Coalescent-Based Genome Analyses Resolve the Early Branches of the Euarchontoglires

Vikas Kumar[1], **Björn M. Hallström**[1,3], **Axel Janke**[1,2]*

[1] Biodiversity and Climate Research Centre (BiK-F) and Senckenberg Gesellschaft für Naturforschung, Frankfurt am Main, Germany, [2] Goethe University Frankfurt, Institute for Ecology, Evolution and Diversity, Frankfurt am Main, Germany, [3] Science for Life Laboratory, Royal Institute of Technology, Stockholm, Sweden

Abstract

Despite numerous large-scale phylogenomic studies, certain parts of the mammalian tree are extraordinarily difficult to resolve. We used the coding regions from 19 completely sequenced genomes to study the relationships within the super-clade Euarchontoglires (Primates, Rodentia, Lagomorpha, Dermoptera and Scandentia) because the placement of Scandentia within this clade is controversial. The difficulty in resolving this issue is due to the short time spans between the early divergences of Euarchontoglires, which may cause incongruent gene trees. The conflict in the data can be depicted by network analyses and the contentious relationships are best reconstructed by coalescent-based analyses. This method is expected to be superior to analyses of concatenated data in reconstructing a species tree from numerous gene trees. The total concatenated dataset used to study the relationships in this group comprises 5,875 protein-coding genes (9,799,170 nucleotides) from all orders except Dermoptera (flying lemurs). Reconstruction of the species tree from 1,006 gene trees using coalescent models placed Scandentia as sister group to the primates, which is in agreement with maximum likelihood analyses of concatenated nucleotide sequence data. Additionally, both analytical approaches favoured the Tarsier to be sister taxon to Anthropoidea, thus belonging to the Haplorrhine clade. When divergence times are short such as in radiations over periods of a few million years, even genome scale analyses struggle to resolve phylogenetic relationships. On these short branches processes such as incomplete lineage sorting and possibly hybridization occur and make it preferable to base phylogenomic analyses on coalescent methods.

Citation: Kumar V, Hallström BM, Janke A (2013) Coalescent-Based Genome Analyses Resolve the Early Branches of the Euarchontoglires. PLoS ONE 8(4): e60019. doi:10.1371/journal.pone.0060019

Editor: Hans Ellegren, University of Uppsala, Sweden

Received December 14, 2012; **Accepted** February 20, 2013; **Published** April 1, 2013

Funding: The present study was supported by the research funding program LOEWE: "Landes-Offensive zur Entwicklung Wissenschaftlich-ökonomischer Exzellenz" of Hesse's Ministry of Higher Education, Research, and the Arts. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

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

* E-mail: axel.janke@senckenberg.de

Introduction

Improvements of phylogenetic methods and the availability of numerous placental mammalian genomes provide an invaluable resource to investigate controversial relationships with increasingly larger datasets and refined methods. The first phylogenomic studies utilising protein-coding sequences of mammals were optimistic to fully resolve the placental mammalian tree [1,2]. However, some of the branches proved to be difficult to resolve even with these increasingly gene and taxon rich datasets and some relationships did not hold up to closer inspection [3–5]. Revisiting such groups using phylogenomic and network analyses provided new insights into placental mammalian evolution [4,5]. One placental mammalian group with uncertain phylogenetic relationships among its orders is Euarchontoglires. This clade has been intensively studied through the use of concatenated genomic sequence data, retroposon insertion analyses, and multi-locus analyses of genomic data [6–9]. While some studies agree and give convincing support for internal branches [3,4,8,9], the major difference was the placement of the tree shrews relative to primates and Glires. In phylogenomic analyses a grouping with the primates was preferred, but the alternatives could not formally be rejected [3,4]. The conflicting results of their relationships in earlier studies makes the Euarchontoglires an interesting group to revisit with a significantly increased genomic sequence data set using a multi-locus analysis approach. Recent coalescent-based analyses seem to have solved the question of resolving Euarchontoglires [9], with results being congruent with previous phylogenomic studies [3].

Euarchontoglires is a super-clade of placental mammals that includes primates (apes, monkeys and allies), rodents (mouse, rat and guinea pig), lagomorphs, (rabbit and hares), dermopterans (flying lemurs) and scandentians (tree shrew), and has recently been established exclusively through molecular analyses [6,10]. The original proposal of the Archonta clade [11] initially included Chiroptera (bat and flying foxes) and then Macroscelidea (elephant shrews), but these groups were later moved to other parts of the placental mammalian tree [12–14]. While the signal from molecular data for the monophyly of Euarchontoglires is strong, some details of the relationships within this clade remain ambiguous. In particular, the position of Scandentia differed in the numerous molecular studies, from being sister group to the primates [6,8,9,14,15], sister group to lagomorphs [16,17], sister group to Glires (rodents plus lagomorphs) [4,18] to being the first ordinal branch among the Euarchontoglires [19]. One large phylogenomic analysis using concatenated data from 2.9 million nucleotides (16 species) could reject a sister group relationship of Scandentia to primates plus Glires, but not Scandentia and primates as sister groups as an alternative to a first divergence among Euarchontoglires [4]. The branches of the placental mammalian tree that have been problematic to resolve are typically short with subsequent divergences occurring within 1–3 million years (Myr) [4,20]. This duration is approximately the time-span establish a mammalian species [21–23] and a time frame in which introgression can occur and incomplete lineage sorting can later complicate phylogenomic analyses [24]. Speciation processes appear to be the main reason for the incongruent results of phylogenomic attempts to resolve the mammalian species tree.

Therefore, we revisit the Euarchontoglires with a new and larger set of genomic sequence data focusing on ordinal relationships within the Euarchontoglires along with the relationship of the tarsiers (haplorhine primates) relative to the other two major primate lineages, Strepsirrhini (lemurs, lorises and allies) and Anthropoidea (platyrrhines and catarrhines). The widely used concatenation method to infer the species tree can mislead inferences, especially when gene trees are in the so-called anomaly zone [25]. In addition to this, species trees generated using concatenated data can sometimes have very high bootstrap support values for incorrect relationships [26]. Coalescent theory tries to trace back alleles to the most recent common ancestor [27]. Recently developed tree reconstruction methods such as STAR, species tree based on ranks of coalescent [28] and MP-EST, maximum pseudo-likelihood estimates method [29] estimates a species tree from multiple gene trees. STAR and MP-EST methods can be applied to large data sets because of their faster and simpler algorithms. They provide reliable results in reasonable computational time compared to more computationally intensive and exact Bayesian species tree methods (such as BEST), which are preferable for smaller data sets [28].

Compared to recent phylogenomic studies [4,9,30] the new dataset now includes sequences of the Gibbon genome (*Nomascus leucogenys*). Genome data of the cow (*Bos taurus*) is used to root the tree. Among the possible outgroups to the Euarchontoglires [4] the bovine genome is arguably the one that has the best sequence coverage and annotation. The root is especially crucial for the coalescence-based species tree analyses such as STAR and MP-EST. Choosing a well-assembled genome of a closely related species facilitates the alignment and maximizes the amount of data that can be utilized. The placement of Dermoptera is also uncertain and previous studies have placed this order with the primates [7] or with Scandentia [6]. However, until a genome of this order is available, their position in the mammalian tree cannot be studied by phylogenomic approaches.

Materials and Methods

The coding sequences (CDS) for the 19 species that are included in this study, listed in Table 1, were retrieved using Biomart from the Ensembl version 65 (http://www.ensembl.org/biomart/martview/). The procedure for alignment and phylogenomic analyses were similar to that previously described [4]. Therefore it is only briefly described here, detailing only additional procedures.

Human sequences longer than 300 nucleotides were used to find the orthologs from 18 different species using the recursive BLAST approach [3]. Only genes represented by at least 16 out of 19 species were kept for further analyses. All sequences were translated into corresponding amino acids and aligned using MAFFT version 6.833b [31]. Any alignment showing an overall nucleotides difference larger than 25% between any two species were discarded to ensure a conservative approach and further reduce the potential for incorrect alignments.

It has been shown that the quality of multiple sequence alignment is essential in phylogenetic inference [32]. Therefore, we used BMGE, Block Mapping and Gathering with Entropy, [33] that utilizes similarity matrices such as BLOSUM and PAM to remove ambiguously aligned regions. We used the option of stringent trimming based on the scoring scheme BLOSUM 95. The selected amino acid alignments were then back translated to nucleotide-alignments. Both types of sequence data were analysed.

Base composition analysis was done using the Treefinder (TF) version of March 2011 [34] to test for the compositional equilibrium of the bases across the species. For this we applied the sliding window approach using the default size of 500.

It is prohibitive to run the model-test for large genomic scale data sets because of the large computational demands [3]. Similar phylogenomic studies [3,5] estimated the preferred model to be GTR [35] and WAG2000 [36] with 4 gamma rate categories (4G) for nucleotide and amino acids respectively. These models were also estimated for a smaller dataset of 1,006 loci using the model test of TF. Assuming the remaining data set to have similar properties, these models were used subsequently for all maximum-likelihood (ML) analyses. Initially concatenated data sets for both nucleotide and amino acid sequences were used to infer the species tree using TF, reconstructing a ML tree using the GTR and WAG2000 models, respectively, and 4 gamma rate categories (4G) as rate heterogeneity parameter. ML hypothesis testing of different topologies for Scandentia and the tarsier within the Euarchontoglires were done with the approximately unbiased test (AU) [37] and Shimodaira- Hasegawa (SH) [38], as implemented in TF. These analyses are consistent with previous phylogenomic studies, thereby enabling comparisons to previous results.

In the second step we analysed the nucleotide data using coalescent methods. We constructed the nucleotide species tree using the coalescent model of evolution implemented in the STAR method and with the MP-EST method implemented as an R package in Phybase [39]. The STAR method is motivated by multispecies coalescent model [40], which assumes deep coalescence to be the major factor for the differences between the gene trees and species tree. STAR uses the rank of coalescence to coalesce the gene trees into species tree. For this analysis each taxon must be represented by a sequence for each gene. From the 5,875 gene data set, 1,006 gene alignments fulfilled this criterion. We also constructed the concatenated tree using the nucleotide and amino acid data.

The 1,006 (Table S1) loci were used for the species tree reconstruction. The longest sequence length is 25,728 base pairs and average length was 1,757 base pairs (Figure S1). Computational constraints limited the multi-locus bootstrapping [26] to 100 replicates, using the bootstrap 'mulgene' method implemented in Phybase. Each gene was bootstrapped and the combined trees (100,600) served as input for the STAR program. Individual ML gene trees were generated for each alignment using PhyML 3.0 [41] applying the GTR model with gamma distribution. From these trees, 100 re-sampled species trees were generated using the STAR method implemented in R. The species trees were rooted with the cow and finally a consensus tree was made using the “consense” module in Phylip, version 3.69 [42].

A second coalescent-based method “MP-EST”, which implements a pseudo maximum likelihood method under the coalescent model to estimate the species tree, was also used to generate a species tree. The MP-EST method has been shown to be more accurate than STAR when inferring short branches in a species tree [29]. All 1,006 gene trees were again used as input for the MP-EST method to generate the species tree. The procedures of

Table 1. List of species included in the study and the percent coverage of alignment.

Common name	Binomial name	Order	Coverage of alignment (%)
Chimpanzee	*Pan troglodytes*	Primates	95.4
Human	*Homo sapiens*	Primates	99.9
Gorilla	*Gorilla gorilla*	Primates	92.6
Orangutan	*Pongo abelii*	Primates	91.5
Gibbon	*Nomascus leucogenys*	Primates	94.0
Macaque	*Macaca mulatta*	Primates	89.9
Marmoset	*Callithrix jacchus*	Primates	92.3
Tarsier	*Tarsius syrichta*	Primates	66.2
Bushbaby	*Otolemur garnettii*	Primates	93.6
Mouse lemur	*Microcebus murinus*	Primates	71.2
Tree shrew	*Tupaia belangeri*	Scandentia	82.5
Mouse	*Mus musculus*	Rodentia	96.5
Rat	*Rattus norvegicus*	Rodentia	88.6
Kangaroo rat	*Dipodomys ordii*	Rodentia	70.2
Guinea pig	*Cavia porcellus*	Rodentia	91.5
Squirrel	*Ictidomys tridecemlineatus*	Rodentia	54.4
Pika	*Ochotona princeps*	Lagomorpha	71.6
Rabbit	*Oryctolagus cuniculus*	Lagomorpha	82.7
Cow	*Bos taurus*	Artiodactyla	94.5

Coverage of alignment is the percent sequence coverage in 9,799,170 nucleotide long alignment.
doi:10.1371/journal.pone.0060019.t001

tree construction, including the bootstrap analyses, were similar to that described for the STAR method.

We also performed network analyses to depict conflicting signal. From the 1,006 amino acid sequence alignment, individual ML trees were generated by TF. A consensus network was built from the individual gene trees using the SplitsTree4 program [43], with a threshold of 10%. In addition, we selected the best ML tree from the three alternative hypotheses regarding the position of the tree shrew. Only gene trees that were separated from the second best ML tree by an arbitrary value >0.7 standard deviations (s.d.) were retained and used for network construction. Choosing a cut-off of 0.7 s.d. allows best depicting the conflict in the networks and retains only topologies with some moderate support from single gene analyses. The three different placements of the tree shrew as earlier described were evaluated for all the three different topologies of tree shrew. A consensus network in Splits Tree4 then summarized the trees.

Results

The final dataset consisted of 5,875 orthologous gene alignments from 19 different species, including ten primates, seven rodents and lagomorphs, one scandentian, and one outgroup species, the cow (Table 1). After trimming the alignments using BMGE [33], we generated a concatenated alignment of 9,799,170 nucleotides with an average sequence coverage of 85% for each species (Table 1). This resulted in 71% more sites than a previous phylogenomic studies that included Euarchontoglires [4]. The base composition showed high homogeneity between the species both for all codon positions (NT123) and first and second codon position alone (NT12) (Table S2, Table S3).

The ML consensus trees from concatenated amino acid and nucleotide data supported different topologies for the position of Scandentia. The nucleotide analysis supported a sister group relationship between Scandentia and the primates, with 100% support both including and excluding the third codon position (Figure 1). The amino acid ML analysis, involved 3,266,390 sites and found Scandentia as the outgroup to both the primates and Glires, albeit with negligible support (Figure 2). The three proposed hypotheses for the position of Scandentia were tested by ML analyses and the results are summarized in Table 2. The support from both the resampling and comparative likelihood tests for the position of Scandentia is ambiguous when analysing the amino acid sequences, supporting the hypothesis of Scandentia either as outgroup to primates or to Glires. Yet, nucleotide data analyses unambiguously support the Scandentia and Primate sister group relationship both when tested with all the three codon positions and first two codon positions.

The concatenated tree analysis of the nucleotide sequence data from 1,006 genes was congruent with the analysis of the 5,875 genes, but lacked strong support most likely due to the reduced amount of data. The species tree topology constructed by STAR, using the 1,006 gene trees, supported the concatenated nucleotide consensus species tree using the 5,875 loci (Figure 3). The bootstrap support for the Scandentia-primates relationship was 94%. MP-EST yielded the same topology for the species tree with a bootstrap support of 86% for the Scandentia-primates branch (Figure 3). Thus, both coalescent multi gene analyses of nucleotide sequences support the Scandentia-primates relationship with high bootstrap support values.

The position of the tarsier within Euarchontoglires is established with high support in all analyses. The tarsier is placed as sister taxon to the Simiformes with both the concatenated nucleotide and amino acid data analysis (Figure 1, 2). Statistical ML analyses of the three different hypotheses for the position of the tarsier and for all types of sequences data reject alternative positions (Table 3). Also, the coalescent based methods supported the same topology

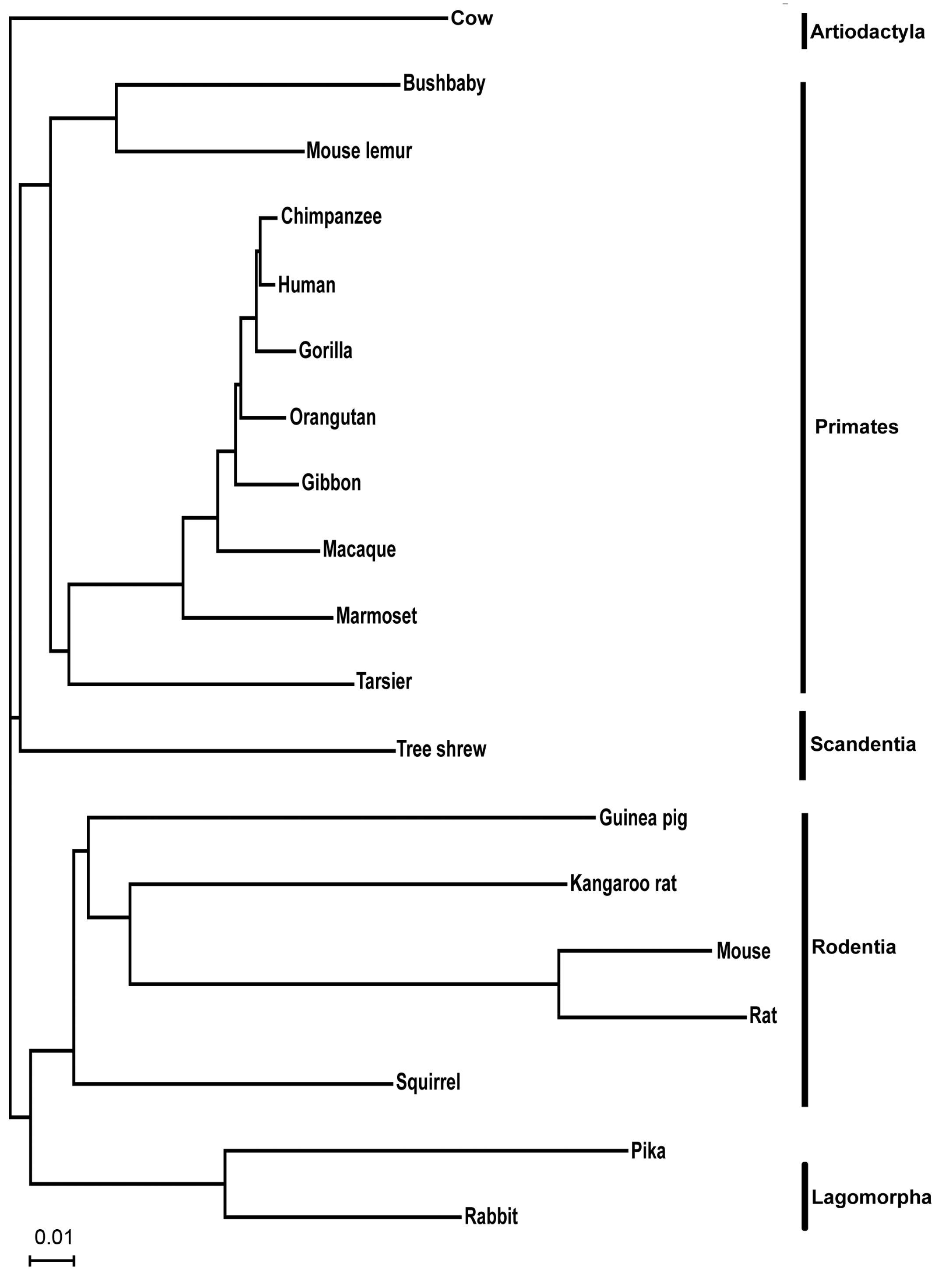

Figure 1. The ML tree of concatenated nucleotides data from 5,875 genes with all the branches being unanimously supported by TF.
doi:10.1371/journal.pone.0060019.g001

as the concatenated analysis with 100% bootstrap support, suggesting the tarsier as sister taxon to the Simiiformes (Figure 3).

The consensus network analysis of all 1,006 genes that are represented in all species shows conflicting branches when limited to splits that are present in at least 10% of the data (10% threshold value) shown in Figure 4. The major signal from this single gene analyses places the tree shrew as sister group to the primates, or basal to all Euarchontoglires. Depicting the position of the tree shrew in a network based on the best ML trees from three alternative hypotheses of the tree shrew position, yielded no further resolution or insight into the evolutionary process (Figure S2).

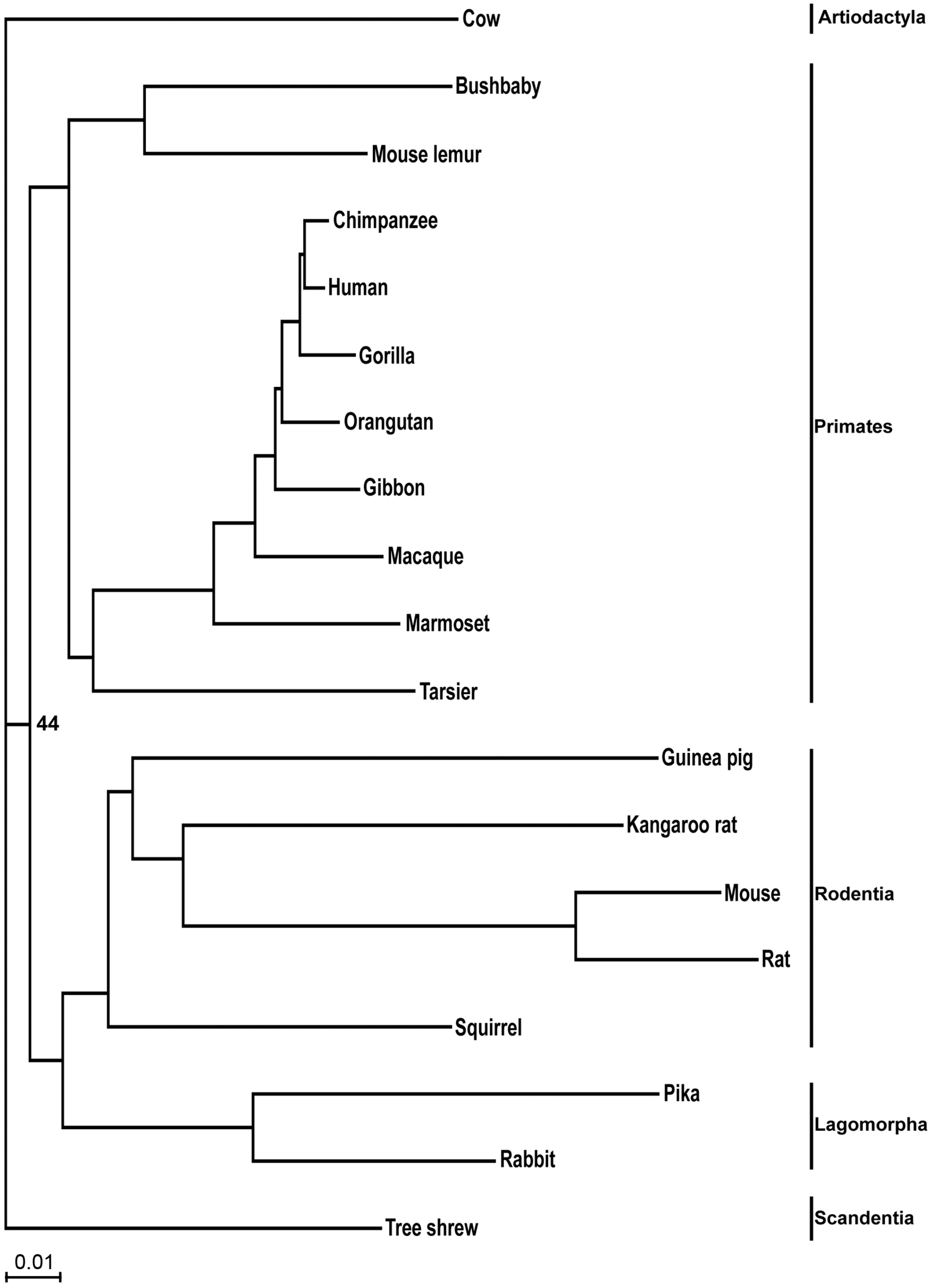

Figure 2. The ML tree based on amino acid data from 5,875 genes representing the best option for a bifurcating topology. Only the TF support values <99 are shown.
doi:10.1371/journal.pone.0060019.g002

Discussion

The increase in the availability of complete mammalian genomes has been seen as an opportunity to fully resolve all branches in the mammalian tree. Yet, numerous studies using genome scale data [3–5,44] yielded the insight that species tree reconstruction is complicated by incomplete lineage sorting and possibly introgression. One major group that was problematic to resolve is the Euarchontoglires clade. In many studies the branches connecting Scandentia and tarsier to the tree were found to be difficult to place within this super-order. The reason for this is that short internal branches have been identified to be a major cause of uncertainty in most phylogenomic studies [3,4,45]. By increasing

Table 2. ML tests statistics for different relationships of the Scandentia within the Euarchontoglires.

Topologies	pSH (AA)	pAU (AA)	pSH (NT12)	pAU (NT12)	pSH (NT123)	pAU (NT123)
((Scandentia,Primates),Glires)	0.23	<0.001	1	0.11	1	0.23
((Scandentia,Glires),Primates)	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
(Scandentia,(Primates,Glires))	1	0.94	<0.001	<0.001	<0.001	<0.001

pSH (probability Shimodaira Hasegawa) and pAU (probability Approximate Unbiased) ML test values are shown.
doi:10.1371/journal.pone.0060019.t002

the taxon sampling and taking advantage of coalescent-based analytical approaches this study investigated the relationships within Euarchontoglires in more detail and with new analytical approaches.

It has been shown that in multi-locus data analysis, gene tree heterogeneity and the conflicts arising due to different gene trees complicate the process of correctly inferring the species tree [24]. Gene duplication, incomplete lineage sorting and deep coalescence are obstacles in correctly inferring the species tree. Also, the entire process of inferring the species tree becomes more complicated as the amount of multi-locus data increases. To avoid analytical problems due to gene duplication we selected strictly orthologous genes. For getting a comprehensive representation of the ML tree, we began the analysis with the normal concatenation method, which has been used in numerous studies of mammalian evolution [1,2,4,9,44]. The resulting ML tree supported the general consensus that Scandentia is the sister group to primates. Thus, focusing on the resolution of the Euarchontoglires and increasing the dataset, yielded robust results [6,9,15]. Only the analysis of amino acid sequences failed to provide the statistical support for the position of Scandentia. Yet, testing different hypotheses for the position of Scandentia yielded high support from nucleotide sequences for the topology shown in Figure 1 and 3.

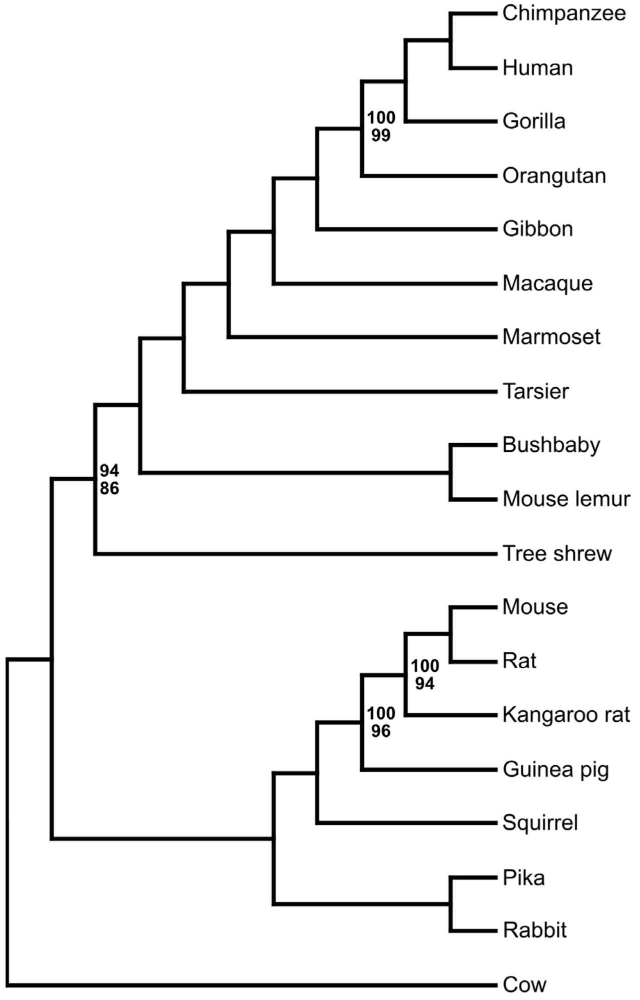

Figure 3. Species tree based on 1,006 gene trees with bootstrap support values (>99% not shown). Above value indicates the STAR support value and MP-EST values are indicated below.
doi:10.1371/journal.pone.0060019.g003

It has been debated [46–49] whether nucleotide or amino acid data contains more reliable phylogenetic information. Generally, the use of amino acid data is advised, because amino acid sequences are expected to be less randomized than nucleotide sequences for ancient divergences [48]. However, by using a set of selected genes it has been shown that nucleotide sequence data can outperform the amino acid sequence data for phylogenetic information on time scales of less than 500 million years [48]. In this study we analyse time scales in the range of 80–90 Ma and with short divergence intervals, where coalescence and introgression can complicate phylogenetic analysis. Under these conditions amino acid sequence data may be too conserved to contain sufficient phylogenetic information.

It has been observed that the approach of using concatenated sequences under certain conditions can obscure important phylogenetic signal [50]. Incongruent gene-trees can mislead phylogenetic analyses of concatenated sequences and result in erroneous interpretations of the species relationships and sometimes the incorrect species trees receive high support values [51]. The solution to this is the analysis of individual genes and their associated evolutionary signal (gene-tree) in a coalescence based framework to recover a final species tree [40]. STAR [28] is one such method for species tree reconstruction and has been successfully used for studying the phylogeny of placental mammals from protein coding genes [9] and from ultra conserved element sequences within mammals [45]. Similar to a recent phylogenomic study using multi-locus analyses [9], albeit with more than twice as much data and using a strict approach to identify orthologs, we find a clear support of Scandentia being the sister group to primates. The large data set used in this study found the same species tree in the multi locus STAR analysis as in the concatenated analysis from nucleotide sequence data. Thus, analyses of large concatenated data sets can yield the same phylogenetic results as multi-locus analyses. Congruence between the approaches increases the confidence to have identified the historic species tree, despite the conflict in individual gene analyses that is revealed by networks. For theoretical reasons a full multi-locus coalescent analysis had been preferable, but this approach is

Table 3. ML tests statistics for different relationships of the tarsier within the Euarchontoglires.

Topologies	pSH (AA)	pAU (AA)	pSH (NT12)	pAU (NT12)	pSH (NT123)	pSH (NT123)
((Tarsier,Anthropoidea),Strepsirrhini)	1	0.23	1	0.23	1	0.23
((Tarsier,Strepsirrhini),Anthropoidea)	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
(Tarsier,(Anthropoidea,Strepsirrhini))	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001

pSH (probability Shimodaira Hasegawa) and pAU (probability Approximate Unbiased) ML test values are shown.
doi:10.1371/journal.pone.0060019.t003

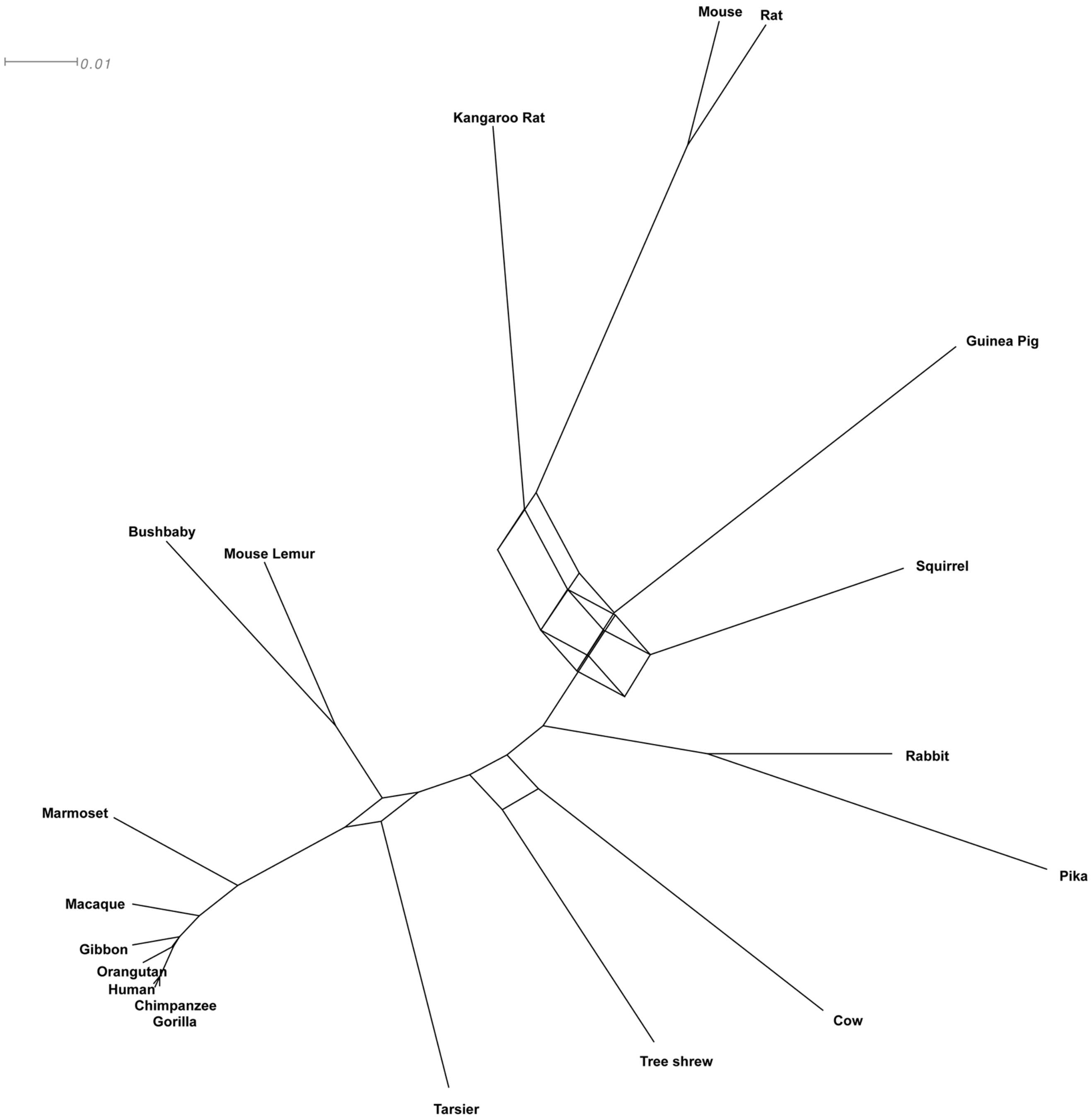

Figure 4. Consensus network in which at least 10% of the 1006 ML gene trees have common branches (threshold value 10%).
doi:10.1371/journal.pone.0060019.g004

still prohibitive for large datasets. The dramatic increase in data was only possible by focussing on the phylogeny of the Euarchontoglires phylogeny and choice of a well sequenced outgroup.

To further consolidate the results obtained from the STAR analysis we employed MP-EST analysis [29] as an additional coalescent-based method. It has been shown to be equally reliable as STAR [28], but uses a pseudo-likelihood method in the environment of coalescence theory. The results from MP-EST are congruent with the STAR method, yielding identical topology with high support values. Both methods, STAR and MP-EST require more data to reconstruct confident species trees, because they use a partial parametric method and summary statistics [28,29]. However, compared to fully parametric methods, STAR and MP-EST allow analysing large and taxon-rich datasets within reasonable time. In contrast to the congruence of the concatenated and multi-locus coalescent analyses, the analysis of a smaller data set (447 genes) on the whole mammalian tree [9] found differences in the two approaches. The Scandentia grouped with the primates using multi-locus analyses, a result favoured by the authors for the new methodology [9]. The concatenated analyses grouped the Scandentia with Rodentia [9], however, this grouping that has been rejected previously [3,4].

Regardless of the analytical approach, conflict in phylogenetic data needs to be shown either by careful ML analyses of alternative trees or by the phylogenetic signal from single gene analyses. The phylogenetic signal of conflicting data from single gene trees can be ideally depicted by network analyses [52]. The network depicts the previous difficulties to resolve the relationship of the Scandentia by sequence analyses even from concatenating genome sequence data [4,5,9] with nearly equally long edges, but no connection to the rodents. Interestingly, retroposon insertion analyses have so far yielded a clear signal [8] with no conflicting data for this branch. In comparison, in other studies of deep mammalian divergences some splits were problematic to resolve from this data as bifurcating tree, because of conflicting signal [4,5,53], suggesting that incomplete lineage sorting and/or hybridization obscure short branches [4,53]. It remains to be shown, if conflicting retroposon insertions are present for Euarchontoglires. However, the high evolutionary rate in the rodents will make it difficult to study neutral sequences like that of retroposon insertions in further detail, because sequence similarity in rodents is highly eroded over the 80 million years of their evolution [4].

Another challenging to resolve relationship has been the tarsier's grouping with anthropoids (platyrrhines and catarrhines) or Strepsirrhini (Lemuriformes and Lorisiforms). It has remained controversial, because contrasting phylogenetic signals from molecular data support different topologies [30]. This conflict is also visible in the network analysis (Figure 4). Our coalescent based analyses, however, confirm tarsier as sister taxon to the Simiiformes [30], which together with the Tarsiiformes form the Haplorhini clade. This relationship was identified by analyses of concatenated data with unanimous support.

Conclusions

The presently largest data set for a multi-locus analyses of mammalian relationships resolved the long challenge of placing Scandentia as the sister group to the primates, as has been previously suggested [6,9,15]. Multi-locus analyses settled the grouping of the tarsier with anthropoids. This leaves the dermopterans as the last order to be placed in the euarchontogliran tree. New mammalian genomes, and further development of methods will soon finalize the ordinal relationships among the Euarchontoglires.

Network analyses are a valuable tool to depict and evaluate conflict in gene trees that can only be identified in genome-scale phylogenetic analyses. These conflicts from multiple gene trees can now be resolved into a reliable species tree by recent implementations of coalescence-based methods into phylogenetic analysis programs [28,29]. The necessity of using methods developed for population genetics for deep divergences is a surprising development, because higher-level relationships have been expected to be deeper than the coalescent times of most genes. However, phylogenomic studies have shown that this is not always the case and speciation related processes interfere with phylogenetic analyses [3–5]. While the use of concatenated sequences generally improves the resolution of the phylogenetic tree, the current development in mammalian evolutionary studies [9] show that this approach may in some cases be uncertain and multi-locus species tree analyses are preferred to yield a reliable and sound species phylogeny even for divergences as deep as that among mammalian orders.

Supporting Information

Figure S1 Length Distribution of the 1006 longest gene trees.
(TIF)

Figure S2 Consensus Network of selected 661 genes. Gene selected on basis of supporting the best topology of tree shrew position with a LogL value larger than standard deviation of >0.7 compared to best tree.
(TIF)

Table S1 List of 1,006 loci (.xls file, Ensembl transcript id) from all the species.
(DOC)

Table S2 Base composition for each species for all the three nucleotide positions (NT123).
(DOC)

Table S3 Base Composition for each species for all the first two nucleotide positions (NT12).
(DOC)

Acknowledgments

We are grateful to Dr. Nilsson for comments on the retroposon insertion analyses.

Author Contributions

Conceived and designed the experiments: AJ BH. Performed the experiments: VK. Analyzed the data: VK. Wrote the paper: VK BH AJ.

References

1. Hallström 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.
2. Wildman DE, Uddin M, Opazo JC, Liu G, Lefort V, et al. (2007) Genomics, biogeography, and the diversification of placental mammals. Proc Natl Acad Sci U S A 104: 14395–14400.

3. Hallström BM, Janke A (2008) Resolution among major placental mammal interordinal relationships with genome data imply that speciation influenced their earliest radiations. BMC Evol Biol 8: 162.
4. Hallström BM, Janke A (2010) Mammalian evolution may not be strictly bifurcating. Mol Biol Evol 27: 2804–2816.
5. Hallström BM, Schneider A, Zoller S, Janke A (2011) A genomic approach to examine the complex evolution of laurasiatherian mammals. PLoS ONE 6: e28199.
6. 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.
7. Janecka JE, Miller W, Pringle TH, Wiens F, Zitzmann A, et al. (2007) Molecular and genomic data Identify the closest living relative of primates. Science 318: 792–794.
8. Kriegs JO, Churakov G, Jurka J, Brosius J, Schmitz J (2007) Evolutionary history of 7SL RNA-derived SINEs in Supraprimates. Trends Genet 23: 158–161.
9. Song S, Liu L, Edwards SV, Wu S (2012) Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. Proc Natl Acad Sci U S A 109: 14942–14947.
10. Asher RJ, Helgen KM (2010) Nomenclature and placental mammal phylogeny. BMC Evol Biol 10: 102.
11. Gregory WK (1910) The Orders of Mammals. Bulletin of the American Museum of Natural History. 1–524 p.
12. Pumo DE, Finamore PS, Franek WR, Phillips CJ, Tarzami S, et al. (1998) Complete mitochondrial genome of a neotropical fruit bat *Artibeus jamaicensis* and a new hypothesis of the relationships of bats to other eutherian mammals. J Mol Evol 47: 709–717.
13. Arnason U, Adegoke JA, Gullberg A, Harley EH, Janke A, et al. (2008) Mitogenomic relationships of placental mammals and molecular estimates of their divergences. Gene 421: 37–51.
14. McKenna MC (1975) Toward a phylogeny and classification of the Mammalia. Phylogeny of the Primates: a Multidisciplinary Approach Plenum, New York Luckett WP, Szalay FS. 21–46 p.
15. Novacek MJ (1992) Mammalian phylogeny: shaking the tree. Nature 356: 121–5.
16. Schmitz J, Ohme M, Zischler H (2000) The complete mitochondrial genome of *Tupaia belangeri* and the phylogenetic affiliation of scandentia to other eutherian orders. Mol Biol Evol 17: 1334–1343.
17. Arnason U, Adegoke JA, Bodin K, Born EW, Esa YB, et al. (2002) Mammalian mitogenomic relationships and the root of the eutherian tree. Proc Natl Acad Sci U S A 99: 8151–8156.
18. 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.
19. Kjer KM, Honeycutt RL (2007) Site specific rates of mitochondrial genomes and the phylogeny of eutheria. BMC Evol Biol 7: 8.
20. Kumar S, Filipski AJ, Battistuzzi FU, Pond SLK, Tamura K (2012) Statistics and truth in phylogenomics. Mol Biol Evol 29: 457–472.
21. Avise JC, Walker D, Johns GC (1998) Speciation durations and Pleistocene effects on vertebrate phylogeography. Proc Biol Sci 265: 1707–1712.
22. Curnoe D, Thorne A, Coate JA (2006) Timing and tempo of primate speciation. J Evol Biol 19: 59–65.
23. van Dam JA, Abdul Aziz H, Alvarez Sierra MA, et al. (2006) Long-period astronomical forcing of mammal turnover. Nature 443: 687–691.
24. Maddison WP (1997) Gene trees in species trees. Syst Biol 46: 523–536.
25. Degnan JH, Rosenberg NA (2006) Discordance of species trees with their most likely gene trees. PLoS Genet 2: e68.
26. Seo TK (2008) Calculating bootstrap probabilities of phylogeny using multilocus sequence data. Mol Biol Evol 25: 960–971.
27. Kingman JFC (1982) The coalescent. Stochastic Processes and their Applications 13: 235–248.
28. Liu L, Yu L, Pearl DK, Edwards SV (2009) Estimating species phylogenies using coalescence times among sequences. Syst Biol 58: 468–477.
29. Liu L, Yu L, Edwards SV (2010) A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. BMC Evol Biol 10: 302.
30. Jameson NM, Hou ZC, Sterner KN, Weckle A, Goodman M, et al. (2011) Genomic data reject the hypothesis of a prosimian primate clade. J Hum Evol 61: 295–305.
31. Katoh K, Misawa K, Kuma K, Miyata T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res 30: 3059–3066.
32. Ogden TH, Rosenberg MS (2006) Multiple sequence alignment accuracy and phylogenetic inference. Syst Biol 55: 314–328.
33. Criscuolo A, Gribaldo S (2010) BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. BMC Evol Biol 10: 210.
34. Jobb G, Haeseler A von, Strimmer K (2004) TREEFINDER: a powerful graphical analysis environment for molecular phylogenetics. BMC Evol Biol 4: 18.
35. Lanave C, Preparata G, Sacone C, Serio G (1984) A new method for calculating evolutionary substitution rates. J Mol Evol 20: 86–93.
36. Whelan S, Goldman N (2001) A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol Biol Evol 18: 691–699.
37. Shimodaira H (2002) An approximately unbiased test of phylogenetic tree selection. Syst Biol 51: 492–508.
38. Shimodaira H, Hasegawa M (1999) Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol Biol Evol 16: 1114.
39. Liu L, Yu L (2010) Phybase: an R package for species tree analysis. Bioinformatics 26: 962–963.
40. Rannala B, Yang Z (2003) Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. Genetics 164: 1645–1656.
41. Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, et al. (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. Syst Biol 59: 307–321.
42. Felsenstein J (2005) PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.
43. Huson DH, Bryant D (2006) Application of phylogenetic networks in evolutionary studies. Mol Biol Evol 23: 254–267.
44. Dos Reis M, Inoue J, Hasegawa M, Asher RJ, Donoghue PC, et al. (2012) Phylogenomic datasets provide both precision and accuracy in estimating the timescale of placental mammal phylogeny. Proc Biol Sci 279: 3491–3500.
45. McCormack JE, Faircloth BC, Crawford NG, Gowaty PA, Brumfield RT, et al. (2012) Ultraconserved elements are novel phylogenomic markers that resolve placental mammal phylogeny when combined with species-tree analysis. Genome Res 22: 746–754.
46. Russo CA, Takezaki N, Nei M (1996) Efficiencies of different genes and different tree-building methods in recovering a known vertebrate phylogeny. Mol Biol Evol 13: 525–536.
47. Simmons MP, Carr TG, O'Neill K (2004) Relative character-state space, amount of potential phylogenetic information, and heterogeneity of nucleotide and amino acid characters. Mol Phylogenet Evol 32: 913–926.
48. Townsend JP, López-Giráldez F, Friedman R (2008) The phylogenetic informativeness of nucleotide and amino acid sequences for reconstructing the vertebrate tree. J Mol Evol 67: 437–447.
49. Fong JJ, Fujita MK (2011) Evaluating phylogenetic informativeness and data-type usage for new protein-coding genes across Vertebrata. Mol Phylogenet Evol 61: 300–307.
50. Kubatko LS, Degnan JH (2007) Inconsistency of phylogenetic estimates from concatenated data under coalescence. Syst Biol 56: 17–24.
51. Edwards SV, Liu L, Pearl DK (2007) High-resolution species trees without concatenation. Proc Nat Acad Sci U S A 104: 5936–5941.
52. Huson HD, Rupp R, Scornavacca C (2010) Phylogenetic Networks Concepts, Algorithms and Applications. Cambridge University Press.
53. Churakov G, Kriegs JO, Baertsch R, Zemann A, Brosius J, et al. (2009) Mosaic retroposon insertion patterns in placental mammals. Genome Res 19: 868–875.