

Importance of synonymous substitutions under dense taxon sampling and appropriate modeling in reconstructing the mitogenomic tree of Eutheria

Jiaqi Wu¹, Masami Hasegawa^{1,2}, Yang Zhong^{1,3} and Takahiro Yonezawa^{1*}

¹*School of Life Sciences, Fudan University, SongHu Rd. 2005, Shanghai 200438, China*

²*The Institute of Statistical Mathematics, Midori-cho 10-3, Tachikawa, Tokyo 190-8562, Japan*

³*Institute of Biodiversity Science and Geobiology, Tibet University, JiangSu Rd. 36, Lhasa 850000, China*

(Received 13 August 2014, accepted 9 September 2014)

Although molecular phylogenetics is a strong tool for reconstructing the tree of life, many problems persist due to systematic errors caused by model mis-specifications. Resolving misconstructured trees should lead us to better understand the processes of molecular evolution. Mammalian mitogenomes provide us with a good opportunity in this respect, because the mammalian tree is well established on the basis of multiple nuclear genes, and mitogenome trees are sometimes in conflict with it, for example concerning the positions of tarsiers and colugos. The utility of mitogenomes as a phylogenetic marker is therefore sometimes questioned, and an important problem is whether any method can overcome the misleading phylogenetic signals of mitogenomes. Here we show that the maximum likelihood tree of 463 eutherian mitogenomes reconstructed from nucleotide sequences of protein-encoding genes gives positions of tarsiers and colugos that are consistent with the well-established nuclear tree; this is the first study to obtain a consistent tree with respect to the positions of tarsiers and colugos using mitogenomes. Furthermore, our mitogenome tree of the 463 eutherians is mostly consistent with the nuclear gene tree. Previous mitogenomic studies have been hampered by sparse taxon sampling, and our analysis demonstrates the importance of dense taxon sampling to relieve the misleading phylogenetic signals of mitogenomes. However, because there are many convergent and parallel substitutions in the amino acid sequences, the effect of dense taxon sampling on the accuracy of tree reconstruction seems to be very limited. We further show the importance of using synonymous substitutions with dense taxon sampling as well as with appropriate modeling in recovering the well-established tree from lower to even higher levels of eutherian phylogeny.

Key words: Taxon sampling, synonymous substitution, non-synonymous substitution, colugo, tarsier

INTRODUCTION

To understand biological events during evolution, reliable phylogenetic trees provide us with prerequisite information. Molecular phylogenetics has been playing an important role in clarifying the whole picture of the tree of life. At the same time, however, we have also faced difficulties in reconstructing phylogenetic trees, due not only to a shortage of information but also to systematic errors inherent in phylogenetic methods such as the long branch attraction artifact, the heterogeneity of

nucleotide (or amino acid) composition among lineages (Phillips et al., 2004), heterotachy or covarion shift (Inagaki et al., 2004; Lopez et al., 2002), convergent evolution at the molecular level (Li et al., 2010; Liu et al., 2010; Yonezawa and Hasegawa, 2010) and heterogeneity of the tempo and mode of evolution among different genes (Nishihara et al., 2007). These systematic errors are due to model mis-specification: that is, the models assumed in phylogenetic inference are not realistic. To overcome these issues, many attempts have been made to improve the models (e.g., the nucleotide substitution model, amino acid substitution model, codon substitution model, gamma distribution model and covarion model) as well as to develop new approaches (e.g., removing fast-evolving

Edited by Hidenori Nishihara

* Corresponding author. E-mail: cyclotis@gmail.com

genes or sites, excluding sequences with long branches or extremely biased nucleotide compositions, and partitioning among sites or genes with different tempos and modes of evolution). In such a situation, the effect of dense taxon sampling on phylogenetic reconstruction has been argued extensively on the basis of simulated data and/or real sequencing data (Debry, 2005; Fiala and Sokal, 1985; Huelsenbeck, 1991; Kim, 1998; Poe, 2003; Rosenberg and Kumar, 2001; San Mauro et al., 2012; Townsend and Leuenberger, 2011; Townsend and Lopez-Giraldez, 2010; Zhao et al., 2013; Zwickl and Hillis, 2002). Many studies have indicated that increased taxon sampling can be one of the most practical and feasible approaches for improving phylogenetic estimates (Nabhan and Sarkar, 2012; Zwickl and Hillis, 2002). A positive correlation between increased taxon sampling and the accuracy of tree topology has been demonstrated mainly on the basis of simulated data (Hedtke et al., 2006; Zwickl and Hillis, 2002). A better performance of denser taxon sampling has also been demonstrated with real data (Dimitrov et al., 2012). However, several studies (Rokas and Carroll, 2005; Zhao et al., 2013; Corl and Ellegren, 2013) failed to indicate such a tendency, and questioned the better performance of dense taxon sampling. Since we still do not fully understand the actual processes of molecular evolution, the existing molecular evolutionary models may be regarded as immature, and may be violated when applied to real sequence data. For example, as we discuss later, there are a considerable number of parallel and convergent substitutions in mammalian mitochondrial genomes (mitogenomes). Because current molecular evolutionary models do not take account of these substitutions or of the interaction among loci, they can strongly mislead the inference of tree topologies. Therefore, it is important to investigate the impact of dense taxon sampling on tree inference with real sequence data. Since dense taxon sampling can increase the accuracy of ancestral state reconstructions (Heath et al., 2008), extensive analyses of real sequence data using increased taxon sampling can also provide us with an opportunity to better understand the processes of molecular evolution.

Mammalian mitogenomes provide us with a good example for this purpose. Thanks to extensive analyses in the last two decades, the whole picture of the mammalian tree has mostly been clarified (Arnason et al., 2008; Meredith et al., 2011; Murphy et al., 2001; Nishihara et al., 2006; Prasad et al., 2008). According to these studies, eutherian mammals are divided into three major lineages: Boreotheria (sometimes called Boreoeutheria; human, mouse, shrew, bat, cat, horse, cow, whale, etc.), Afrotheria (elephant, dugong, tenrec, armadillo, etc.) and Xenarthra (armadillo, sloth, anteater, etc.). They are thought to originate from the Laurasian, African and South American continents, respectively. Moreover, Boreotheria is divided into Euarchontoglires and

Laurasiatheria. The former consists of Primates, Scandentia, Dermoptera, Lagomorpha and Rodentia; the latter consists of Eulipotyphla, Chiroptera, Cetartiodactyla, Perissodactyla, Carnivora and Pholidota (Asher and Helgen, 2010). These findings were mainly based on analyses of multiple nuclear gene loci. However, mitogenome trees are sometimes in conflict with the well-established nuclear tree.

For example, in the phylogeny of Euarchontoglires, analyses of nuclear sequences, indels and SINEs established that Dermoptera (colugos) is sister to Primates and that Tarsiidae (tarsiers) belongs to Haplorhini; i.e., tarsiers are closer to simians than to prosimians (Jameson et al., 2011; Janečka et al., 2007; Perelman et al., 2011; Schmitz et al., 2001; Schmitz and Zischler, 2003). On the other hand, many studies using protein-encoding genes of mitogenomes suggested a different scenario: Dermoptera is sister to Simiiformes (simians), rejecting the monophyly of Primates, and Tarsiidae does not form a clade with Simiiformes (Arnason et al., 2002; Chatterjee et al., 2009; da Fonseca et al., 2008; Gibson et al., 2005; Kjer and Honeycutt, 2007; Matsui et al., 2009; Reyes et al., 2004; Xu et al., 2012). Figure 1 shows several trees from previous studies that have used different genetic markers. Trees (a–c) in Fig. 1 are based on mitogenomes, while tree (d) is based on nuclear markers. The conflicting positioning of Dermoptera and Tarsiidae in the mitogenome tree may be due to similar nucleotide compositions of mitogenomes between Dermoptera and Simiiformes, and between Tarsiidae and Strepsirrhini (Gibson et al., 2005; Matsui et al., 2009; Schmitz et al., 2002a).

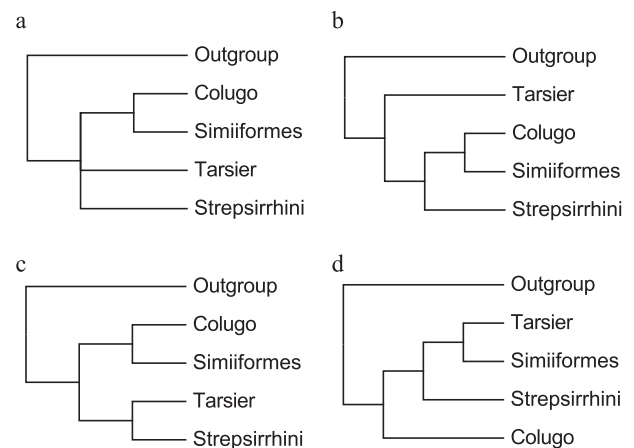


Fig. 1. Previously proposed phylogenetic relationships within the Euarchontoglires based on mitogenome sequences (trees a–c) and nuclear markers (tree d). (tree a) Arnason et al., 2008, 14 Primates. (tree b) Arnason et al., 2002, 9 Primates; Reyes et al., 2004, 10 Primates; Gibson et al., 2005, 12 Primates; Matsui et al., 2009, 24 Primates; Chatterjee et al., 2009, 19 Primates. (tree c) Kjer and Honeycutt, 2007, 11 Primates (tree d) Janečka et al., 2007, nuclear genes; Jameson et al., 2011, nuclear genes; Schmitz et al., 2001, SINEs.

In this report, by focusing on the phylogenetic positions of Dermoptera and Tarsiidae, as well as on other problematic taxa such as Pholidota and Cetacea, as case studies, we show that dense taxon sampling can relieve the misleading phylogenetic signals of mitogenomes when the maximum likelihood (ML) method is used with an appropriate model. Furthermore, we demonstrate a better performance of codon-based partition analyses than amino acid analyses, and show the importance of synonymous substitutions during denser taxon sampling. Since natural selection operates directly on amino acid sequences, there may be many parallel or convergent substitutions at the amino acid level due to strong functional constraints. In such a case, no matter how dense the taxon sampling, there would be no remarkable increase in the accuracy of tree reconstruction. In contrast, synonymous substitutions can be regarded as nearly neutral, and therefore signatures of phylogenetic traits may be recorded more explicitly. However, the effect of multiple substitutions could present a serious problem in using synonymous substitutions to clarify higher taxonomic relationships because of their high evolutionary rate. We show that, if dense taxon sampling as well as an appropriate model are available, information from synonymous substitutions is useful even in the study of higher taxonomic relationships.

MATERIALS AND METHODS

Sequence data In this study, only protein-encoding genes in the mitogenomes were analyzed. Hudelot et al. (2003) instead analyzed mitogenome tRNAs and rRNAs. Despite sparse taxon sampling, their Bayesian tree of 69 mammals, obtained using both paired and unpaired regions of mitochondrial rRNA and tRNA genes, was mostly consistent with the well-established nuclear tree including the sister group position of Dermoptera with Primates, although the sister group relationship of Tarsiidae with Strepsirrhini was preferred. Matsui et al. (2009) did not include Dermoptera in their analyses, but their rRNA analysis of mitochondria supported the sister group relationship of Tarsiidae with Simiiformes, although weakly. Thus, it seems that the misleading phylogenetic signal of the mitogenome is confined to the protein-encoding genes.

Mitogenome data from 463 eutherians were retrieved from GenBank, and species names and accession numbers are listed in Supplementary Table S1. All protein-encoding genes in the mitogenome except *nd6* were aligned separately by the MUSCLE program (Edgar, 2004) and were concatenated for the analysis. In total, 12 genes were used and the total length of the alignment was 10,704 bp.

Phylogenetic inference A phylogenetic tree was inferred

by the ML method (Felsenstein, 1981) with RAxML ver. 7.04 (Stamatakis, 2006; Stamatakis et al., 2008) for the heuristic tree search. The GTR + Γ + I model (Hasegawa et al., 1985; Rodriguez et al., 1990; Yang, 1996) and the mtMAM + F + Γ + I model (Hasegawa et al., 1985; Yang, 1996; Yang et al., 1998) were applied for nucleotide and amino acid sequence data, respectively. To evaluate the confidence of the internal branches, bootstrap probabilities (BPs) were estimated by the standard bootstrap method and the rapid bootstrap method (Stamatakis et al., 2008), with 1,000 replications for the nucleotide sequences and 500 replications for the amino acid sequences. In the nucleotide ML analyses with RAxML, partition was applied to distinguish among the three codon positions (ML_CodonPartition) in our main analysis, but an analysis without partition (ML_NoPartition) was also carried out for comparison. Furthermore, ML analyses using only third codon positions (ML_3), using 1st + 2nd positions (ML_12) and using translated amino acid sequences (ML_aa) were also carried out. When partition was applied, the parameters of the model including the branch lengths were separately estimated for each codon position.

The ancestral states of internal nodes were inferred by the CODEML program in PAML ver. 4 (Yang, 2007) using the empirical Bayesian approach with the mtMAM + F + Γ model (Yang, 1996; Yang et al., 1998). Parallel and convergent substitution sites were identified by comparing the estimated ancestral and the extant sequences.

Evaluation of the effect of dense taxon sampling

The effect of dense taxon sampling on the accuracy of tree inference was evaluated in the following way. First, to reduce the computational burden, one representative species was selected from each genus, yielding 273 species in total. The ML tree of the 273 species using GTR + I + Γ + codon partition model was inferred and was consistent with the ML_CodonPartition tree based on 463 species (Supplementary Figs. S1 and S2 for the 463 and 273 species, respectively). Next, in addition to the five sequences of tarsier, colugo, pangolin, sperm whale and pygmy sperm whale, we randomly sampled 50, 100, 150, 200, and 240 sequences from the remaining 268 sequences. These datasets contain one tarsier and one colugo sequence, and are named T1C1. Furthermore, to take advantage of the deep branchings between the two species of tarsiers and among individuals of colugo, two extra colugo sequences and one extra tarsier sequence were added to each sample, giving two tarsier and three colugo sequences in total; this data set is named T2C3. The ML trees were inferred using RAxML with the GTR + I + Γ model with ML_CodonPartition for each sample. These procedures were repeated 10 times.

The effect of taxon sampling densities of the in-group and out-group was evaluated, taking Euarchontoglires as

an example. The dataset includes 64 sequences of Primates and Dermoptera as the in-group, and 59 sequences of Rodentia, Lagomorpha and Scandentia as an out-group. A ML tree of all 123 Euarchontoglires sequences was inferred using the GTR + I + Γ + codon partition model (Supplementary Fig. S3) and found to be basically consistent with the ML_CodonPartition tree using the 463-species data set, except for several nodes,

such as those concerning the phylogenetic position of *Tupaia* and the relationship between *Lepilemur* and *Propithecus*. In addition to the two tarsier and three colugo sequences from the in-group and five tree shrew sequences from the out-group, we randomly selected 10, 20, 30, 40 or 50 sequences from the in-group other than tarsiers and colugos and 10, 20, 30 or 40 sequences from the out-group other than tree shrews. For each in-group

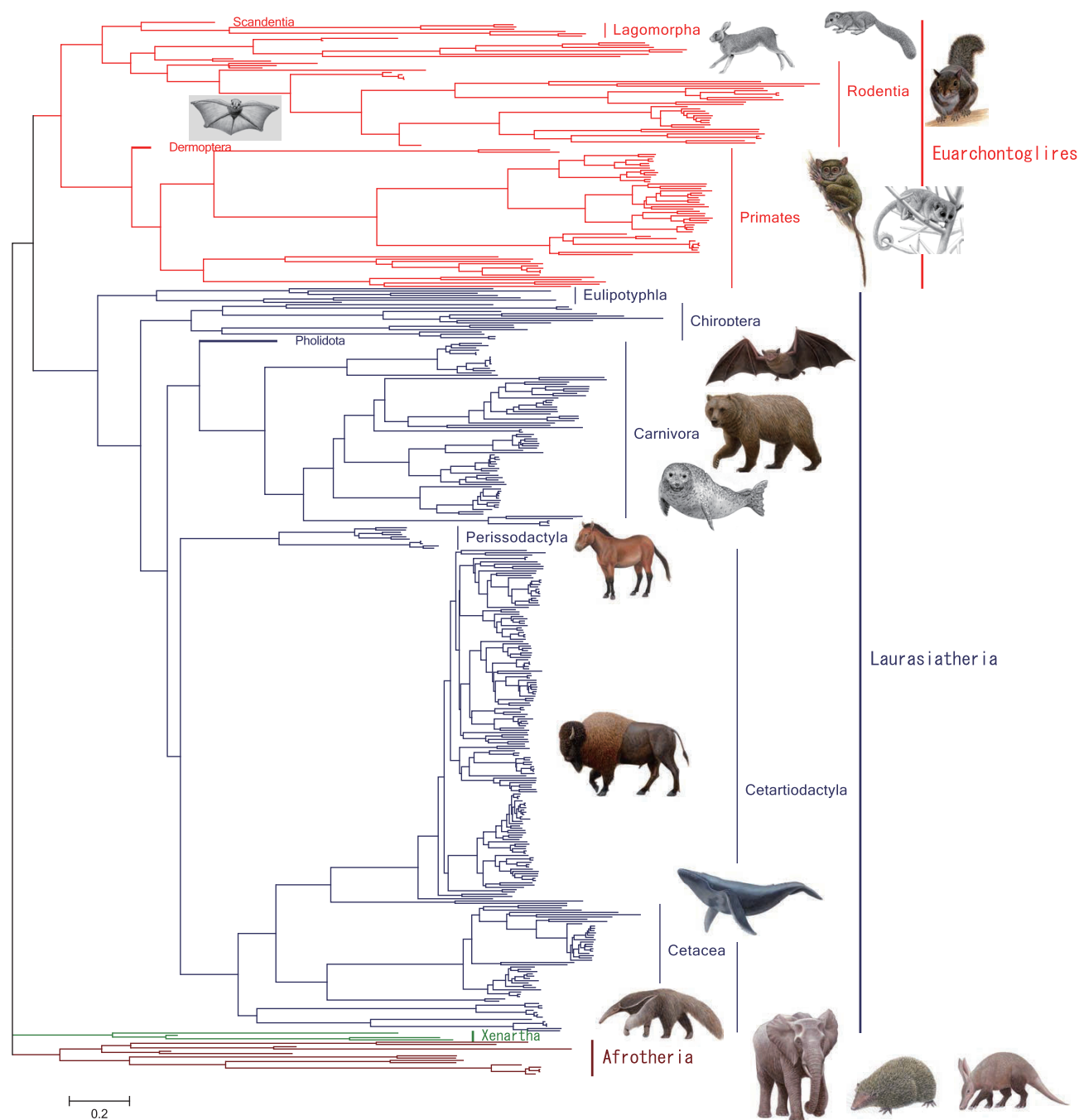


Fig. 2. The ML_CodonPartition tree of 463 eutherian mitogenomes. The tree was estimated by RAxML ver. 7.04 (Stamatakis, 2006; Stamatakis et al., 2008) using the GTR + Γ + I model (Hasegawa et al., 1985; Rodriguez et al., 1990; Yang, 1996) with partition among three codon positions. The scale bar indicates 0.2 substitutions per site. The pictures were drawn by Utako Kikutani.

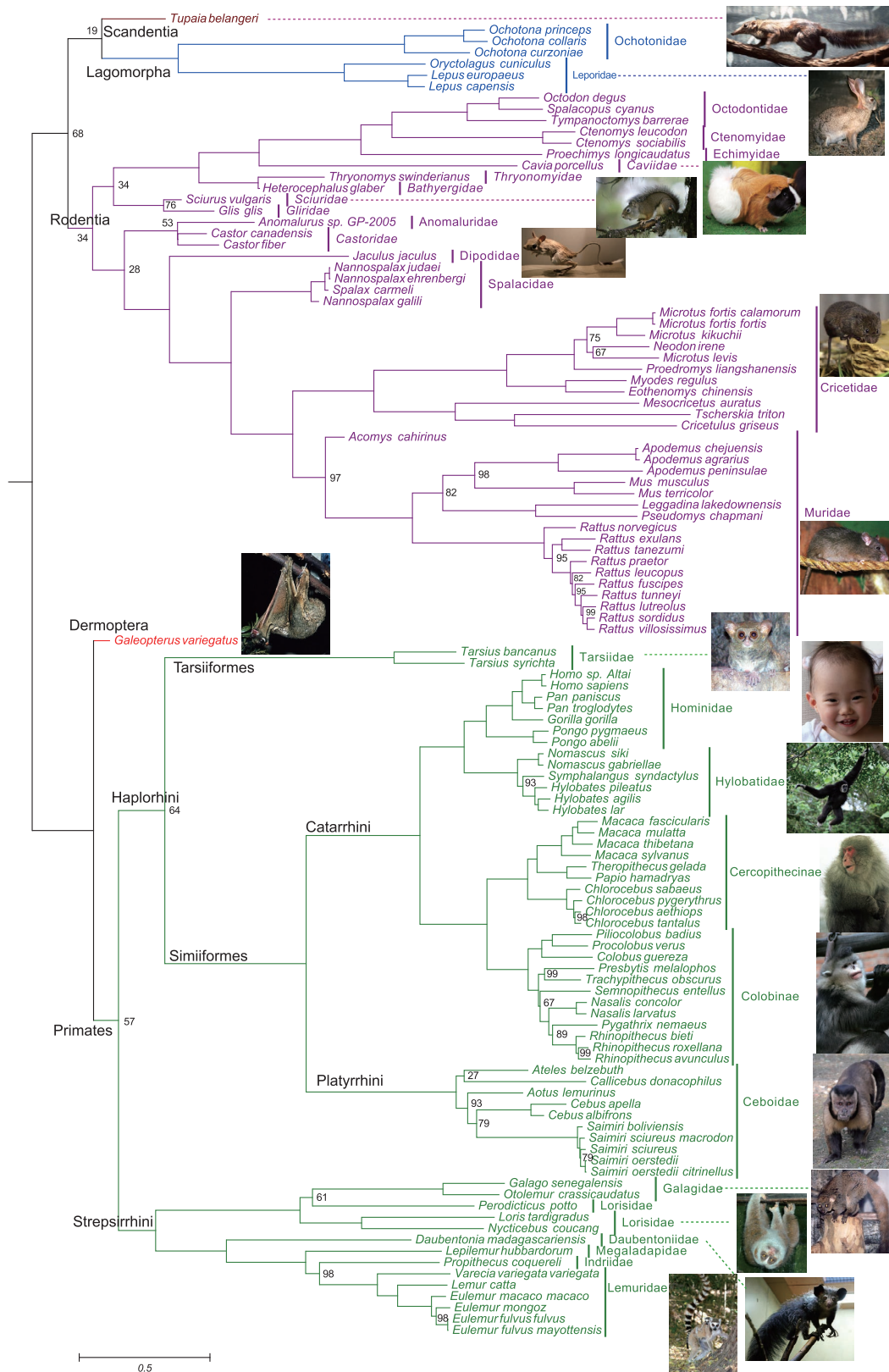


Fig. 3. The Euarchontoglires part of the ML_CodonPartition tree. Numbers at nodes indicate bootstrap probabilities (BPs) in % (nodes without BP indications are 100% supported). The scale bar indicates 0.5 substitutions per site. Photos show representatives of the families, and do not necessarily correspond to the indicated species. The photos were all taken by M. H. except for *Galeopterus variegatus* (courtesy of Teruyuki Komiya).

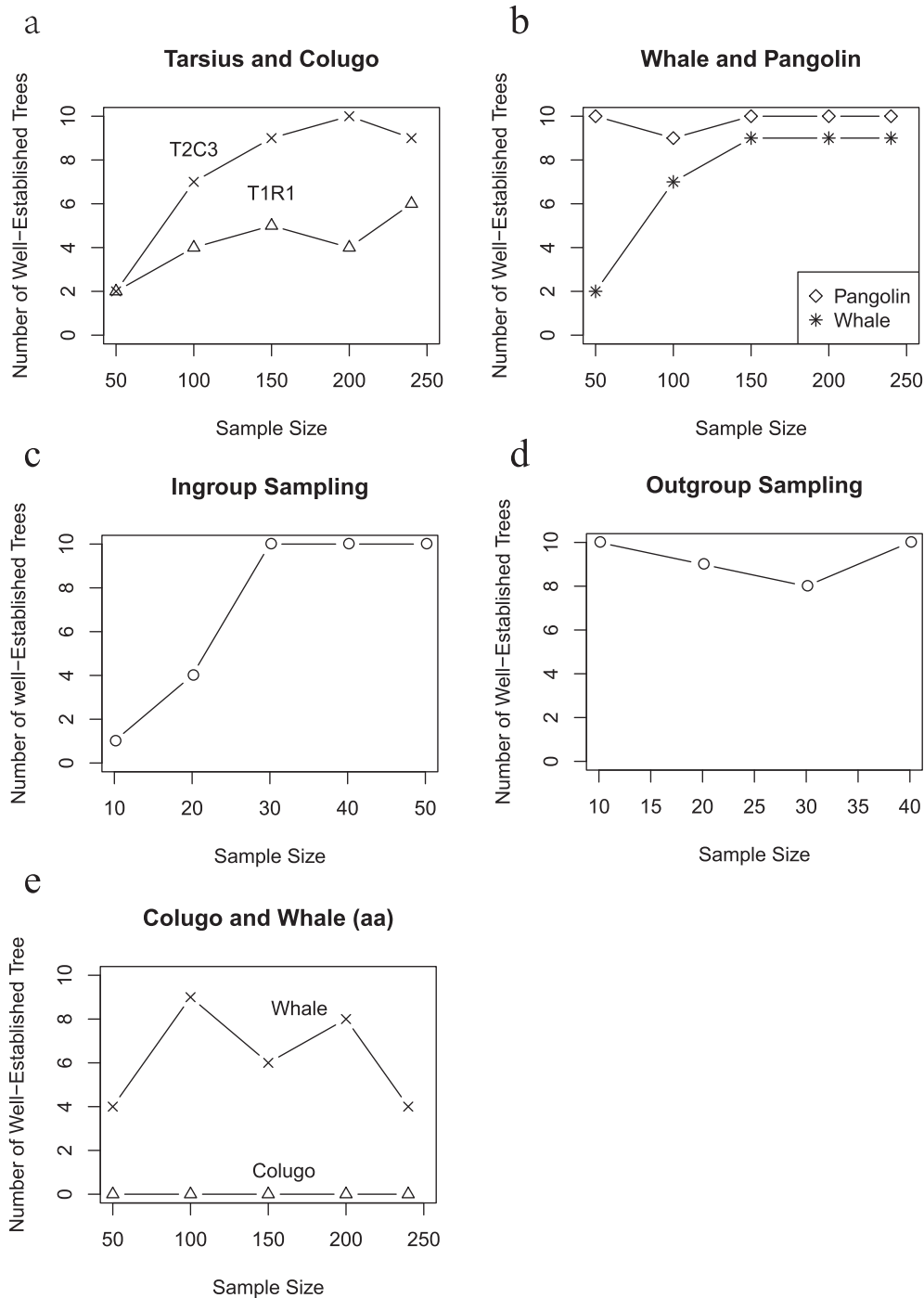


Fig. 4. Effect of denser taxon sampling on the accuracy of inferring Eutheria and Euarchontoglires trees. Horizontal axes indicate sample size (number of sequences) and vertical axes indicate the number of times well-established trees were recovered among 10 trials. (a) Phylogenetic positions of the tarsier and colugo as inferred from +5 test data (one species from one genus) and +8 test data (one species from one genus, except for two species of *Tarsius* and three individuals of colugo). (b) Phylogenetic positions of the sperm whale and pangolin as inferred from T1C1. For detailed procedures, see the main text. (c) Phylogenetic positions of the tarsier and colugo as inferred from in-group sampling test data. (d) Phylogenetic positions of the tarsier and colugo as inferred from out-group sampling test data. (e) Phylogenetic positions of the colugo and the sperm whale inferred from out-group sampling test data on the basis of amino acid (aa) sequence data. Heuristic searches of the ML tree by the RAxML program were carried out for (a) to (d). However, since inferring the ML tree from the amino acid sequences was computationally too expensive, two topologies were compared by PAML for (e) for the primate+colugo, and whale, respectively. For the Primates, (((Simiiformes, *Tarsius*), Strepsirrhini), Colugo) and (((Simiiformes, Colugo), *Tarsius*), Strepsirrhini) were compared. For the whales, (((Delphinoidea + Inoidea + Lipotoidea, Ziphiioidea + Platanistoidea), Physeteroidea), Mysticeti) and (((Mysticeti, Ziphiioidea + Platanistoidea), Physeteroidea), Delphinoidea + Inoidea + Lipotoidea) were compared.

test, all taxa of the out-group were used, while for each out-group test, all in-group taxa were used. A ML tree was inferred using RAxML with the GTR + I + Γ + codon partition model. The procedures were repeated 10 times, and numbers of trees displaying the well-established positions of colugo and tarsier in the 10 trials were counted.

RESULTS

Phylogenetic relationships and effect of dense taxon sampling The ML tree of 463 taxa with the partition among different codon positions (ML_CodonPartition) is summarized in Fig. 2, and the detailed tree is shown in Supplementary Figs. S1 (463 species) and S2 (273 species: one representative species selected from each genus). Remarkably, the tree is mostly consistent with the well-established tree derived from analyses of nuclear DNA. The Euarchontoglires part of the tree is given in Fig. 3 (the detailed tree is shown in Supplementary Fig. S3), in which Dermoptera is sister to Primates, and Tarsiidae is sister to Simiiformes, consistent with the nuclear DNA analyses (Jameson et al., 2011; Janečka et al., 2007; Perelman et al., 2011; Schmitz et al., 2001; Schmitz and Zischler, 2003). Previous mitogenome analyses all failed to obtain such relationships (Arnason et al., 2002; Chatterjee et al., 2009; da Fonseca et al., 2008; Gibson et al., 2005; Kjer and Honeycutt, 2007; Matsui et al., 2009; Reyes et al., 2004; Xu et al., 2012), and our study is the first to obtain them using protein-encoding genes of mitogenomes. As will be discussed later, the phylogenetic positions of the pangolins (Pholidota), which is sister to Carnivora, and the sperm whales (Physeteridae, Cetacea), which represent the basal lineage of Odontoceti, are consistent with the well-established relationships already shown by previous studies (Arnason et al., 2008; Hassanin et al., 2012). Concerning the positions of tarsier and colugo, the main difference between our and previous mitogenome analyses is denser taxon sampling in

our analysis. Therefore, we think the improvement of our result is mainly due to this denser taxon sampling. To test this hypothesis, we evaluated the effect of dense taxon sampling (Fig. 4a) on BPs of each node in each trial (Supplementary Table S2). For the positions of tarsier and colugo, as taxon sample size increases, the probability of obtaining the well-established tree also increases, especially in the T2C3 (two tarsier and three colugo sequences) data set.

We also evaluated the effect of dense taxon sampling on the phylogenetic positions of pangolins (*Manis*) and sperm whales. Figure 4b shows the number of trees having the well-established positions of sperm whale and pangolin. The number of trees with the well-established position of sperm whale increases when the taxon sample size becomes larger, but the increase of taxon sample size has little effect on the position of pangolin.

Comparing different methods and datasets The results obtained from various models and methods are compared in Table 1 with respect to the four phylogenetic issues (the tree topologies by each model and method were indicated in Supplementary Figs. S4-S10). Contrary to the high performance of ML_CodonPartition with nucleotide sequences for tree reconstruction, the performance of the amino acid model was poor, and the amino acid tree (ML_aa; Supplementary Fig. S7) gave a sister group relationship of Dermoptera to Simiiformes as in previous mitogenome analyses, and a sister group relationship of Tarsiidae to the clade formed by Dermoptera and Simiiformes. The position of the sperm whale was also erroneous.

Interestingly, the ML analysis using only third codon positions (ML_3) performs unexpectedly well in resolving relatively shallow branchings such as the phylogenetic relationships within orders (the Primates and Cetacea parts of ML_3 are shown in Figs. 5 and 6, respectively); ML_3 gave sister group relationships between Tarsiidae

Table 1. Phylogenetic positions estimated by various methods and models

	tarsier Tarsiiformes	colugo Dermoptera	pangolin Pholidota	sperm whale Physeteridae
ML_CodonPartition	*Sister to Simiiformes (64)	*Sister to Primates (57)	*Sister to Carnivora (79)	*Basal Odontoceti (64)
ML_NoPartition	*Sister to Simiiformes (59)	Sister to Haplorhini (47)	*Sister to Carnivora (66)	*Basal Odontoceti (62)
ML_3	*Sister to Simiiformes (88)	*Sister to Primates (50)	Sister to Cetartiodactyla (37)	*Basal Odontoceti (80)
ML_12	Sister to Dermoptera + Simiiformes (40)	Sister to Simiiformes (48)	*Sister to Carnivora (80)	*Basal Odontoceti (34)
ML_aa	Sister to Dermoptera + Simiiformes (65)	Sister to Simiiformes (94)	*Sister to Carnivora (81)	Sister to <i>Platanista</i> + Ziphiidae + Mysticeti (41)

Asterisks (*) indicate agreement with putatively correct position established by nuclear gene sequences, indels and SINEs. Figures in parentheses indicate BPs in %.

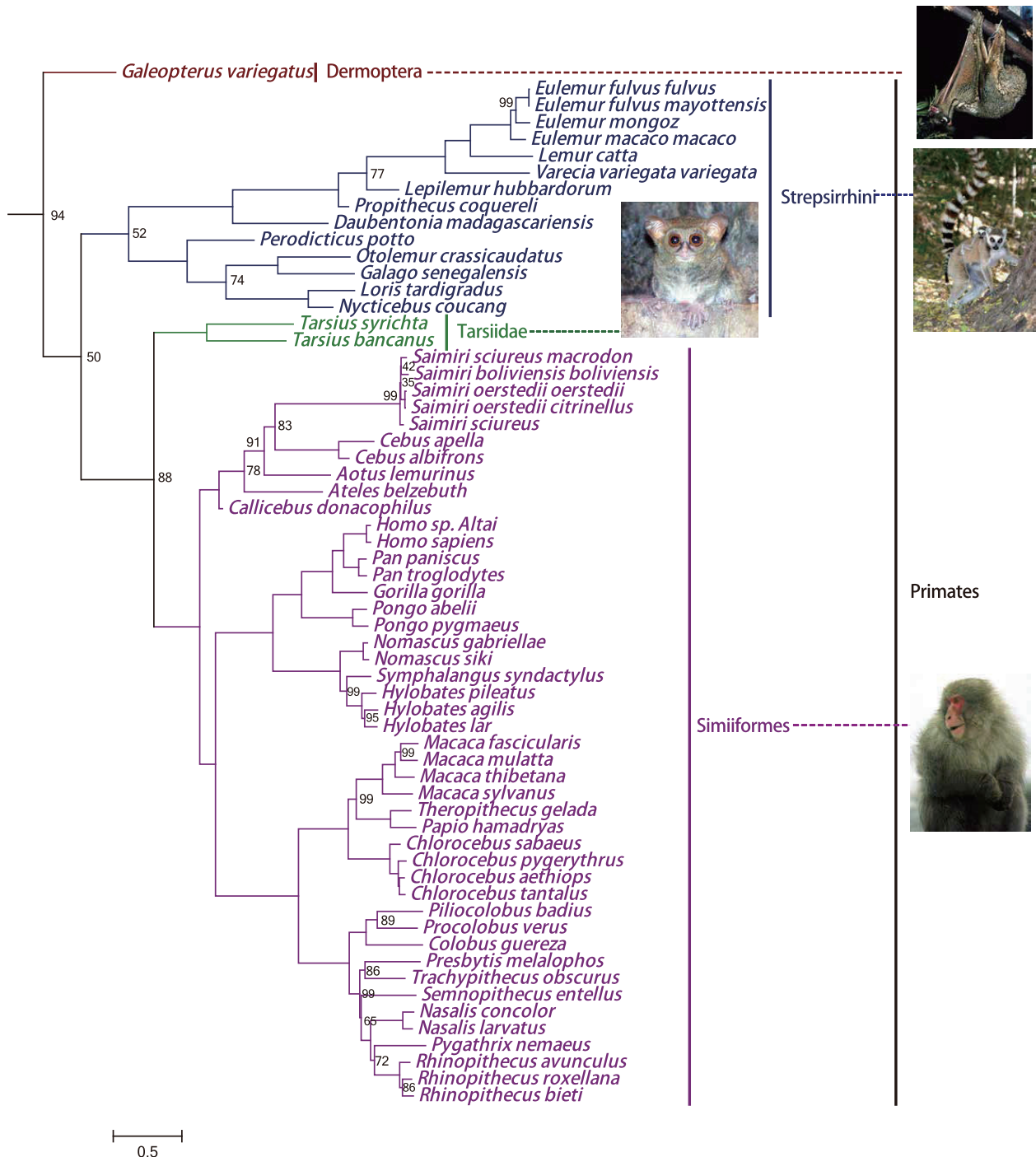


Fig. 5. The Primates part of the ML_3 tree. Numbers at nodes indicate % BPs (nodes without BP indications are 100% supported). The scale bar indicates 0.5 substitutions per site. Photos show representatives of the families, and do not necessarily correspond to the indicated species. The photos were all taken by M. H.

and Simiiformes (88% BP) and between Dermoptera and the monophyletic Primates (50% BP) (Table 1), consistent with the ML_CodonPartition analysis, while the other analyses (ML_12, ML_aa; see MATERIALS AND METHODS for details) failed to recover these relationships. Since ML_3 uses only the third codon positions, which

evolve faster than other positions, it gave misleading relationships for several deep branchings as expected, but still gave agreement with the well-established relationships for most shallow branchings, sometimes showing even better performance than ML_aa and ML_12. ML NoPartition gave a result consistent with

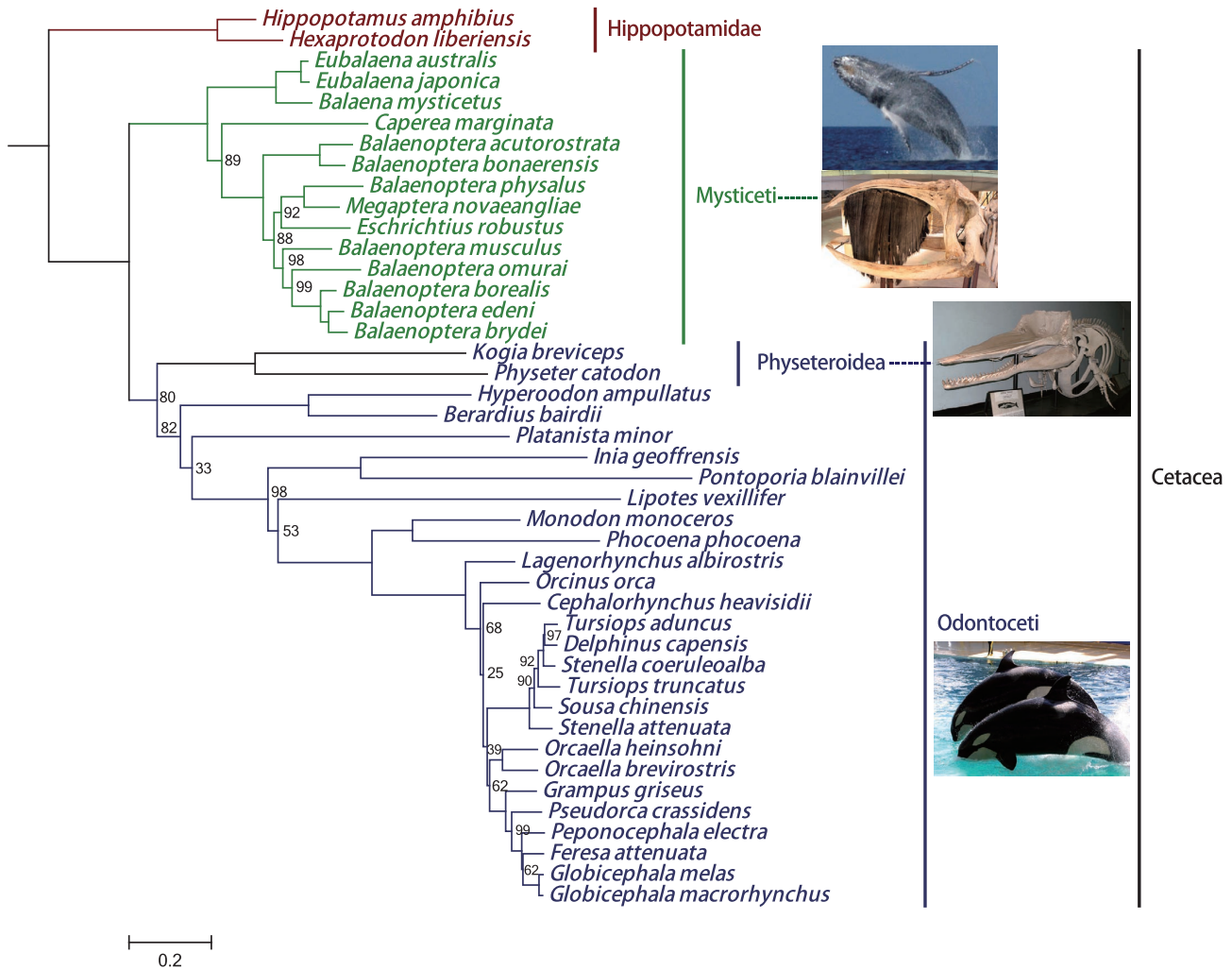


Fig. 6. The Cetacea part of the ML_3 tree. Numbers at nodes indicate % BPs (nodes without BP indications are 100% supported). The scale bar indicates 0.2 substitutions per site. Photos show representatives of the families, and do not necessarily correspond to the indicated species. The photos were all taken by M. H.

ML_CodonPartition concerning the position of Tarsiiformes, but the position of Dermoptera was erroneous, indicating that the appropriate treatment of third codon positions by partitioning is essential for recovering the well-established tree. The ML_CodonPartition model is much better than the ML_NoPartition model in terms of a model selection criterion such as AIC (Akaike, 1973): 2,614,915.4 *vs.* 2,684,863.3 for ML_CodonPartition and ML_NoPartition, respectively.

DISCUSSION

Impact of dense taxon sampling and its pitfalls

Some authors have argued that increased taxon sampling could greatly reduce phylogenetic errors (Goloboff et al., 2009; Heath et al., 2008; Hillis, 1996; Zwickl and Hillis, 2002), while others have concluded that it has little benefit for phylogenetic inference compared to increased

sequence length (Kim, 1996; Rosenberg and Kumar, 2001; Zhao et al., 2013). Our analysis convincingly showed that dense taxon sampling is important (Heath et al., 2008; Hillis, 1996; Zwickl and Hillis, 2002). Previously, it was generally considered that adding more taxa to a phylogenetic inference creates a more complex problem of tree topology searching among a huge number of trees, and Kim (1996) argued that increasing the number of characters is a more practical way to improve phylogenetic accuracy. However, as Nishihara et al. (2007) demonstrated, systematic errors present a serious problem in such a large dataset with respect to the number of characters, but with sparse taxon sampling (see also the Supplementary text, "Bias of phylogenetic inference and importance of taxon sampling").

Recent improvements in phylogenetic algorithms and in computational power have removed many constraints on analyzing large, thoroughly sampled datasets (Heath

et al., 2008).

The ML method takes account of the ancestral states of the nodes, and therefore dense taxon sampling should be of benefit in overcoming misleading phylogenetic signals such as the long-branch attraction (LBA) artifact (Felsenstein, 1978; Kück et al., 2012). The increase of taxon sampling could cut long branches and helps to correctly evaluate multiple substitutions. The ML method has an obvious advantage over the distance method, which does not take account of the ancestral states. The parsimony method does take them into account, and the accuracy of this method is reportedly increased by denser taxon sampling (Debry, 2005). However, this method ignores multiple substitutions in a site and is therefore prone to be more seriously affected by the LBA than is the ML (see also Supplementary text). Although the ML method is thus superior to the alternative methods, it is extremely computationally demanding, and was therefore previously impractical for analyzing a dataset with dense taxon sampling. Due to the development of computer technology and of powerful algorithms (Stamatakis, 2006; Stamatakis et al., 2008), the ML method has now become practical even with a large number of taxa.

According to the taxon sampling test, when the number of samples increases, the probability of obtaining the well-established tree also increases. This result strongly suggests that increased taxon sampling can reduce systematic error. Noticeably, in most of the previous studies using mitogenomes (Arnason et al., 2002; Gibson et al., 2005; Kjer and Honeycutt, 2007; Reyes et al., 2004), only around ten species of Primates were included. According to Fig. 4d and Supplementary Tables S3 and S4, when the in-group sample size is around 10, the probability of recovering the well-established tree is around 0.1. This may explain why the previous studies failed to obtain the well-established relationship. By increasing the number of in-group sequences, the probability of obtaining the well-established positions of tarsier and colugo increased, while increasing the number of out-group species did not affect the probability of obtaining the well-established tree because it was already obtained with only 10 out-group species. Although Matsui et al. (2009) and Chatterjee et al. (2009) used more (around 20) Primate species, they still failed to obtain the well-established relationships. When a phylogenetic tree was inferred with the same out-group as used by Matsui et al. (2009) (tree shrew, mouse, and rat) and all 64 species of the in-group, it failed to match the well-established tree (Supplementary Fig. S11). However, if distantly related species from the mouse and rat clade among the Glires were added (e.g., two extra species of rabbit), the well-established tree was obtained, although the bootstrap support was low (Supplementary Fig. S12). Thus, the selection of out-group species is also important.

It is also notable that although denser taxon sampling

can increase the probability of obtaining the well-established tree, it has little effect on the level of nodal support. Despite the dense taxon sampling, misleading signals caused by several factors as mentioned above still apparently exist, and therefore nodal support is low in many problematic taxa (see also the Supplementary text “Comparison of the standard and rapid bootstrap methods” as well as Supplementary Fig. S13).

In contrast to the usefulness of dense taxon sampling for the nucleotide sequence data, the tree topology on the basis of the amino acid sequence data was highly inaccurate even under dense taxon sampling (Fig. 4e). This may be caused by the information loss of synonymous substitutions, and the well-established tree could not be recovered simply due to the lack of informative sites. However, the strong nodal support for the colugo/Simiiformes clade (94%) implies the existence of signals that strongly mislead the tree topology, probably arising from convergent or parallel substitutions at the amino acid level. We reconstructed the ancestral states of amino acids using the ML_CodonPartition tree topology and found several parallel amino acid substitutions between the colugo branch and the ancestral branches within Haplorhini, and between the tarsier branch and the ancestral branches of Strepsirrhini, as well as between the ancestral baleen whale and other toothed whales (Table 2). Current models of phylogenetics do not assume parallel and convergent evolution. Therefore, in such a case, it is possible that however dense the taxon sampling was, the well-established tree might not be achievable.

Importance of synonymous substitutions Another finding of our analyses was the importance of synonymous substitutions in inferring the eutherian mitogenome tree. The substitution rate of the mitogenome is very high, and the number of transitional nucleotide differences in third codon positions (always synonymous in the eutherian mitochondrial code) is almost the same in human/chimpanzee and human/orangutan comparisons (732/751 and 856/864 in our dataset) (Hasegawa et al., 1985), although the orangutan divergence from human is roughly twice as old as the chimpanzee divergence (dos Reis et al., 2012). This means that a substantial number of multiple substitutions have accumulated between human and orangutan. For this reason, it was presumed that phylogenetic analyses might be hampered by synonymous noise, and third codon positions were often discarded from the analyses (Reyes et al., 2004) or amino acid sequences were used instead (Arnason et al., 2002; Cao et al., 2000; Xu et al., 2012).

Even in this situation, our analyses showed that synonymous substitutions substantially improve phylogenetic inference if the taxon sampling is sufficiently dense, consistent with a suggestion from previous work (Seo and

Table 2. Numbers of convergent and parallel substitution sites between branches

(a) Primates and Dermoptera			
Ancestral branch leading to		Number of convergent sites	Number of parallel substitutions
Dermoptera	Haplorhini	1	7
Dermoptera	Simiiformes	0	45
Dermoptera	Tarsiiformes	0	37
Dermoptera	Platyrrhini	0	32
Dermoptera	Catarrhini+Hominoidea		35
Dermoptera	Catarrhini	6	13
Dermoptera	Hominoidea	0	11
	Total	8	180
Tarsiiformes	Strepsirrhini	0	11
Tarsiiformes	Lemuroidea	0	2
Tarsiiformes	Lorisoidea	0	23
	Total	0	36
(b) Cetacea			
Ancestral branch leading to		Number of convergent sites	Number of parallel substitutions
Mysticeti	Physeteroidea	0	2
Mysticeti	Ziphioidae+Platanistoidea	1	1
Mysticeti	Platanistoidea	0	5
Mysticeti	Delphinoidea+Inoidea+Lipotoidea	1	1
	Total	2	9
Physeteroidea	Platanistoidea	0	12
Physeteroidea	Delphinoidea+Inoidea+Lipotoidea	0	3
	Total	0	15
Platanistoidea	Delphinoidea+Inoidea+Lipotoidea	0	10
	Total	0	10

Kishino, 2008; Simmons et al., 2006).

It should be noted that we are not directly comparing third codon positions of two arbitrary sequences which diverged 50–80 million years ago. In such a comparison, it is difficult to evaluate correctly the number of substitutions during the long time period because of abundant multiple substitutions. The ML method we are using cuts long branches short and can correctly infer the ancestral states of intermediate splitting when taxon sampling becomes dense. A possible concern is that only one species from Dermoptera is used in Fig. 3. In this case, an extremely slow evolutionary rate in colugos, represented by a short branch in Fig. 3, might have contributed to recovery of the phylogenetic signals contained in the third codon positions. As mentioned above, there are parallel and convergent substitutions at the amino acid level. In addition to parallel and convergent evolution, there appear to be other unknown factors that mislead tree inference, because, even when such amino acid sites were excluded from the analysis, an erroneous tree was still

reconstructed (data not shown).

Parallel and convergent substitutions of amino acids are probably due to functional constraints on protein structure. In contrast, synonymous substitutions can be regarded as nearly neutral, and signatures of the phylogenetic traits may thus be recorded more explicitly. Misleading signals at the amino acid level may be relieved by the inclusion of third codon positions in the analysis. This suggests that selection of nearly neutral sites was also preferred as well as the phylogenetic representativeness of the taxon sampling.

Further perspectives This study has shown an unexpectedly high resolution of deep branchings as inferred from densely sampled mitogenome data, indicating the importance of dense taxon sampling as well as the use of appropriate models. However, several issues remain to be resolved.

First, the position of Scandentia (tree shrews) is problematic (Fig. 2). In this study, Scandentia is represented

only by *Tupaia belangeri*, and it was placed as a sister group to Lagomorpha. With a BP of only 19%, this placement is very uncertain from the dataset we analyzed. The placement of Scandentia with Dermoptera and Primates in the grouping Euarchonta has been supported by nuclear evidence including retroposon-insertion analysis (Kreigs et al., 2007; McCormack et al., 2012; Song et al., 2012) as well as genome-scale analyses (Fan et al., 2013; however, their data did not include Dermoptera). On the other hand, the relationships among Scandentia, Dermoptera and Primates within the monophyletic Euarchonta were left as an unresolved trichotomy (Janečka et al., 2007; Nie et al., 2008; Martin, 2008). We extensively examined the position of Scandentia within Euarchontoglires on the basis of mitgenome data (Supplementary text). As a result, the relationships of “the sister to Lagomorpha (the ML tree)” and “the most basal among Euarchontoglires (the second best tree)” were indistinguishable in terms of the likelihood score. However, as mentioned above, these topologies are unlikely. On the other hand, the tree topologies that support the “Euarchonta hypothesis” were marginally rejected, probably because of the sparse sampling of Scandentia and Dermoptera in this study. Although our study included only one species each of Scandentia and Dermoptera, Scandentia is a diversified order that includes distant groups from Tupaia, such as Ptilocercus and Dendrogale (Roberts et al., 2011), and Dermoptera has another genus, Cynocephalus, that is distant from Galeopithecus (Janečka et al., 2007). Inclusion of these additional taxa in the mitgenome analysis should contribute to resolving the phylogenetic relationships among orders in Euarchonta.

Second, the rooting problem of the eutherian mammals still remains. When the marsupial mammals and monotremes were used as out-groups, the Rodentia was placed as the most basal group of eutherian mammals, and moreover was recognized as a paraphyletic group (Supplementary Fig. S14). The monophyly of Rodentia and the phylogenetic position of this order are well established on the basis of nuclear evidence including retroposon-insertion analysis and genome scale analysis (Murphy et al., 2001; Nishihara et al., 2006, 2007; Fan et al., 2013). Therefore, this topology of the rooted tree on the basis of the mitgenome is unlikely.

Third, the position of Erinaceidae (hedgehog and moonrat) still remains unsettled. We did not include Erinaceidae in our analysis, because the position of Erinaceidae has not been settled, as noted in the Supplementary text. Although Erinaceidae was placed as the most basal lineage among eutherian mammals on the basis of the mitgenome analysis (Supplementary text), this phylogenetic position is unlikely from nuclear evidence (Murphy et al., 2001; Douady et al., 2002; Hallström et al., 2011; Zhou et al., 2012). The aberrant

placement of Erinaceidae was attributed to their distinct nucleotide compositions and rapid rate of mitgenome evolution, particularly in hedgehogs (Lin et al., 2002; Nikaido et al., 2003; Gibson et al., 2005). Applying three-state (AGY) and four-state (AGTC) general time-reversible models for the first and second codon positions improved the position of Erinaceidae, placing them at the base of Laurasiatheria and making Eulipotyphla paraphyletic (Gibson et al., 2005). Even if a eutherian unrooted tree was inferred, Erinaceidae was highly nested within rodents (data not shown).

The second and third problems are caused by the LBA artifact, and also by the convergently acquired nucleotide and/or amino acid composition similarity (Yonezawa and Hasegawa, 2010) between rodents and marsupials (data not shown). To solve them, we tried several strategies including (1) application of more realistic models such as the UNREST model that does not assume “time reversibility” (Yang, 1994), a modified version of the HKY85 model (Yang, 2007) that allows heterogeneity of the nucleotide composition bias among lineages, and a more accurate amino acid substitution model derived from a huge amount of mammalian mitgenome data (Wu et al., in preparation), and (2) exclusion of fast-evolving sites and heterotachy sites that can cause the LBA. However, none of these approaches resolved the rooting problem of eutherian mammals or the phylogenetic position of hedgehogs.

Since the overall tree topology of eutherian mammals has been established by nuclear gene analyses, we can conclude that the relationships inferred from mitgenomes, such as the early branching of hedgehogs in Eutheria and the paraphyly of Rodentia, were caused by an artifact of phylogenetic inference. Therefore, further studies using mitgenomes to refine the well-established tree are needed to reveal more accurately the mechanism of mitgenome evolution and to improve the models used in phylogenetic analyses.

CONCLUSIONS

By taking advantage of the rapid accumulation of mitgenome sequences in the database and of the development of powerful software for phylogenetic inference based on ML (Stamatakis, 2006; Stamatakis et al., 2008), we estimated the ML tree of 463 eutherian species from their mitgenome sequences. With such a densely sampled data set, we could obtain relationships that were consistent with the well-established nuclear DNA tree, whereas all previous mitgenome analyses failed to obtain such relationships. This is probably due to the denser taxon sampling used in our analysis, which thus emphasizes the importance of dense taxon sampling in phylogenetic inference. Furthermore, we showed the importance of using synonymous substitutions with an

appropriate model for inferring the well-established tree up to the subordinal level. However, we found that there are many parallel and convergent substitutions in the amino acid sequences, and, in such a case, in contrast to the high performance of dense taxon sampling concerning the accuracy of tree reconstruction at the nucleotide sequence level, it has almost no effect at the amino acid sequence level. Another point raised by our analysis is the importance of using an appropriate model in the ML analysis. ML analyses with poor models and analyses using alternative methods such as MP and NJ could not take full advantage of dense taxon sampling, and failed to obtain relationships that were consistent with the well-established nuclear DNA tree. Reconstructing the tree of life is prerequisite to understanding how Earth's biodiversity evolved, and it is important to choose an adequate phylogenetic marker for this purpose. Although mitogenomes sometimes give a misleading phylogeny, we have shown that this shortcoming can be partially overcome by increasing the density of taxon sampling in conjunction with adequate phylogenetic models and methods, and we emphasize the usefulness of mitogenomes in reconstructing the tree of life.

We thank Utako Kikutani for drawing the pictures used in Fig. 2 and Teruyuki Komiya for providing the colugo photo used in Fig. 3. This work was supported by NSFC 30925004 to Y.Z. and by Grants-in-Aid for Scientific Research C22570099 and C25440219 to M.H. from JSPS.

REFERENCES

- Akaike, H. (1973) Information theory as an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory* (eds.: Petrov, B. N., and Csaki, F.), pp. 267–281. Akademiai Kiado, Budapest.
- Arnason, U., Adegoke, J. A., Bodin, K., Born, E. W., Esa, Y. B., Gullberg, A., Nilsson, M., Short, R. V., Xu, X., and Janke, A. (2002) Mammalian mitogenomic relationships and the root of the eutherian tree. *Proc. Natl. Acad. Sci. USA* **99**, 8151–8156.
- Arnason, U., Adegoke, J. A., Gullberg, A., Harley, E. H., Janke, A., and Kullberg, M. (2008) Mitogenomic relationships of placental mammals and molecular estimates of their divergences. *Gene* **421**, 37–51.
- Asher, R. J., and Helgen, K. M. (2010) Nomenclature and placental mammal phylogeny. *BMC Evol. Biol.* **10**, 102.
- Cao, Y., Fujiwara, M., Nikaido, M., Okada, N., and Hasegawa, M. (2000) Interordinal relationships and timescale of eutherian evolution as inferred from mitochondrial genome data. *Gene* **259**, 149–158.
- Chatterjee, H. J., Ho, S. Y., Barnes, I., and Groves, C. (2009) Estimating the phylogeny and divergence times of primates using a supermatrix approach. *BMC Evol. Biol.* **9**, 259.
- Corl, A., and Ellegren, H. (2013) Sampling strategies for species trees: The effects on phylogenetic inference of the number of genes, number of individuals, and whether loci are mitochondrial, sex-linked, or autosomal. *Mol. Phylogenet. Evol.* **67**, 358–366.
- da Fonseca, R. R., Johnson, W. E., O'Brien, S. J., Ramos, M. J., and Antunes, A. (2008) The adaptive evolution of the mammalian mitochondrial genome. *BMC Genomics* **9**, 119.
- DeBry, R. W. (2005) The systematic component of phylogenetic error as a function of taxonomic sampling under parsimony. *Syst. Biol.* **54**, 432–440.
- Dimitrov, D., Lopardo, L., Giribet, G., Arnedo, M. A., Álvarez-Padilla, F., and Hormiga, G. (2012) Tangled in a sparse spider web: single origin of orb weavers and their spinning work unravelled by denser taxonomic sampling. *Proc. Roy. Soc. B* **279**, 1341–1350.
- dos Reis, M., Inoue, J., Hasegawa, M., Asher, R. J., Donoghue, P. C., and Yang, Z. (2012) Phylogenomic datasets provide both precision and accuracy in estimating the timescale of placental mammal phylogeny. *Proc. Roy. Soc. B* **279**, 3491–3500.
- Douady, C. J., Chatelier, P. I., Madsen, O., de Jong, W. W., Catzefflis, F., Springer, M. S., and Stanhope, M. J. (2002) Molecular phylogenetic evidence confirming the Eulipotyphla concept and in support of hedgehogs as the sister group to shrews. *Mol. Phylogenet. Evol.* **25**, 200–209.
- Edgar, R. C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797.
- Fan, Y., Huang, Z. Y., Cao, C. C., Chen, C. S., Chen, Y. X., Fan, D. D., He, J., Hou, H. L., Hu, L., Hu, et al. (2013) Genome of the Chinese tree shrew. *Nat. Commun.* **4**, 1426.
- Felsenstein, J. (1978) Cases in which parsimony or compatibility methods will be positively misleading. *Syst. Zool.* **27**, 401–410.
- Felsenstein, J. (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* **17**, 368–376.
- Fiala, K. L., and Sokal, R. R. (1985) Factors determining the accuracy of cladogram estimation: evaluation using computer simulation. *Evolution* **39**, 609–622.
- Gibson, A., Gowri-Shankar, V., Higgs, P. G., and Rattray, M. (2005) A comprehensive analysis of mammalian mitochondrial genome base composition and improved phylogenetic methods. *Mol. Biol. Evol.* **22**, 251–264.
- Goloboff, P. A., Catalano, S. A., Mirande, J. M., Szumik, C. A., Arias, J. S., Kallersjo, M., and Farris, J. S. (2009) Phylogenetic analysis of 73 060 taxa corroborates major eukaryotic groups. *Cladistics* **25**, 211–230.
- Hallström, B. M., Schneider, A., Zoller, S., and Janke, A. (2011) A genomic approach to examine the complex evolution of Laurasiatherian mammals. *PLoS One* **6**, e28199.
- Hasegawa, M., Kishino, H., and Yano, T. (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* **22**, 160–174.
- Hassanin, A., Delsuc, F., Ropiquet, A., Hammer, C., Jansen van Vuuren, B., Matthee, C., Ruiz-Garcia, M., Catzefflis, F., Areskoug, V., Nguyen, T. T., et al. (2012) Pattern and timing of diversification of Cetartiodactyla (Mammalia, Laurasiatheria), as revealed by a comprehensive analysis of mitochondrial genomes. *C. R. Biol.* **335**, 32–50.
- Heath, T. A., Hedtke, S. M., and Hillis, D. M. (2008) Taxon sampling and the accuracy of phylogenetic analyses. *J. Syst. Evol.* **46**, 239–257.
- Hedtke, S. M., Townsend, T. M., and Hillis, D. M. (2006) Resolution of phylogenetic conflict in large data sets by increased taxon sampling. *Syst. Biol.* **55**, 522–529.
- Hillis, D. M. (1996) Inferring complex phylogenies. *Nature* **383**, 130–131.
- Hudlot, C., Gowri-Shankar, V., Jow, H., Rattray, M., and Higgs, P. G. (2003) RNA-based phylogenetic methods: application to mammalian mitochondrial RNA sequences. *Mol.*

- Phylogenet. Evol. **28**, 241–252.
- Huelsenbeck, J. P. (1991) When are fossils better than extant taxa in phylogenetic analysis? *Syst. Zool.* **40**, 458–469.
- Inagaki, Y., Susko, E., Fast, N. M., and Roger, A. (2004) Covariance shifts cause a long-branch attraction artifact that unites microsporidia and archaeobacteria in EF-1 α phylogenies. *Mol. Biol. Evol.* **21**, 1340–1349.
- Jameson, N. M., Hou, Z. C., Sterner, K. N., Weckle, A., Goodman, M., Steiper, M. E., and Wildman, D. E. (2011) Genomic data reject the hypothesis of a prosimian primate clade. *J. Hum. Evol.* **61**, 295–305.
- Janečka, J. E., Miller, W., Pringle, T. H., Wiens, F., Zitzmann, A., Helgen, K. M., Springer, M. S., and Murphy, W. J. (2007) Molecular and genomic data identify the closest living relative of primates. *Science* **318**, 792–794.
- Kim, J. (1996) General inconsistency conditions for maximum parsimony: Effects of branch lengths and increasing number of taxa. *Syst. Biol.* **45**, 363–374.
- Kim, J. (1998) Large-scale phylogenies and measuring the performance of phylogenetic estimators. *Syst. Biol.* **47**, 43–60.
- Kjer, K. M., and Honeycutt, R. L. (2007) Site specific rates of mitochondrial genomes and the phylogeny of eutheria. *BMC Evol. Biol.* **7**, 8.
- Kriegs, J. O., Churakov, G., Jurka, J., Brosius, J., and Schmitz, J. (2007) Evolutionary history of 7SL RNA-derived SINEs in Supraprimates. *Trends Genet.* **23**, 158–161.
- Kück, P., Mayer, C., Wägele, J. W., and Misof, B. (2012) Long branch effects distort maximum likelihood phylogenies in simulations despite selection of the correct model. *PLoS One* **7**, e36593.
- Li, Y., Liu, Z., Shi, P., and Zhang, J. (2010) The hearing gene *Prestin* unites echolocating bats and whales. *Curr. Biol.* **20**, R55–R56.
- Lin, Y. H., McLenachan, P. A., Gore, A. R., Phillips, M. J., Ota, R., Hendy, M. D., and Penny, D. (2002) Four new mitochondrial genomes and the increased stability of evolutionary trees of mammals from improved taxon sampling. *Mol. Biol. Evol.* **19**, 2060–2070.
- Liu, Y., Cotton, J. A., Shen, B., Han, X., Rossiter, S. J., and Zhang, S. (2010) Convergent sequence evolution between echolocating bats and dolphins. *Curr. Biol.* **20**, R53–R54.
- Lopez, P., Casane, D., and Philippe, H. (2002) Heterotachy, an important process of protein evolution. *Mol. Biol. Evol.* **19**, 1–7.
- Martin, R. D. (2008) Colugos: obscure mammals glide into the evolutionary limelight. *J. Biol.* **7**, 13.
- Matsui, A., Rakotondraparany, F., Munechika, I., Hasegawa, M., and Horai, S. (2009) Molecular phylogeny and evolution of prosimians based on complete sequences of mitochondrial DNAs. *Gene* **441**, 53–66.
- McCormack, J. E., Faircloth, B. C., Crawford, N. G., Gowaty, P. A., Brumfield, R. T., and Glenn, T. C. (2012) Ultraconserved elements are novel phylogenomic markers that resolve placental mammal phylogeny when combined with species-tree analysis. *Genome Res.* **22**, 746–754.
- Meredith, R. W., Janečka, J. E., Gatesy, J., Ryder, O. A., Fisher, C. A., Teeling, E. C., Goodbla, A., Eizirik, E., Simão, T. L., Stadler, T., et al. (2011) Impacts of the Cretaceous Terrestrial Revolution and KPg extinction on mammal diversification. *Science* **334**, 521–524.
- Murphy, W. J., Eizirik, E., O'Brien, S. J., Madsen, O., Scally, M., Douady, C. J., Teeling, E., Ryder, O. A., Stanhope, M. J., de Jong, W. W., et al. (2001) Resolution of the early placental mammal radiation using Bayesian phylogenetics. *Science* **294**, 2348–2351.
- Nabhan, A. R., and Sarkar, I. N. (2012) The impact of taxon sampling on phylogenetic inference: a review of two decades of controversy. *Brief. Bioinform.* **13**, 122–134.
- Nie, W., Fu, B., O'Brien, P. C., Wang, J., Su, W., Tanomtong, A., Volobouev, V., Ferguson-Smith, M. A., and Yang, F. (2008) Flying lemurs—the ‘flying tree shrews’? Molecular cytogenetic evidence for a Scandentia-Dermoptera sister clade. *BMC Biol.* **6**, 18.
- Nikaido, M., Cao, Y., Harada, M., Okada, N., and Hasegawa, M. (2003) Mitochondrial phylogeny of hedgehogs and monophyly of Eulipotyphla. *Mol. Phylogenet. Evol.* **28**, 276–284.
- Nishihara, H., Hasegawa, M., and Okada, N. (2006) Pegasoferae, an unexpected mammalian clade revealed by tracking ancient retroposon insertions. *Proc. Natl. Acad. Sci. USA* **103**, 9929–9934.
- Nishihara, H., Okada, N., and Hasegawa, M. (2007) Rooting the eutherian tree: the power and pitfalls of phylogenomics. *Genome Biol.* **8**, R199.
- Perelman, P., Johnson, W. E., Roos, C., Seuanez, H. N., Horvath, J. E., Moreira, M. A., Kessing, B., Pontius, J., Roelke, M., Rumpler, Y., et al. (2011) A molecular phylogeny of living primates. *PLoS Genet.* **7**, e1001342.
- Phillips, M. J., Delsuc, F., and Penny, D. (2004) Genome-scale phylogeny and the detection of systematic biases. *Mol. Biol. Evol.* **21**, 1455–1458.
- Poe, S. (2003) Evaluation of the strategy of long-branch subdivision to improve the accuracy of phylogenetic methods. *Syst. Biol.* **52**, 423–428.
- Prasad, A. B., Allard, M. W., Program, N. C. S., and Green, E. D. (2008) Confirming the phylogeny of mammals by use of large comparative sequence data sets. *Mol. Biol. Evol.* **25**, 1795–1808.
- Reyes, A., Gissi, C., Catzeflis, F., Nevo, E., Pesole, G., and Saccone, C. (2004) Congruent mammalian trees from mitochondrial and nuclear genes using Bayesian methods. *Mol. Biol. Evol.* **21**, 397–403.
- Roberts, T. E., Lanier, H. C., Sargis, E. J., and Olson, L. E. (2011) Molecular phylogeny of treeshrews (Mammalia: Scandentia) and the timescale of diversification in Southeast Asia. *Mol. Phylogenet. Evol.* **60**, 358–372.
- Rodriguez, F., Oliver, J. L., Marin, A., and Medina, J. R. (1990) The general stochastic model of nucleotide substitution. *J. Theor. Biol.* **142**, 485–501.
- Rokas, A., and Carroll, S. B. (2005) More genes or more taxa? the relative contribution of gene number and taxon number to phylogenetic accuracy. *Mol. Biol. Evol.* **22**, 1337–1344.
- Rosenberg, M. S., and Kumar, S. (2001) Incomplete taxon sampling is not a problem for phylogenetic inference. *Proc. Natl. Acad. Sci. USA* **98**, 10751–10756.
- San Mauro, D., Gower, D. J., Cotton, J. A., Zardoya, R., Wilkinson, M., and Massingham, T. (2012) Experimental design in phylogenetics: testing predictions from expected information. *Syst. Biol.* **61**, 661–674.
- Schmitz, J., and Zischler, H. (2003) A novel family of tRNA-derived SINEs in the colugo and two new retrotransposable markers separating dermopterans from primates. *Mol. Phylogenet. Evol.* **28**, 341–349.
- Schmitz, J., Ohme, M., and Zischler, H. (2001) SINE insertions in cladistic analyses and the phylogenetic affiliations of *Tarsius bancanus* to other primates. *Genetics* **157**, 777–784.
- Schmitz, J., Ohme, M., Suryobroto, B., and Zischler, H. (2002a) The colugo (*Cynocephalus variegatus*, Dermoptera): the primates’ gliding sister? *Mol. Biol. Evol.* **19**, 2308–2312.
- Schmitz, J., Ohme, M., and Zischler, H. (2002b) The complete

- mitochondrial sequence of *Tarsius bancanus*: evidence for an extensive nucleotide compositional plasticity of primate mitochondrial DNA. *Mol. Biol. Evol.* **19**, 544–553.
- Seo, T. K., and Kishino, H. (2008) Synonymous substitutions substantially improve evolutionary inference from highly diverged proteins. *Syst. Biol.* **57**, 367–377.
- Simmons, M. P., Zhang, L. B., Webb, C. T., and Reeves, A. (2006) How can third codon positions outperform first and second codon positions in phylogenetic inference? An empirical example from the seed plants. *Syst. Biol.* **55**, 245–258.
- Song, S., Liu, L., Edwards, S. V., and Wu, S. (2012) Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. *Proc. Natl. Acad. Sci. USA* **109**, 14942–14947.
- Stamatakis, A. (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**, 2688–2690.
- Stamatakis, A., Hoover, P., and Rougemont, J. (2008) A rapid bootstrap algorithm for the RAxML Web servers. *Syst. Biol.* **57**, 758–771.
- Townsend, J. P., and Lopez-Giraldez, F. (2010) Optimal selection of gene and ingroup taxon sampling for resolving phylogenetic relationships. *Syst. Biol.* **59**, 446–457.
- Townsend, J. P., and Leuenberger, C. (2011) Taxon sampling and the optimal rates of evolution for phylogenetic inference. *Syst. Biol.* **60**, 358–365.
- Xu, L., Chen, S. Y., Nie, W. H., Jiang, X. L., and Yao, Y. G. (2012) Evaluating the phylogenetic position of Chinese tree shrew (*Tupaia belangeri chinensis*) based on complete mitochondrial genome: implication for using tree shrew as an alternative experimental animal to primates in biomedical research. *J. Genet. Genomics* **39**, 131–137.
- Yang, Z. (1994) Estimating the pattern of nucleotide substitution. *J. Mol. Evol.* **39**, 105–111.
- Yang, Z. (1996) Among-site rate variation and its impact on phylogenetic analyses. *Trends Ecol. Evol.* **11**, 367–372.
- Yang, Z. (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591.
- Yang, Z., Nielsen, R., and Hasegawa, M. (1998) Models of amino acid substitution and applications to mitochondrial protein evolution. *Mol. Biol. Evol.* **15**, 1600–1611.
- Yonezawa, T., and Hasegawa, M. (2010) Was the universal common ancestry proved? *Nature* **468**, E9; discussion E10.
- Yonezawa, T., and Hasegawa, M. (2012) Some problems in proving the existence of the universal common ancestor of life on Earth. *The Scientific World Journal* **2012**, 479824.
- Zhao, L., Annie, A. S., Amrita, S., Yi, S. K., and Rudolf, M. (2013) Does better taxon sampling help? A new phylogenetic hypothesis for Sepsidae (Diptera: Cyclorrhapha) based on 50 new taxa and the same old mitochondrial and nuclear markers. *Mol. Phylogenet. Evol.* **69**, 153–164.
- Zhou, X., Xu, S., Xu, J., Chen, B., Zhou, K., and Yang, G. (2012) Phylogenomic analysis resolves the interordinal relationships and rapid diversification of the Laurasiatherian mammals. *Syst. Biol.* **61**, 150–164.
- Zwickl, D. J., and Hillis, D. M. (2002) Increased taxon sampling greatly reduces phylogenetic error. *Syst. Biol.* **51**, 588–598.