 3:e2.
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.
Nishihara H, Okada N, Hasegawa M. 2007. Rooting the eutherian tree: the power and pitfalls of phylogenomics. Genome Biol. 8:R199.
Nishihara H, Satta Y, Nikaido M, Thewissen JG, Stanhope MJ, Okada N. 2005. A retroposon analysis of Afrotherian phylogeny. Mol Biol Evol. 22:1823–1833.
Nylander JAA. 2004. MrModeltest. Program distributed by the author. Uppsala (Sweden): Evolutionary Biology Centre, Uppsala University.
Opazo JC, Wildman DE, Prychitko T, Johnson RM, Goodman M. 2006. Phylogenetic relationships and divergence times among New World monkeys (Platyrrhini, Primates). Mol Phylogenet Evol. 40:274–280.
Pecon-Slattery J, Pearks Wilkerson AJ, Murphy WJ, O'Brien SJ. 2004. Phylogenetic assessment of introns and SINEs within the Y chromosome using the cat family felidae as a species tree. Mol Biol Evol. 21:2299–2309.
Philippe H, Delsuc F, Brinkmann H, Lartillot N. 2005. Phylogenomics. Annu Rev Ecol Evol. 36:541–562.
Philippe H, Snell EA, Bapteste E, Lopez P, Holland PW, Casane D. 2004. Phylogenomics of eukaryotes: impact of missing data on large alignments. Mol Biol Evol. 21: 1740–1752.
Phillips MJ, Delsuc F, Penny D. 2004. Genome-scale phylogeny and the detection of systematic biases. Mol Biol Evol. 21:1455–1458.
Phillips MJ, Lin YH, Harrison GL, Penny D. 2001. Mitochondrial genomes of a bandicoot and a brushtail possum confirm the monophyly of australidelphian marsupials. Proc Biol Sci. 268:1533–1538.
Phillips MJ, Penny D. 2003. The root of the mammalian tree inferred from whole mitochondrial genomes. Mol Phylogenet Evol. 28:171–185.
Pollard DA, Moses AM, Iyer VN, Eisen MB. 2006. Detecting the limits of regulatory element conservation and divergence estimation using pairwise and multiple alignments. BMC Bioinformatics. 7:376.
Posada D, Crandall KA. 1998. MODELTEST: testing the model of DNA substitution. Bioinformatics. 14:817–818.
Poux C, Douzery EJ. 2004. Primate phylogeny, evolutionary rate variations, and divergence times: a contribution from the nuclear gene IRBP. Am J Phys Anthropol. 124:1–16.
Ray DA, Xing J, Hedges DJ, et al. (13 co-authors). 2005. Alu insertion loci and platyrrhine primate phylogeny. Mol Phylogenet Evol. 35:117–126.
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.
Rosenberg MS, Kumar S. 2003. Taxon sampling, bioinformatics, and phylogenomics. Syst Biol. 52:119–124.
Scally M, Madsen O, Douady CJ, de Jong WW, Stanhope MJ, Springer MS. 2002. Molecular evidence for the major clades of placental mammals. J Mammal Evol. 8:239–277.
Schwartz S, Elnitski L, Li M, Weirauch M, Riemer C, Smit A, Green ED, Hardison RC, Miller W. 2003. MultiPipMaker and supporting tools: alignments and analysis of multiple genomic DNA sequences. Nucleic Acids Res. 31:3518–3524.
Shimamura M, Yasue H, Ohshima K, Abe H, Kato H, Kishiro T, Goto M, Munechika I, Okada N. 1997. Molecular evidence from retroposons that whales form a clade within even-toed ungulates. Nature. 388:666–670.
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.
Siepel A, Bejerano G, Pedersen JS, et al. (16 co-authors). 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 15:1034–1050.
Springer MS, de Jong WW. 2001. Phylogenetics. Which mammalian supertree to bark up? Science. 291:1709–1711.
Springer MS, Murphy WJ, Eizirik E, O'Brien SJ. 2003. Placental mammal diversification and the cretaceous-tertiary boundary. Proc Natl Acad Sci USA. 100:1056–1061.
Springer MS, Stanhope MJ, Madsen O, de Jong WW. 2004. Molecules consolidate the placental mammal tree. Trends Ecol Evol. 19:430–438.
Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics. 22:2688–2690.
Stanhope MJ, Waddell VG, Madsen O, de Jong W, Hedges SB, Cleven GC, Kao D, Springer MS. 1998. Molecular evidence for multiple origins of Insectivora and for a new order of endemic African insectivore mammals. Proc Natl Acad Sci USA. 95:9967–9972.
Sullivan J, Swofford DL. 2004. Are guinea pigs rodents? The importance of adequate models in molecular phylogenetics. J Mammal Evol. 4:77–86.
Swofford DL. 2003. PAUP*: Phylogenetic analyisis using parsimony (*and other methods). Sunderland (MA): Sinauer Associates.
Thomas JW, Prasad AB, Summers TJ, Lee-Lin SQ, Maduro VV, Idol JR, Ryan JF, Thomas PJ, McDowell JC, Green ED. 2002. Parallel construction of orthologous sequence-ready clone contig maps in multiple species. Genome Res. 12: 1277–1285.
Thomas JW, Touchman JW, Blakesley RW, et al. (71 co-authors). 2003. Comparative analyses of multi-species sequences from targeted genomic regions. Nature. 424: 788–793.
van de Lagemaat LN, Gagnier L, Medstrand P, Mager DL. 2005. Genomic deletions and precise removal of transposable elements mediated by short identical DNA segments in primates. Genome Res. 15:1243–1249.

van Rheede T, Bastiaans T, Boone DN, Hedges SB, de Jong WW, Madsen O. 2006. The platypus is in its place: nuclear genes and indels confirm the sister group relation of monotremes and Therians. Mol Biol Evol. 23:587–597.

Waddell PJ, Cao Y, Hauf J, Hasegawa M. 1999. Using novel phylogenetic methods to evaluate mammalian mtDNA, including amino acid-invariant sites-LogDet plus site stripping, to detect internal conflicts in the data, with special reference to the positions of hedgehog, armadillo, and elephant. Syst Biol. 48:31–53.

Waddell PJ, Kishino H, Ota R. 2001. A phylogenetic foundation for comparative mammalian genomics. Genome Inform. 12:141–154.

Waddell PJ, Kishino H, Ota R. 2002. Very fast algorithms for evaluating the stability of ML and Bayesian phylogenetic trees from sequence data. Genome Inform. 13:82–92.

Waddell PJ, Okada N, Hasegawa M. 1999. Towards resolving the interordinal relationships of placental mammals. Syst Biol. 48:1–5.

Waddell PJ, Shelley S. 2003. Evaluating placental inter-ordinal phylogenies with novel sequences including RAG1, gamma-fibrinogen, ND6, and mt-tRNA, plus MCMC-driven nucleotide, amino acid, and codon models. Mol Phylogenet Evol. 28:197–224.

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.

Wiens J. 2005. Can incomplete taxa rescue phylogenetic analyses from long-branch attraction? Syst Biol. 54:731–742.

Wildman DE, Uddin M, Opazo JC, Liu G, Lefort V, Guindon S, Gascuel O, Grossman LI, Romero R, Goodman M. 2007. Genomics, biogeography, and the diversification of placental mammals. Proc Natl Acad Sci USA. 104: 14395–14400.

Yang Z. 2007. Fair-balance paradox, star-tree paradox, and Bayesian phylogenetics. Mol Biol Evol. 24:1639–1655.

Yu L, Zhang YP. 2005. Evolutionary implications of multiple SINE insertions in an intronic region from diverse mammals. Mamm Genome. 16:651–660.

Scott Edwards, Associate Editor

Accepted April 7, 2008