Early History of Mammals Is Elucidated with the ENCODE Multiple Species Sequencing Data

Sergey Nikolaev[1*], Juan I. Montoya-Burgos[2], Elliott H. Margulies[3], NISC Comparative Sequencing Program[3,4], Jacques Rougemont[5], Bruno Nyffeler[5], Stylianos E. Antonarakis[1*]

1 Department of Genetic Medicine and Development, University of Geneva, Geneva, Switzerland, **2** Department of Zoology and Animal Biology, University of Geneva, Geneva, Switzerland, **3** Genome Technology Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland, United States of America, **4** Intramural Sequencing Center, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland, United States of America, **5** Swiss Institute of Bioinformatics, Lausanne, Switzerland

Understanding the early evolution of placental mammals is one of the most challenging issues in mammalian phylogeny. Here, we addressed this question by using the sequence data of the ENCODE consortium, which include 1% of mammalian genomes in 18 species belonging to all main mammalian lineages. Phylogenetic reconstructions based on an unprecedented amount of coding sequences taken from 218 genes resulted in a highly supported tree placing the root of Placentalia between Afrotheria and Exafroplacentalia (Afrotheria hypothesis). This topology was validated by the phylogenetic analysis of a new class of genomic phylogenetic markers, the conserved noncoding sequences. Applying the tests of alternative topologies on the coding sequence dataset resulted in the rejection of the Atlantogenata hypothesis (Xenarthra grouping with Afrotheria), while this test rejected the second alternative scenario, the Epitheria hypothesis (Xenarthra at the base), when using the noncoding sequence dataset. Thus, the two datasets support the Afrotheria hypothesis; however, none can reject both of the remaining topological alternatives.

Citation: Nikolaev S, Montoya-Burgos JI, Margulies EH, NISC Comparative Sequencing Program, Rougemont J, et al. (2007) Early history of mammals is elucidated with the ENCODE multiple species sequencing data. PLoS Genet 3(1): e2. doi:10.1371/journal.pgen.0030002

Introduction

The relationships among mammalian lineages have recently been revised using multiple gene phylogenetic analyses leading to the recognition of four main placental clades: Euarchontoglires, Laurasiatheria, Xenarthra, and Afrotheria. Humans and other Primates and Dermoptera, Scandentia, and Glires (Rodentia and Lagomorpha) form the Euarchontoglires [1,2]. The Laurasiatheria comprise the Cetartiodactyla, Perissodactyla, Carnivora, Chiroptera, Eulipotyphla, and Pholidota [3]. The Euarchontoglires and Laurasiatheria are sister groups, and together they form the Boreoeutheria [2]. The American lineage Xenarthra groups anteaters, armadillos, and sloths [4]. The remaining placental lineage is called Afrotheria due to the African origin of its stem members and includes elephants, Sirenia, Hyracoidea, Tubulidentata, Macroscelida, Tenrecidae, and Chrysochloridae [2]. It is still unclear, however, how Afrotheria, Xenarthra, and Boreoeutheria are interrelated, though most molecular studies favor Afrotheria as the basal placental lineage (e.g., [2,5,6]).

The three possible scenarios for the position of the root of placental mammals are: (1) Afrotheria at the base of Placentalia (Xenarthra grouping with Boreoeutheria in the Exafroplacentalia); this hypothesis emerged from molecular phylogenetic studies [2,7]. Due to a limited amount of available comparative information this topology was not significantly supported, and Shimodaira-Hasegawa tests did not reject alternative topologies [8]. (2) Epitherian hypothesis, where Xenarthra are at the base of placental mammals and Afrotheria are grouped with other placental mammals. This hypothesis is based on morphological features, uniting Afrotheria and Boreoeutheria, such as developed penis and absence of vaginal longitudinal divisions [9]. Recently, molecular characters consisting of two LINE insertions were found to be shared by Boreoeutheria and Afrotheria and absent in Xenarthra [10]. (3) Atlantogenata hypothesis, based on the analysis of vertebrate mtDNA protein sequences favors the grouping of Afrotheria and Xenarthra [11].

Understanding the early evolution of placental mammals remains a major challenge not only for evolutionary biology but also for genomic, developmental, and biomedical research [12]. Resolving the placental root is an important question that has not been unambiguously answered even by the analysis of numerous genes and a wide taxonomic sampling [2,13]. The reason may lie in the limited amount of phylogenetic information tracing back to early placental divergences that may have been compressed in time [14] as suggested by the fossil record [15]. Adding a substantial amount of sequence data may help in resolving the order of these closely spaced cladogenetic events.

Editor: Andrew G. Clark, Cornell University, United States of America

Received July 17, 2006; **Accepted** November 20, 2006; **Published** January 5, 2007

Abbreviations: aa, amino acid; bp, base pair; BP, bootstrap proportion; CDS, coding sequence; CNC, conserved noncoding sequence; G, gamma; GTR, general time-reversible; I, proportion of invariable sites; JP, jackknife proportion; Kaa, kilo–amino acids; kb, kilobase; ML, maximum likelihood

* To whom correspondence should be addressed. E-mail: Sergey.Nikolaev@medecine.unige.ch (SN), Stylianos.Antonarakis@medecine.unige.ch (SEA)

Author Summary

Application of molecular phylogenetic methods drastically changed the conception of relationships within mammals. Recent molecular phylogenetic studies have shown that living placental mammals belong to one of the three subgroups: Boreoeutheria, Afrotheria, or Xenarthra, but the relations between these are still unknown. In a previous analysis using 16 genes, Boreoeutheria and Xenarthra grouped together. However, a study based on LINE insertions supported the grouping of Boreoeutheria and Afrotheria. To resolve this discrepancy, we applied sequence data from 1% of a genome in a subset of 18 mammalian species. We used concatenated coding sequence data from 218 genes encompassing 205 kilobases of DNA sequence. Phylogenetic analyses have shown Afrotheria as a basal group of Placentalia with high statistical support. To further validate these results, we analyzed a new phylogenetic marker: conserved noncoding sequence alignments (430 kilobases), which resulted in the same position of the placental root. Topological tests rejected the possibility of Afrotheria-Xenarthra grouping with the coding sequence dataset and Boreoeutheria-Afrotheria grouping with the noncoding sequence dataset. Ascertaining the relationships between mammals is of great importance for the investigation of evolutionary behavior of the different functional genomic elements.

To investigate the early evolution of placental mammals, we analyzed the ENCODE [16] dataset of coding sequences (CDSs) comprising 218 orthologous genes taken from 18 mammalian species representing all main lineages (Table S1). In addition, we have investigated the phylogenetic properties of evolutionarily constrained DNA sequences that do not overlap known CDSs. Such conserved noncoding sequences (CNCs) have been established by genomic scale comparisons between human and mouse [17–19]. CNCs are mostly not repetitive and cover approximately 3% of mammalian genomes, which is roughly two times more than CDSs [19–23]. CNCs seem to be under an equal or even stronger purifying selection than CDSs [21,23]. These qualities could make CNCs a potentially powerful class of phylogenetic markers to solve ancient cladogenetic events.

Results/Discussion

To settle the debate about the position of the root of placental mammals, we used the ENCODE consortium sequencing data, covering 1% of human genome and orthologous genomic regions in other mammals [23,24]. We created two independent alignments: the first from CDSs, the second from CNCs. Both alignments were prepared in a way in which in every column there are positions for at least one representative of each major mammalian group (Primates, Glires, Laurasiatheria, Xenarthra, Afrotheria, Marsupialia, and Monotremata). The final alignment of concatenated CDSs contained 204,786 base pairs (bp) belonging to 218 genes while the concatenated CNC alignment included 429,675 bp coming from all individual CNCs longer than 50 bp (the CNC alignment is more than twice as large as the CDS alignment). With this unprecedented amount of sequence information (at least 39 times larger than previous studies) we assessed mammalian phylogeny using *Monodelphis* (Metatheria) and *Platypus* (Monotremata) as an out-group.

The two independent datasets, CDS and CNC, were analyzed using maximum likelihood (ML) methods as implemented in PHYML and a general time-reversible (GTR) + gamma (G) + proportion of invariable sites (I) model of sequence evolution as defined by ModelTest [25,26]. Both datasets converged to the same highly supported topology (Figure 1). All major mammalian lineages known so far are reconstructed: Primates, Glires, Eurarchontoglires, Laurasiatheria, Boreoeutheria (B), Xenarthra (X), and Afrotheria (A). The bootstrap support for those nodes is 100% on trees from both phylogenetic markers. Moreover, our phylogenetic analyses provide a clear topological solution to the long-standing question of the position of mammalian root: placental mammals are split onto Afrotheria on one side and Exafroplacentalia on the other side with a high statistical support (CDS amino acids [aa]: 95% bootstrap proportion [BP]; CDS nucleotides: 88% BP; CNC: 73% BP). The phylogeny based on CNCs fully corroborates the CDS-based phylogeny. The consistent results of the two non-intersecting datasets provide additional support for the settlement of the debate regarding the position of the placental root.

The two remaining alternative topologies of the position of the root of placentals: (1) the Epitheria hypothesis (X [A,B]), and (2) the Atlantogenata hypothesis ([A,X] B), were confronted to the ML topology (Figure 1) using approximately unbiased, Kishino-Hasegawa, and Shimodaira-Hasegawa topological tests as implemented in the CONSEL package [27]. The results of ML analysis with GTR + G + I model performed with baseml program implemented in PAML package [28] (Table 1) using the CDS__DNA dataset show that none of the alternative topologies can be significantly rejected. Using the CDS__AA dataset the Atlantogenata theory can be rejected ($p < 0.00001$) while the Epitheria theory cannot (Table 1). Only using the CNC dataset, we can reject the Epitheria theory ($p < 0.01$) while the two other topologies are almost equally possible. The combined dataset from CNC and CDS__DNA 1 and 2 (first and second positions from CDS__DNA) alignments places Afrotheria at the base and allows us to reject only the Epitheria but not the Atlantogenata hypotheses.

Taken together, these results suggest that the root of the placental lineage lies between the Afrotheria and the group formed by all other Placentalia. Thus, it is only by the combined use of the CNC and CDS datasets (taken from 1% of mammalian genome) that a reasonable amount of evidence supporting the root of the tree is provided. We conclude that CNCs are powerful phylogenetic markers that can be complementary to CDS markers in phylogenetic reconstructions.

By concatenating the CNC and CDS__DNA 1 and 2 datasets, we obtain a DNA matrix that is approximately four times larger than when using CDS__DNA 1 and 2 alignments alone (592,027 bp, comprised of one fourth of CDS data and three fourths of CNC data). The bootstrap support for the placental root between Afrotheria and Exafroplacentalia using the combined data provides a confidence of 98% of BP (the individual BP supports were 95% and 73% using CDS__DNA 1 and 2 and CNC data alone, respectively). Thus, the inclusion of the CNC data has a significant impact in determining the most likely topology of Mammalia.

In order to test if a partitioning of our concatenated dataset might affect our results, we calculated the ML scores for the three alternative topologies using PAML and partitioning CNC versus CDS__DNA 1 and 2 (Table 1). We found that the

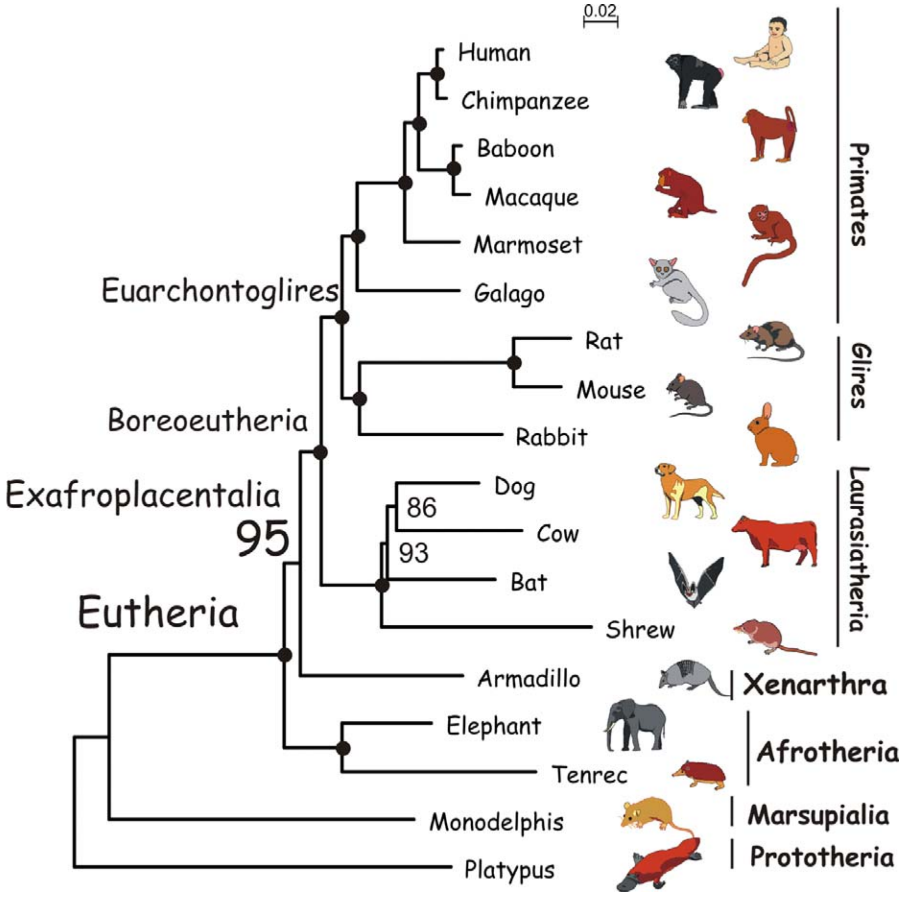

Figure 1. ML Phylogenetic Tree of Mammals

The topology and branch lengths shown here are based on concatenated alignments of coding exons (68,262 codons). Branch lengths are scaled to the number of substitutions per site. Bootstrap support values are indicated at each node, with solid-color dots indicating 100% support.

doi:10.1371/journal.pgen.0030002.g001

topology with Afrotheria at the base is the best; and using CONSEL, the Epitheria hypothesis is rejected while the Atlantogenata hypothesis is not significantly rejected.

Third codon positions of CDSs are known to saturate over evolutionary time, possibly at the placental evolutionary scale. To test this, we first applied a codon model (GTR + G) that assigns different values to all parameters for first, second, and third codon positions, as implemented in baseml. We also tested the exclusion of the third position from the analysis (CDS__DNA 1 and 2) (Table 1). Both analyses increased the robustness of our results. In all analyses, the Afrotheria clade remained at the base of Placentalia (with 95% BP support if excluding third position); the Atlantogenata hypothesis was rejected with $p < 0.01$ in both cases (codon model or excluding third position). In both analyses, topological tests did not reject the Epitheria hypothesis, yet with the CDS__DNA 1 and 2 data, the p-value is close to the significance threshold of 5% (see Table 1).

Problems with base compositional differences are important to address in studies where taxonomic sampling is sparse rather than dense. In the concatenated dataset CNC + CDS__DNA 1 and 2, the homogeneity chi-square test for base composition rejected base stationarity. Therefore, we assessed the impact of base composition by using baseml to perform likelihood estimates of the three competing topologies under a non-stationarity model (nhomo = 3 option) with TN93 + G + I. The best ML score was obtained for the topology with Afrotheria at the base (Table 1), suggesting that our results are not sensitive to non-stationarity of base composition.

Distal out-groups may influence the branching order of the basal in-group lineages. One way of exploring the potential impact of the out-group sampling on the rooting of Placentalia is to delete either *Monodelphys* or *Platypus.* Deleting *Monodelphys* favors the topology with the Afrotheria at the base for all datasets, while deleting *Platypus* favors the Epitheria hypothesis (for CDS-derived datasets) or the Atlantogenata hypothesis (CNC dataset). We further tested for the potential impact of the long branch of *Tenrec* on the rooting of the Placentalia by its deletion. All three alternative possibilities were found depending on the three datasets. There is, therefore, no clear evidence that a long-branch attraction artifact favors the topology with Afrotheria at the base.

To test if our phylogenetic searches were sensitive to the starting tree, we repeated the analysis for CDS__AA, CDS__DNA, and CNC using different starting trees (Afrotheria, Epitheria, and Atlantogenata). In all cases the tree with Afrotheria at the base was retrieved. To further test if PHYML NNI could result in the best topology, we performed PHYML SPR analysis for three datasets, CDS__DNA,

Table 1. Topological Tests of Alternative Hypotheses of Placental Root

Matrix	Afrotheria	Epitheria	Atlantogenata
CDS_AA	Best	0.079/ 0.080/ 0.097 (+)	0.00003/ 0/ 0 (–)
CDS_DNA (stationarity GTR + G + I)	Best	0.133/ 0.119/ 0.186 (+)	0.102/ 0.083/ 0.128 (+)
CDS_DNA (non-stationarity TN + G + I)	Best	0.137/ 0.122/ 0.180 (+)	0.004/ 0.001/ 0.01 (–)
CDS_DNA 1 and 2 positions	Best	0.061/ 0.053/ 0.064 (+)	0.0001/ 0.001/ 0.01 (–)
CNC	Best	0.001/ 0.008/ 0.013 (–)	Best
CNC + CDS_DNA 1 and 2 positions (two partitions)	Best	0.020/ 0.024/ 0.041 (–)	0.283/ 0.254/ 0.371 (+)

Numbers in cells correspond to *p*-values of the approximately unbiased, Kishino-Hasegawa, and Shimodaira-Hasegawa tests. (+), topology is not rejected; (–), topology is rejected.
doi:10.1371/journal.pgen.0030002.t001

CDS__AA, and CNC. All three resulted in the trees with Afrotheria at the base with support of 92%, 95%, and 65%, respectively.

Our analysis demonstrated that mammalian CNCs contain abundant signal for phylogenetic studies. In order to assess the phylogenetic signal as compared to CDS, we used two approaches: (1) jacknife analysis, (Figure 2); and (2) likelihood mapping (Figure S1). With jacknife analysis, the relative amount of phylogenetic signal is measured in CNC and CDS datasets (both DNA and AA) by systematically reducing the length of the initial alignment and measuring jacknife supports. We generated a CNC alignment of equivalent length to that of CDS__DNA alignment (205 kilobases [kb]) and for both sets we generated four gradually reduced datasets, comprising 100 kb, 50 kb, 20 kb, and 10 kb (for CDS__AA the length was calculated for the corresponding DNA alignment). We observed that for almost all nodes of the tree, even 5% of initial alignment comprising 10 kb (3,300 aa) is sufficient to assess highly supported phylogenetic relationship with jacknife proportion (JP) of near 100%. Among the less stable nodes of the tree is that of Exafroplacentalia. In the CNC and CDS datasets, the JP support of the Exafroplacentalia declines drastically with the reduction of the length of the alignments (Figure 2). Only with around a 100-kb sequence alignment (for both coding and noncoding sequence) is it possible to reconstruct the basal Exafroplacentalia group with support between 60% and 90% JP. This result explains why the majority of previous studies addressing the question of placental root were unable to give a conclusive answer, since these studies were conducted using a maximum alignment length of 16.4 kb [2], that is, approximately six times less than needed according to our estimates.

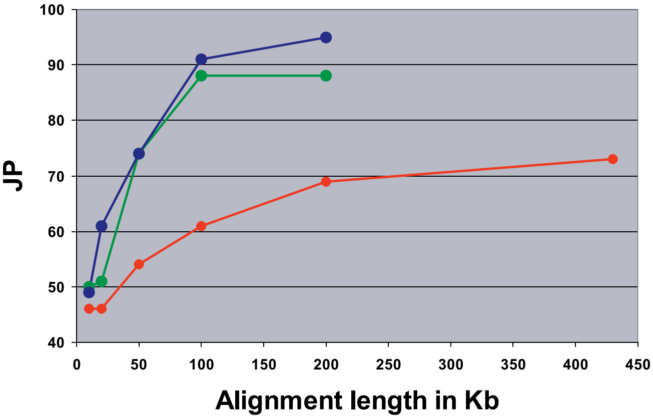

Figure 2. Jacknife Support for Exafroplacentalia Depending on the Alignment Length

Red, CNC; green, CDS_DNA; blue, CDS_AA.
doi:10.1371/journal.pgen.0030002.g002

Likelihood mapping [29] was used as a second test of assessing the quality of the phylogenetic signal contained in CNCs as compared to CDSs. In the 18 species datasets used in this study, the 3,060 possible quartets were phylogenetically analyzed. The proportion of resolved quartets indicates the amount of information in the dataset. For this test, the length of the CDS and CNC alignments were 205 kb and 430 kb, respectively.

The results showed a similar performance of the CNC dataset which gave 99.97% of resolution of 3,060 quartets (one quartet remained unresolved, 0.03%); the CDS__DNA dataset (for which six quartets were unresolved, 0.2%); and CDS__AA (one quartet unresolved, 0.03%).

Overall, the results suggest that CNCs are equally powerful phylogenetic markers as CDSs, and hence they could be used in parallel with CDSs, to maximize the statistical support of phylogenetic trees.

The recent study of retroposed LINE elements in mammals revealed a number of insertions supporting all major mammalian clades that are also supported by our analyses [10]. Two insertions (L1MB5) common for Boreoeutheria and Afrotheria that are absent in Xenarthra were found supporting the Epitheria hypothesis (X [A,B]). The analysis of rare genomic changes, such as the insertion of retroposed elements, are thought to be exceptionally useful markers due to their ambiguity-free phylogenetic information, because the coincidence of orthologous insertions of retroposed elements belonging to the same type is unlikely [30]. However, little is known about the frequency of retroposon loss by small-scale deletions. Because extant Xenarthra radiated quite recently (during the Tertiary) from a 35-million-y-long standing stem lineage [31], the probability of deletion of one or more retroposons in this 35-million-y period of time is not negligible. This explanation may reconcile the findings by Kriegs et al. [10] and the phylogenomic results presented here. Another possible explanation comes from the fact that the splitting among Afrotheria, Xenarthra, and Boreoeutheria occurred in a relatively short period of time (estimated 5–10 million y [2]), and therefore incomplete lineage sorting [32,33] may also explain the observation of Kriegs et al.

Although our study provides strong evidence for the rooting of Placentalia between Afrotheria and Exafroplacentalia, the phylogenetic signal supporting this hypothesis in 1% of mammalian genome is not sufficiently conclusive. The final resolution of the placental root might come with the addition in the genomic datasets of Xenarthra and Afrotheria species with short branch lengths. Previous phylogenetic studies suggest that short branches are expected for some xenarthrans: *Choloepus* spp. (two-toed sloth), *Cyclopes didactylus* (silky anteater), *Cabassous* spp. (naked-tailed armadillo); and afrotherians: *Dugong dugong* (Dugong), *Chrysochloris* spp. (golden mole), *Talpa* spp. (mole).

Materials and Methods

The phylogenetic analyses were based on the ENCODE TBA alignments [23]. The CDS_DNA alignment (Dataset S1) was created by concatenating the longest transcript per gene for all ENCODE targets and keeping a single reading frame. The CDS_DNA dataset was translated into amino acids. The CNC alignment (Dataset S2) was prepared by the concatenation of all CNCs >50 bp, as identified by the ENCODE Multi-Species Sequence Analysis group [23,24]. For the CNC dataset, we have screened visually for clearly misaligned parts of sequences and have deleted them. For the CDSs, we translated the DNA alignment, and the stretches with stop codons we deleted from both DNA and AA alignments. Due to the large amount of missing data, we kept only sites for which at least one representative of each main lineage was present. In this way we preserved 100% of data for Xenarthra, Marsupialia, and Monotremata, and we have full coverage across taxa for Primates, Glires, Laurasiatheria, and Afrotheria. The individual maximal amount of missing data is 32%. The amount of data included in the alignments for each species is shown in Table S1 [23].

To gain the maximum amount of positions, we selected only mammalian species with substantial amount of data. A total of 18 species representing the major mammalian groups (Table S2) were included; *Platypus* was used as an out-group. For the CDS, the length of DNA alignment is 204,786 bp, and the length of amino acid alignment is 68,262 aa positions; CNC alignments comprise 429,675 bp.

All phylogenies were performed using ML method [34] with PHYML software using NNI and SPR branch swapping methods, or imposing different starting topologies [25,35]. The GTR + G + I model of sequence evolution [36–38] was selected as the best fitting model to the data using ModelTest 3.7 program [26]. Statistical support was assessed with bootstrap analysis [39].

Baseml and Codeml programs implemented in the PAML package [28] were used to calculate ML scores for three competing topologies (Afrotheria, Epitheria, and Atlantogenata) using different models of sequence evolution. Approximately unbiased, Kishino-Hasegawa, and Shimodaira-Hasegawa topological tests (as implemented in the CONSEL package [27]) were used to test alternative rooting of placental animals.

To assess the amount of phylogenetic signal in CDS and CNC datasets we performed a jacknife analysis. For that, with the seqboot.exe program implemented in PHYLIP package [40], we created 100 jacknife replicates of the following lengths: 205 kb (68 kilo–amino acids [Kaa]), 100 kb (33 Kaa), 50 kb (17 Kaa), 20 kb (7 Kaa), and 10 kb (3.3 Kaa) from CDS_DNA, CDS_AA, and CNC matrices. With PHYML program we reconstructed trees from those replicates, and with consense.exe (PHYLIP package) we calculated JPs (Figure 2).

Using the likelihood mapping method [29], we performed a direct comparison of the amount of signal between CNC, CDS_DNA, and CDS_AA matrices by the analysis of ML for the fully resolved tree topologies that could be computed for four sequences.

Supporting Information

Dataset S1. The CDS_DNA Alignment with the Length 204,786 bp in a Subset of 18 Species

This alignment was generated by concatenating the longest transcript per gene for all ENCODE targets and by keeping a single reading frame. Clearly misaligned DNA parts were manually deleted.

Found at 10.1371/journal.pgen.0030002.sd001 (4.9 MB DOC).

Dataset S2. The CNC Alignment with the Length 429,675 bp in a Subset of 18 Species

This alignment was prepared by the concatenation of all CNCs >50 bp, as identified by the ENCODE Multi-Species Sequence Analysis group [23,24]. Clearly misaligned DNA parts were manually deleted.

Found at 10.1371/journal.pgen.0030002.sd002 (10.5 MB DOC).

Figure S1. Direct Comparison of Amount of Phylogenetic Signal between CDS_DNA, CDS_AA, and CNC Datasets

Numbers in the corners of triangles indicate the percentage of resolved quartets of species in the alignment. Numbers in the middle of the triangle and on the sides indicate the percentage of unresolved quartets.

Found at 10.1371/journal.pgen.0030002.sg001 (133 KB JPG).

Table S1. Amount of Data Included in CNC and CDS Datasets per Species

Found at 10.1371/journal.pgen.0030002.st001 (51 KB DOC).

Table S2. Latin Names of the Species

Found at 10.1371/journal.pgen.0030002.st002 (34 KB DOC).

Acknowledgments

We thank members of the ENCODE Multi-Species Sequence Analysis group (leadership provided by G. M. Cooper and EHM) for providing us with the sequence data, alignments, and CNCs used here. Leadership for the National Intramural Sequencing Center (NISC) Comparative Sequencing Program is provided by E. D. Green, R. W. Blakesley, G. G. Bouffard, N. F. Hansen, B. Maskeri, P. J. Thomas, J. C. McDowell, and M. Park. We also thank the National Swiss Foundation and the National Center of Competence in Research (NCCR) Frontiers in Genetics program.

Working in the NISC Comparative Sequencing Program are Gerard G. Bouffard, Jacquelyn R. Idol, Valerie V. B. Maduro, and Robert W. Blakesley at the Genome Technology Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland, United States of America; and Gerard G. Bouffard, Xiaobin Guan, Nancy F. Hansen, Baishali Maskeri, Jennifer C. McDowell, Morgan Park, Pamela J. Thomas, Alice C. Young, and Robert W. Blakesley at the Intramural Sequencing Center, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland, United States of America.

Author contributions. SN and JIMB conceived and designed the experiments. SN performed the experiments and analyzed the data. JR and BN contributed analysis tools. SN and JIMB wrote the paper. EHM worked with the ENCODE consortium for the collection and availability of the data. SEA initiated the project and worked with the ENCODE consortium for the collection and availability of the data. NISC Comparative Sequencing Program provided genomic sequence data.

Funding. This research was supported by the Intramural Research Program of the National Human Genome Research Institute, National Institutes of Health.

Competing interests. The authors have declared that no competing interests exist.

References

1. Waddell PJ, Kishino H, Ota R (2001) A phylogenetic foundation for comparative mammalian genomics. Genome Inform Ser Workshop Genome Inform 12: 141–154.
2. Murphy WJ, Eizirik E, O'Brien SJ, Madsen O, Scally M, et al. (2001) Resolution of the early placental mammal radiation using Bayesian phylogenetics. Science 294: 2348–2351.
3. Lin YH, McLenachan PA, Gore AR, Phillips MJ, Ota R, et al. (2002) Four new mitochondrial genomes and the increased stability of evolutionary trees of mammals from improved taxon sampling. Mol Biol Evol 19: 2060–2070.
4. Delsuc F, Scally M, Madsen O, Stanhope MJ, de Jong WW, et al. (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.
5. 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.
6. Reyes A, Gissi C, Catzeflis F, Nevo E, Pesole G, et al. (2004) Congruent

mammalian trees from mitochondrial and nuclear genes using Bayesian methods. Mol Biol Evol 21: 397–403.

7. 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.
8. Murphy WJ, Eizirik E, Johnson WE, Zhang YP, Ryder OA, et al. (2001) Molecular phylogenetics and the origins of placental mammals. Nature 409: 614–618.
9. 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.
10. Kriegs JO, Churakov G, Kiefmann M, Jordan U, Brosius J, et al. (2006) Retroposed elements as archives for the evolutionary history of placental mammals. PLoS Biol 4(4): e91. doi:10.1371/journal.pbio.0040091
11. 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.
12. Springer MS, Stanhope MJ, Madsen O, De Jong WW (2004) Molecules consolidate the placental mammal tree. Trends Ecol Evol 19: 430–438.
13. Misawa K, Nei M (2003) Reanalysis of Murphy et al.'s data gives various mammalian phylogenies and suggests overcredibility of Bayesian trees. J Mol Evol 57 (Suppl 1): S290–S296.
14. Rokas A, Kruger D, Carroll SB (2005) Animal evolution and the molecular signature of radiations compressed in time. Science 310: 1933–1938.
15. Benton MJ (1999) Early origins of modern birds and mammals: Molecules versus morphology. Bioessays 21: 1043–1051
16. ENCODE Project Consortium (2004) The ENCODE (ENCyclopedia Of DNA Elements) Project. Science 306: 636–640.
17. Chiaromonte F, Yang S, Elnitski L, Yap VB, Miller W, et al. (2001) Association between divergence and interspersed repeats in mammalian noncoding genomic DNA. Proc Natl Acad Sci U S A 98: 14503–14508.
18. Dermitzakis ET, Clark AG (2002) Evolution of transcription factor binding sites in mammalian gene regulatory regions: Conservation and turnover. Mol Biol Evol 19: 1114–1121.
19. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, et al. (2002) Initial sequencing and comparative analysis of the mouse genome. Nature 420: 520–562.
20. Thomas JW, Touchman JW, Blakesley RW, Bouffard GG, Beckstrom-Sternberg SM, et al. (2003) Comparative analyses of multi-species sequences from targeted genomic regions. Nature 424: 788–793.
21. Dermitzakis ET, Reymond A, Scamuffa N, Ucla C, Kirkness E, et al. (2003) Evolutionary discrimination of mammalian conserved non-genic sequences (CNGs). Science 302: 1033–1035.
22. Dermitzakis ET, Reymond A, Antonarakis SE (2005) Conserved non-genic sequences: An unexpected feature of mammalian genomes. Nat Rev Genet 6: 151–157.
23. The ENCODE Consortium (2006) The ENCODE pilot project: Functional annotation of 1% of the human genome. Nature. In press.
24. Margulies EH, Cooper GM, Asimenos G, Thomas DJ, et al. (2006) Annotation of the human genome through comparisons of diverse mammalian sequences. Genome Res. In press.
25. Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol 52: 696–704.
26. Posada D, Crandall KA (1998) Modeltest: Testing the model of DNA substitution. Bioinformatics 14: 817–818.
27. Shimodaira H, Hasegawa M (2001) CONSEL: For assessing the confidence of phylogenetic tree selection. Bioinformatics 17: 1246–1247
28. Yang Z (1997) PAML: A program package for phylogenetic analysis by maximum likelihood. Comput Appli BioSci 13: 555–556
29. Strimmer K, von Haeseler A (1997) Likelihood-mapping: A simple method to visualize phylogenetic content of a sequence alignment. Proc Natl Acad Sci U S A 94: 6815–6819.
30. Shedlock AM, Okada N (2000) SINE insertions: Powerful tools for molecular systematics. Bioessays 22: 148–160.
31. 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.
32. Nei M (1987) Molecular evolutionary genetics. New York: Columbia University Press. 512 p.
33. Pamilo P, Nei M (1988) Relationships between gene trees and species trees. Mol Biol Evol 5: 568–583.
34. Felsenstein J (1981) Evolutionary trees from DNA sequences: A maximum likelihood approach. J Mol Evol 17: 368–376.
35. Hordijk W, Gascuel O (2005) Improving the efficiency of SPR moves in phylogenetic tree search algorithms based on maximum-likelihood. Bioinformatics 21: 4338–4347.
36. Lanave C, Preparata G, Saccone C, Serio G (1984) A new method for calculating evolutionary substitution rates. J Mol Evol 20: 86–93.
37. Rodriguez R, Olivier JL, Marin A, Medina JR (1990) The general stochastic model of nucleotide substitution. J Theor Biol 142: 485–501.
38. Jones DT, Taylor WR, Thornton JM (1992) The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci 8: 275–282.
39. Felsenstein J (1985) Confidence limits on phylogenies: An approach using the bootstrap. Evolution 39: 783–791.
40. Felsenstein J (1989) PHYLIP: Phylogeny Inference Package (Version 3.2). Cladistics 5: 164–166.