Genomics, biogeography, and the diversification of placental mammals

Derek E. Wildman, Monica Uddin, Juan C. Opazo, Guozhen Liu, Vincent Lefort, Stephane Guindon, Olivier Gascuel, Lawrence I. Grossman, Roberto Romero, and Morris Goodman

Phylogenetic analyses can elucidate the history of diversification within a group of organisms such as placental mammals (i.e., Placentalia) (1, 2). However, if the analyses use too few characters or taxa, an inaccurate phylogenetic tree can be obtained because of sampling error (3). Here we seek to reduce such error by using abundant character information from mammalian and other vertebrate genomes that have been completely or nearly completely sequenced. On examining phylogenetically the coding nucleotide sequences and translated amino acid sequences for thousands of genes, we attempt to resolve key uncertainties about the ancient branching order of crown placental mammals. Our results are promising, but uncertainties remain concerning the basal diversification of Placentalia. The practice of phylogenomics is still in its infancy and has yet to produce an authoritative model that infallibly predicts all of the real patterns of nucleotide substitutions within evolved genomes.

Among the placental mammals, phylogenetic branching events have been inferred by using either nucleotide sequence data (4–11) or rare insertion/deletion patterns (12–15). Many of the results recognize four primary eutherian groups: Afrotheria, Xenarthra, Laurasiatheria, and Euarchontoglires. Afrotherians (e.g., elephants, hyraxes, manatees, aardvarks, tenrecs, and allies) are a clade of mammals that originated in Africa, and whose extant members still mostly remain on that continent with the exception of Asian elephants and sirens such as the Florida manatee. The Xenarthra includes the sloths, armadillos, and anteaters that today are restricted to South and Central America (although some Xenarthra, such as the nine-banded armadillo, have recently dispersed to North America). The Laurasiatheria (e.g., bats, eulipotyphlans, pangolins, carnivores, perissodactyls, and cetartiodactyls) is a diverse clade including extant lineages that originated in the ancient northern continent of placental mammals. (A) Afrotheria as sister taxon to the other placental mammalian clades. (B) Xenarthra as sister taxon to the other placental mammalian clades. (C) Xenarthra and Afrotheria group together to the exclusion of the other clades. (D) Murid rodents as sister taxon to the other placental mammalian clades; Glires is disrupted by the joining of Lagomorpha to the remaining placental clades.


The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

Abbreviations: MP, maximum parsimony; ML, maximum likelihood; NJ, neighbor-joining.

†To whom correspondence may be addressed. E-mail: prbchiefstaff@med.wayne.edu, dwildman@med.wayne.edu, or mgoodwayne@aol.com.

This article contains supporting information online at www.pnas.org/cgi/content/full/0704342104DC1.

© 2007 by The National Academy of Sciences of the USA
of Laurasia. The Euarchontoglires includes the species from five living mammalian orders (e.g., primates, treeshrews, flying lemurs, rabbits, and rodents). This last group remains the most controversial, and a number of recent studies have suggested it is not valid (4, 6, 7).

The studies that have identified and supported these primary groupings have been able to resolve the branching pattern within some of the clades, but a statistically robust determination of the branching order at the base of the placental clade continues to be elusive. Three primary hypotheses (Fig. 1A–C) supporting the branching orders among the four major placental groups described above have been proposed. In the first scenario (9, 10, 16), Afrotherians split from the other three clades (Notolegia/H11005 Exafroplacentalia) at the base of the placental tree. In the second hypothesis (13), the Xenarthrans split from the other three clades (Epitheria). The third hypothesis (5, 11, 14, 17) proposes that Xenarthra and Afrotheria group together (Atlantogenata) to the exclusion of Laurasiatheria and Euarchontoglires (Boreoeutheria).

Some studies have suggested a fourth hypothesis, that one of the groups, Euarchontoglires, is polyphyletic, and that Glires (rodents and lagomorphs) is also not monophyletic (Fig. 1D). Based on a parsimony analysis of nuclear-encoded genes, it was reported (8) that murid rodents as represented by mouse and rat were the sister group to all other placental mammals, and that rabbits were the next branching clade. More recently, a number of large-scale genomic studies (4, 6) have reported that primates grouped more closely to laurasiatherians than to Glires.

To test the four alternative hypotheses (Fig. 1), we constructed phylogenetic trees and performed topology tests based on a data set that included nucleotide and amino acid sequence data from the complete genome sequences of 11 mammalian species representing the major placental clades depicted in Fig. 1, a marsupial opossum, and two nonmammalian outgroups. The ingroup taxa include three primates (human, chimpanzee, and rhesus macaque), two rodents (mouse and rat), rabbit, dog, cow, armadillo, African elephant, tenrec, and the South American Gray Short-tailed opossum. Chicken and frog were used as outgroups to root the tree. To avoid common pitfalls associated with phylogenomic studies, we focused our analyses on conserved protein coding genes to reduce saturation and long-branch attraction effects. Our alignment method also preserved the codon reading frame for all loci and all taxa, so we were able to remove potentially saturated third codon positions from the parsimony analyses. Because all tree reconstruction methods fail to recover the “true” tree given certain conditions, we used four different inferential tree reconstruction methods: maximum parsimony (MP), maximum likelihood (ML), Bayesian, and...
neighboring (NJ). The probabilistic methods (ML and Bayesian) may be more robust than the others to model assumption violation, although NJ can also often recover the true tree when the wrong model of sequence evolution is used. MP minimizes the amount of evolution (i.e., nucleotide substitutions) and in some cases, outperforms the probabilistic methods (18, 19). Although we would be encouraged if several of these tree reconstruction methods converged on just one of the four hypotheses, we realize that all methods might converge on an incorrect branching pattern for the basal diversification of Placentalia.

Results

When all human RefSeq mRNA transcripts (n = 25,556) were aligned to their putative orthologs, the result was a gapped alignment of 36 Mb. However, orthologs for most genes could not be found for all 14 taxa. Therefore, the analyses included only those loci for which our method could assign orthologous sequences for all 14 taxa. This reduced data set consisted of a multiple sequence alignment of 1,698 protein-coding loci with an alignment length of 1,443,825 bp, including insertions and deletions. The mean composition of nucleotide bases in this alignment was as follows: T = 22.8%, C = 23.8%, A = 27.8%, and G = 25.6%. A table showing the nucleotide base composition at each codon position for each taxon as well as the number of nucleotides analyzed is included in supporting information (SI) Table 3. Notably, for each base at each of the three codon positions, there are at most only very small compositional differences among the 14 taxa. A summary spreadsheet of the putative orthologs used in the main analyses (SI Table 4), as well as the accompanying alignment files are available as SI.

Fig. 2 depicts the optimal phylogenetic branching pattern among the taxa whether a Bayesian, ML, or MP phylogenetic tree reconstruction method is used with the coding nucleotide sequence data or the parsimony amino acid data. The branch lengths in Fig. 2 were obtained by ML analysis. Fig. 3 depicts the optimal NJ trees. The MP nucleotide tree has a length of 267,158 steps, and the ML scores for the tree topology are -lnL 7,238,023.807 (PAUP*) and -lnL 7,240,210.130 (PhyML). The branch support values for each of the clades are 100% (MP and PhyML bootstrap percent) or 1.0 (Bayesian posterior probability) at all nodes in the tree with the single exception of the Atlantogenata clade, which had a parsimony bootstrap value of 95% in the nucleotide analysis. The relatively short branch lengths seen in the catarrhine and, more specifically, the ape clade also provide further evidence for the hominid slowdown hypothesis (20, 21).

The NJ tree confirms the presence of the Atlantogenata clade with bootstrap support of 100%, although it differs from the other three methods by not supporting Euarchontoglires or Glires as monophyletic, indicating rodents as the first branching placental clade and depicting Laurasiatheria as the sister group to primates (Fig. 3). Parsimony and NJ results from the translated amino acid sequences both depict a basal split between Atlantogenata and Boreoeutheria. Whereas the parsimony amino acid and nucleotide sequence data results show an identical topology, the NJ amino acid tree depicts a boreoeutherian clade not detected by using nucleotide sequence data and indicates Lagomorpha is sister to a clade that includes the remaining Boreoeutheria (Fig. 3). Notably, with nucleotide sequences, when only first and second codon positions (positions less likely to be saturated by superimposed mutations) are retained, the NJ tree again depicts the basal Placentalia divergence to be between Atlantogenata and Boreoeutheria.

We conducted parsimony and likelihood topology tests on the phylogenetic trees constructed by using the coding nucleotide sequence data (Tables 1 and 2). The likelihood tests clearly indicate the tree depicted in Figs. 1C and 2 is the optimal tree (P < 0.0001; Table 1). The trees depicted in Fig. 1A, B, and D are rejected as being suboptimal. Results of parsimony tests also reject the topologies depicted in Fig. 1A and D (Table 2); however, the topology in Fig. 1B cannot be rejected (P = 0.0897; Templeton test). The parsimony scores from these two topologies differ by only 73 steps.

Table 1. Likelihood tests of alternative topologies

<table>
<thead>
<tr>
<th>Tree</th>
<th>-ln L</th>
<th>Diff -ln L</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>7,239,065.38</td>
<td>1,041.57228</td>
<td>0.000*</td>
</tr>
<tr>
<td>B</td>
<td>7,239,235.261</td>
<td>1,211.45404</td>
<td>0.000*</td>
</tr>
<tr>
<td>C</td>
<td>7,238,023.807</td>
<td>Best</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>7,243,526.067</td>
<td>5,502.25965</td>
<td>0.000*</td>
</tr>
</tbody>
</table>

Shimodaira-Hasegawa test using RELL bootstrap (one-tailed test).
Number of bootstrap replicates, 1,000.
* P < 0.05.
Long-branch attraction is a classic phylogenetic problem that incorrectly unites long branches together in a clade (28). It is a particular problem in MP, which fails to correct for parallel changes on long branches (25, 26). Long-branch attraction can cause systematic error in all methods used in this study, but there are data that suggest adding more taxa can break up long branches, reducing the probability of error (29). Interestingly, an NJ analysis of the Murphy et al. (9) data set composed of 44 mammalian taxa detected, as in the original study, an Afrotherian clade as sister to all other placental mammals (SI Text), rather than the rodent first separation detected by the nucleotide NJ analysis (all codon positions) performed in this study. The tree topology also shows a monophyletic Glires and Euarchontoglires. To so analyze the Murphy data set by NJ demonstrates long-branch attraction effect gets drastically reduced by denser taxon sampling (which breaks up the rodent long branch). Indeed, if the NJ analysis of the Murphy data set includes only the taxa represented in this study, a “rodent first” topology is once again recovered. The present study sampled more taxa than other recent mammalian phylogenomic studies (4, 6) and is therefore less likely to be affected by the long-branch attraction problem. This point is borne out by the finding that our data set recovered identical topologies to those reported by Hultley et al. (6) and Cannarozzi et al. (4) when we limited our data to include only the taxa in those studies (see SI Text). Furthermore, the parsimony nucleotide results were obtained by using only first and second codon positions, which are less subject to the parallel changes that contribute to the long-branch attraction problem, and the topology obtained was identical to that recovered when MP was applied to the translated amino acid sequences (which also reduces the likelihood of erroneously identifying homoplasies as synapomorphies).

Variation in substitution rate at a single base or amino acid position over evolutionary time is referred to as heterotachy and can result in phylogenetic artifacts (30, 31). Errors resulting from heterotachy are difficult to detect with the methods we used; however, it has been shown that MP methods are sometimes less sensitive to heterotachy than are the probabilistic ML and Bayesian techniques (19), and in this study, results were congruent among these three methods. Nevertheless, further investigation of the cladistic relationships among mammals is now possible, because more eutherian genomes are available (e.g., platypus, cat, horse, bat, galago, tree shrew, and guinea pig). The availability of these genomes will allow further testing of the Atlantogenata/Boreoeutheria split and, within Boreoeutheria, the Glires hypothesis depicted in the majority of analyses in this study.

In our view, the weight of evidence now points to a sister group relationship between Atlantogenata and Boreoeutheria, and a clear scenario of biogeographic diversification emerges (Fig. 4). In this scenario, the placental mammals would have been subdivided into two lineages when the spreading Tethyan seaway widely separated Gondwana in the south from Laurasia in the north during the Cretaceous (32, 33). This process divided the initial members of the clades Boreoeutheria in the north from their southern atlantogenatan counterparts. Also, later in the Cretaceous, the disconnection of the African and South Amer-
and Euarchontoglires remains murky, and it is unclear whether this was primarily because of vicariance between North America and Eurasia, some other vicariant event, or dispersal. Some remarkably similar morphological features that have emerged among the mammalian clades in the different geographic areas led previous workers to group divergent taxa together in polyphyletic assemblages (e.g., ungulates) based on convergent evolution of hoofs or to assume that features emerged only once (e.g., the variant types of gross anatomy in the placenta). With the availability of the mammalian genome sequences ever accumulating, it is now possible to design and test phylogenetic hypotheses about the genetic underpinnings of these and other important aspects of mammalian phenotypes.

Materials and Methods
Details of analysis parameters and settings are available in SI Text.

Data Composition. Orthology extraction and multiple sequence alignment were performed by using Online Codon-Preserved Alignment Tool (OCPAT), an in-house-developed tool that combines the BLAST (34) and Clustal (35) algorithms and preserves the correct protein coding reading frame in all taxa aligned. The tool is available at http://homopan.wayne.edu/Pise/ocpat.html. Details of the analysis pipeline are provided at http://homopan.wayne.edu/ocpat/index.html (54). Taxa included in the study were Homo sapiens (human) (36), Pan troglodytes (chimpanzee) (37), Macaca mulatta (Rhesus monkey) (38), Mus musculus (mouse) (39), Rattus norvegicus (rat) (40), Orctolagus cuniculus (rabbit), Canis familiaris (dog) (41), Bos taurus (cow), Dasyus novemcinctus (armadillo), Loxodonta africana (African elephant), Echinops telfairi (tenrec), Monodelphis domestica (gray short-tailed opossum) (42), Gallus gallus (chicken) (43), and Xenopus tropicalis (Western clawed frog). For the analysis, all data we used were updated at the end of August 2006. Ornithorhynchus anatinus (platypus) and other recently completed draft genomes were not included. Nucleotide composition was calculated by using MEGA 4.0 (44).

Phylogenetic Analyses. ML analysis. ML analyses were conducted in PAUP* Ver. 4.0b10 (45) by using Model settings determined by the program ModelTest Ver. 3.7 (46), as chosen by Akaike Information Criterion (AIC). These settings correspond to the general time reversible (GTR) + I + model with four rate categories. Assumed nucleotide frequencies were: A = 0.27630, C = 0.24200, G = 0.25440, and T = 0.22730. The assumed proportion of invariable sites = 0.3082 and the distribution of rates at variable sites = gamma (discrete approximation) with a shape parameter (\(\alpha\)) = 0.6954. Our heuristic analysis included 10 random sequence additions and tree-bisection-reconnection (TBR) tree searching. Complete settings are available in SI Text. Additional analyses were conducted in a new beta version (3.0) of PhyML (47), which includes fast subtree pruning and regrafting (SPR) tree search (48). We first ran PhyML with an SPR search from 10 random starting trees and with GTR + I model with four rate categories but without bootstrap. The proportion of invariant sites, shape parameter of the gamma distribution and GTR parameters were estimated from the data. Then, we performed a bootstrap analysis with 100 replicates by using BioNJ trees as starting points, SPR tree search, and the model parameter values estimated in the first run. All inferred trees (i.e., 10 with random starting trees and 100 with resampled data) were identical.

Bayesian analysis. The parallel MrBayes [Ver. 3.1.2 (49–51)] analysis was carried out by using the resources of the Computational Biology Service Unit from Cornell University (Ithaca, NY).

With 14 taxa and a concatenated sequence of 1,443,825 bases, the parallel processes ran for \(\approx\)1 week for 1 million generations
of Markov-Chain Monte Carlo. We chose to assume partition homogeneity rather than using mixed models, which potentially would suffer from overparameterization from the 1,600+ loci included in the analysis. For the 1.4-Mb data set, we also ran the analysis using BayesPhylogenies (52). In this case, we assumed a single GTR pattern.

**MP analysis.** We conducted a heuristic search in PAUP* Ver. 4.0b10 consisting of 100 random addition sequence replicates using the tree-bisection-reconnection (TBR) branch-swapping algorithm. We excluded third codon positions in parsimony analysis because of potential saturation at those sites. Thus, 481,275 characters were excluded, leaving a data set that included the remaining 962,550 characters. In addition to searching for the optimal tree, we conducted a bootstrap analysis that was based on 1,000 pseudoreplicates with 10 random addition sequence replicates per pseudoreplicate. The TBR algorithm was also used in this analysis. We also conducted an MP analysis on the translated amino acid sequences. In addition to the full complement of taxa, we ran parsimony analyses to reflect the taxon sampling in refs. 4 and 6.

**NJ analysis.** We conducted an NJ search in PAUP* Ver. 4.0b10 with 1,000 full heuristic bootstrap replicates using the ML distances as selected by ModelTest and in MEGA 4.0 using the ML composite distance. We also conducted an NJ bootstrap analysis (1,000 replicates) in MEGA on the translated amino acid sequences using the Jones–Taylor–Thornton distance. In addition to the full complement of taxa, we ran NJ analyses to reflect the taxon sampling in refs. 4 and 6.

Andrew Meade (University of Reading, Reading, U.K.) provided access to computing resources, and part of this work was carried out by using the resources of the Computational Biology Service Unit from Cornell University, which is partially funded by Microsoft Corporation. Munirul Islam (Wayne State University, Detroit, MI) provided computational support. We thank Wen-Hsiung Li (University of Chicago, Chicago, IL), Wilfried W. de Jong (University of Nijmegen, Nijmegen, The Netherlands), Nathaniel M. Pearson (Wayne State University), two anonymous reviewers, and the PNAS editorial board for comments and suggestions that greatly improved the manuscript. This research was supported in part by the Intramural Research division of National Institute of Child Health and Human Development, National Institutes of Health, Department of Health and Human Services and National Science Foundation Grant BCS-0550209. We acknowledge the sources of the unpublished genomic data, including: Baylor College of Medicine Human Genome Sequencing Center (cow; http://www.hgsc.bcm.tmc.edu/); the Broad Institute (rabbit, elephant, tenrec, and armadillo); and the U.S. Department of Energy Joint Genome Institute (frog).