Phylogenomic Resolution of the Phylogeny of Laurasiatherian Mammals: Exploring Phylogenetic Signals within Coding and Noncoding Sequences

Meng-Yun Chen[†], Dan Liang[†], and Peng Zhang*

State Key Laboratory of Biocontrol, College of Ecology and Evolution, School of Life Sciences, Sun Yat-Sen University, Guangzhou, China

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

Accepted: July 30, 2017

[†]These authors contributed equally to this work.

Abstract

The interordinal relationships of Laurasiatherian mammals are currently one of the most controversial questions in mammalian phylogenetics. Previous studies mainly relied on coding sequences (CDS) and seldom used noncoding sequences. Here, by data mining public genome data, we compiled an intron data set of 3,638 genes (all introns from a protein-coding gene are considered as a gene) (19,055,073 bp) and a CDS data set of 10,259 genes (20,994,285 bp), covering all major lineages of Laurasiatheria (except Pholidota). We found that the intron data contained stronger and more congruent phylogenetic signals than the CDS data. In agreement with this observation, concatenation and species-tree analyses of the intron data set yielded well-resolved and identical phylogenies, whereas the CDS data set produced weakly supported and incongruent results. Further analyses showed that the phylogeny inferred from the intron data is highly robust to data subsampling and change in outgroup, but the CDS data produced unstable results under the same conditions. Interestingly, gene tree statistical results showed that the most frequently observed gene tree topologies for the CDS and intron data are identical, suggesting that the major phylogenetic signal within the CDS data is actually congruent with that within the intron data. Our final result of Laurasiatheria phylogeny is (Eulipotyphla,((Chiroptera, Perissodactyla),(Carnivora, Cetartiodactyla))), favoring a close relationship between Chiroptera and Perissodactyla. Our study 1) provides a well-supported phylogenetic framework for Laurasiatheria, representing a step towards ending the long-standing "hard" polytomy and 2) argues that intron within genome data is a promising data resource for resolving rapid radiation events across the tree of life.

Key words: phylogenomics, phylogeny, intron, noncoding, data subsampling, Laurasiatheria.

Introduction

Reconstructing relationships of clades that experienced rapid radiation in deep geological time has proved to be particularly difficult (Meredith et al. 2011; McCormack et al. 2012; Song et al. 2012; Zhou et al. 2012; Morgan et al. 2013; Romiguier et al. 2013; Jarvis et al. 2014; Prum et al. 2015; Irisarri and Meyer 2016; Tarver et al. 2016). During rapid radiation, speciation events occur within a relatively short time interval, few phylogenetic signals accumulate, and these signals are often obscured by subsequent substitutions long after the speciation event. In this context, recent research has often used genome-scale data sets, aiming to provide more signals to resolve the phylogenetic relationships resulting from rapid radiation (Amemiya et al. 2013; Jarvis et al. 2014; Irisarri and Meyer 2016).

Previous phylogenomic studies have predominantly used coding sequences (CDS) because identifying orthologous sequences of coding regions is relatively easy, and the conservativeness of CDS also makes alignment across divergent species relatively straightforward (Thomson et al. 2010). However, the conservativeness of CDS also means they carry fewer phylogenetic signals that can resolve rapid radiation events compared with other fast-evolving genomic regions. Additionally, CDS are functionally constrained and are thus potentially subject to convergent evolution. Recently, Jarvis et al. (2014) used genome-scale data, including protein-coding sequences, intron sequences, and ultraconserved elements, to reconstruct the Neoaves phylogeny, a well-known ancient radiation. They found that noncoding sequences contributed most phylogenetic signals to the resultant tree, whereas CDS yielded relationships reflecting convergent life history traits. Compared with CDS, noncoding sequences are less prone to convergent evolution and carry more variable characters, making them informative even at shallow timescales (Chojnowski et al. 2008; Yu et al. 2011; Foley et al. 2015). However, the high variability of noncoding sequences also poses a challenge in determining orthology and accurately aligning sequences. Every type of data has advantages and disadvantages, and a deep comparison of the phylogenetic utility of coding and noncoding sequences will provide helpful information for future phylogenomic practices in resolving the relationships of rapidly radiating clades.

Laurasiatheria is one of the most diverse superorders of placental mammals, and its evolution represents a typical ancient rapid radiation. The phylogenetic relationships among the six orders of Laurasiatheria are currently a hotly debated topic (Nikolaev et al. 2007; Prasad et al. 2008; Meredith et al. 2011; Zhou et al. 2012; dos Reis et al. 2012; McCormack et al. 2012; Nery et al. 2012; Song et al. 2012; O'Leary et al. 2013; Romiguier et al. 2013; Tsagkogeorga et al. 2013). After years of intensive studies, there are only two placements with some certainty: Eulipotyphla is sister to all other laurasiatherians, and Pholidota is the sister group of Carnivora. Phylogenetic placements of the other four orders Carnivora, Perissodactyla, Cetartiodactyla, and Chiroptera remain a subject of heated debate, and many hypotheses exist. For example, some authors argue that Chiroptera is sister to other Laurasiatheria, with the exclusion of Eulipotyphla (Murphy et al. 2001; Song et al. 2012; Tsagkogeorga et al. 2013), but there is also evidence that Cetartiodactyla should be placed in this position (dos Reis et al. 2012; McCormack et al. 2012). Some authors have argued that Perissodactyla and Cetartiodactyla is a clade (Meredith et al. 2011; Zhou et al. 2011; O'Leary et al. 2013), whereas other researchers have speculated that Perissodactyla is instead close to Carnivora (Nery et al. 2012; Romiguier et al. 2013). Ruling out any of these hypotheses has been difficult because the divergence among major laurasiatherian lineages occurred over a short time of 1–4 Myr (Hallström and Janke 2010). Therefore, reconstructing the true phylogenetic relationships of Laurasiatheria provides an ideal case study to explore the phylogenetic performance of both coding and noncoding sequences in resolving an ancient rapid radiation.

Here, by data-mining 22 available published placental mammalian genomes, we compiled two genome-scale data sets of coding and noncoding regions to reconstruct the Laurasiatheria phylogeny: a CDS data set of 10,259 loci (20,994,285 bp), and an intron data set of 3,638 genes (19,055,073 bp). Our analyses showed that the intron data produced well-resolved interordinal phylogeny for Laurasiatheria but the CDS data yielded weakly supported results, partially because the intron data contains more homogeneous and stronger phylogenetic signals than the CDS data. With comparative analyses of data subsampling and a change in outgroup, we found that noncoding sequences can provide more robust phylogeny than CDS in resolving the Laurasiatheria phylogeny. Our study highlights the potential phylogenetic utility of noncoding sequences in resolving ancient rapid radiation events in future phylogenomic practices.

Materials and Methods

Data Collection and Orthology Assignment

CDS Data Set

Orthologous coding DNA sequences (CDS) were retrieved from the OrthoMaM database (Douzery et al. 2014) for 21 placental mammalian species comprising two Eulipotyphla, two Chiroptera, one Perissodactyla, four Carnivora, five Cetartiodactyla, and seven outgroups (two Xenarthra, three Afrotheria, and two Euarchontoglires). Because only one species from Perissodactyla is available in the OrthoMaM database (horse; *Equus caballus*), we additionally downloaded all predicted mRNA transcripts sequences of the southern white rhinoceros (Perissodactyla: *Ceratotherium simum simum*) from NCBI. We performed a reciprocal best-hit (RBH) BLAST analysis between transcripts of rhinoceros and all CDS sequences of *Homo sapiens* to identify the rhinoceros orthologs (cut-off: *E* value≤10-5, identity≥70%, coverage≥60%). To avoid introducing paralogs, we used a strict RBH criterion: the score of the second BLAST hit was required to be less than 50% of the best hit for the sequence to be retained for further analysis. The reading frame of obtained sequences was checked with a custom Python script, and the sequences of rhinoceros were incorporated into the supermatrix. In total, 22 species were included in the final CDS data set (supplementary table S1, Supplementary Material online).

Intron Data Set

We identified orthologous introns based on the orthology of their flanking exons on both sides. To do this, predicted exon sequences and genomic sequences were retrieved from the Ensembl database using the Biomart tool for 21 placental mammals (http://www.ensembl.org). For each species, exon sequences were sorted in ascending order by their genomic location and were named using their gene name plus their genomic location information. Exon orthology was also assessed using an RBH BLAST procedure (cut-off: *E* value≤10-5, identity≥70%, and coverage≥60%). We defined orthology as present if bidirectional BLAST best hits were found between the query species and the reference species *Homo sapiens*. Moreover, a candidate orthologous exon was retained only if it satisfied the following criteria: 1) the ratio of the BLAST score of the first best-hit to the BLAST score of the second hit was >1.5; and 2) the difference ratio of exon length between *Homo sapiens* and the query species was

<0.6. Filtered orthologous exons and syntenic adjacent orthologous exons from the same gene (according to their exon names) constituted an exon pair. The exon pair and the intron between the two exons were considered an exon–intron–exon unit. The full-length sequence of an exon–intron–exon unit was extracted from the genome sequences based on the genomic location information of the two exons. Because the rhinoceros genome lacks exon annotation, we mapped all obtained exon–intron–exon units of *Homo sapiens* onto the rhinoceros genome scaffolds using BLASTN and extracted the corresponding sequences of the rhinoceros. To exclude introns with high length variability (difficult to align), we calculated the intron length difference between the other 21 mammals and *Homo sapiens* and filtered out exon–intron–exon units whose intron length difference ratio > 50%. In addition, we also filtered out exon–intron–exon units whose intron length > 10,000 bp. All of the above steps were performed with custom Python scripts (available at https://github.com/chenmy33/IntronGetting).

Alignment and Alignment Refining

For the CDS data set, sequences of 21 mammals (except southern white rhinoceros) from the OrthoMaM database had already been aligned, and poorly aligned regions had been removed. Therefore, we only needed to add the sequences from rhinoceros into the existing alignments using MAFFT *L-INSI* with the option "–add –keeplength." The "–keeplength" option maintains the length of existing alignments, therefore preserving the original codon reading frames (Katoh & Standley 2013).

For the intron data set, each exon–intron–exon unit contained two parts: a highly divergent intron and its two conserved flanking exons. Fast-evolving introns are difficult to align, but the flanking exons acted as two anchoring regions when aligning intron sequences and thus improved the accuracy of alignments to some extent. Multiple sequence alignment was conducted using the program SATé (Liu et al. 2012) with the "—auto" option, because iterative methods such as SATé can reduce error in aligning highly divergent sequences (Lemmon and Lemmon 2013). The resulting alignments were further refined using Gblocks 0.91 b (Castresana 2000), with half gaps allowed (−b5 = h). For each alignment of an exon–intron–exon unit, the flanking exon sequences were trimmed so that the final alignment contained only intron sequences. If several introns were from the same host gene, they were merged and treated as a single gene. To further exclude possible errors in orthology assignment and alignment, we constructed an ML tree under a GTR + GAMMA model using RAxML v8.2.2 (Stamatakis 2014) for each intron alignment. If a tree contained extremely long branches that accounted for >50% of the total tree length, the corresponding sequences were removed from the alignment.

Finally, to focus our data on our question of interest, for both the CDS and intron data sets, only gene alignments that contained at least one sequence for each related order (Chiroptera, Perissodactyla, Carnivora, Cetartiodactyla) were retained. To reduce random (or sampling) error in building gene trees, gene alignments shorter than 600 bp were also discarded. The entire data process resulted in a final set of 10,259 CDS alignments and a final set of 3,638 intron alignments for the next step of analysis.

Gene Tree Statistics

Gene trees for each CDS and intron alignment were constructed using RAxML v8.2.2 with the GTR + GAMMA model. We performed 100 bootstrap replicates and searched for the best-scoring ML tree in a single run for each gene tree in RAxML (-f a option). We then used a custom Python script to analyze all gene trees and classified them into 15 possible hypotheses regarding the phylogenetic positions of Cetartiodactyla, Perissodactyla, Carnivora, and Chiroptera, without taking branch support values into account. Gene trees that did not support any of the 15 alternative hypotheses were categorized as "nonmatching" (Chen et al. 2015). We calculated the proportion of gene trees classified into the "nonmatching" category or 15 alternative hypotheses.

To investigate the degree of incongruence among gene trees, we calculated the pairwise Robinson–Foulds distances (Robinson & Foulds 1981) between gene trees using Gori's (2016) python script. The average RF distance for a gene tree relative to all other gene trees was calculated using a custom Python script. The average RF distance of genes and pairwise RF distances were used to plot the histogram and tree space was visualized using multidimensional scaling (MDS) in R (Hillis et al. 2005).

Data Subsampling with Various Gene Filtering Methods

Many factors influence the accuracy of phylogenetic inference, such as the GC content of genes, evolutionary rates of genes, phylogenetic resolution of genes, missing data, and the choice of outgroup. To explore the influence of aforementioned factors on the resulting phylogenies for both the CDS and intron data sets, we generated a series of data subsets from the original data.

GC Content of Genes

We calculated the average GC content at the third codon position (GC3%) for each gene in the CDS data set and average GC content (GC%) for each gene in the intron data set. The original 10,259-CDS data set was divided into three subsets with different thresholds of GC3%: 7,697 genes ("GC_CDS1" data set: GC3% < 73.1%), 5,130 genes ("GC_CDS2" data set: GC3% < 60.4%), and 2,850 genes ("GC_CDS3" data set: GC3% < 48%). The original 3,638-

intron data set was divided into three subsets with different thresholds of GC%: 3,445 genes ("GC_Intron1" data set: GC% < 56%), 3,125 genes ("GC_Intron2" data set: GC% < 50%), and 2,728 genes ("GC_Intron3" data set: GC% < 44.6%).

Evolutionary Rate of Genes

We used average pairwise identity as an approximation of the evolutionary rate of a gene. For the CDS data set, two subsets of genes were selected: one contained 50% of genes with a slow evolutionary rate (5,127 genes: "Rate_CDS1" data set) and the other contained 50% of genes with a fast evolutionary rate (5,132 genes: "Rate_CDS2" data set). Similarly, for the intron data set, two subsets of genes were selected: one contained 50% of genes with a slow evolutionary rate (1,817 genes: "Rate_Intron1" data set) and the other contained 50% of genes with a fast evolutionary rate (1,821 genes: "Rate_Intron2" data set).

Resolution of Genes

The average bootstrap support of the ML tree is the resolution of a gene. For the CDS data set, we generated three subsets, selecting genes with average bootstrap support values ≥70% (5,698 genes: "Resolution_CDS1" data set), 80% (2,702 genes: "Resolution_CDS2" data set) and 90% (450 genes: "Resolution_CDS3" data set). For the intron data set, we generated three subsets, selecting genes with average bootstrap support values ≥70% (3,279 genes: "Resolution_Intron1" data set), 80% (2,537 genes: "Resolution_Intron2" data set), and 90% (1,068 genes: "Resolution_Intron3" data set).

Missing Data

For the CDS data set, we generated three subsets that allowed for a maximum of 30% missing data per gene (9,870 genes: "Completeness_CDS1" data set), 20% missing data (8,808 genes: "Completeness_CDS2" data set), and 10% missing data (6,029 genes: "Completeness_CDS3" data set). For the intron data set, three gene subsets were selected based on different levels of missing data: 50% (2,521 genes: "Completeness_Intron1" data set), 40% (1,313 genes: "Completeness_Intron2" data set), and 30% (404 genes: "Completeness_Intron3" data set).

Outgroup

For the original CDS and intron data sets, we used seven species from Afrotheria, Xenarthra, and Euarchontoglires as the outgroup to Laurasiatheria. To investigate the influence of the choice of outgroup on phylogenetic reconstruction, we generated three taxon-reduced data sets with the same set of genes but different composition of outgroup species: "X + E_out_CDS" and "X + E_out _Intron" (using only Xenarthra and Euarchontoglires as outgroup), "A + E_out_CDS" and "A + E_out_Intron" (using only Afrotheria and Euarchontoglires as outgroup), and "A + X_out_CDS" and "A + X_out_Intron" (using only Afrotheria and Xenarthra as outgroup).

Phylogenetic Analyses

For the original 10,259-CDS data set, the original 3,638-intron data set, and their data subsets, phylogenetic trees were reconstructed using both maximum likelihood (RAxML; Stamatakis 2014) and coalescent-based species-tree inference (ASTRAL; Mirarab and Warnow 2015). Because our data sets included thousands of genes, using gene partitioning scheme was not applicable. Moreover, in our pilot analyses, we found that different partitioning schemes had almost no effect on the final phylogenetic results. Therefore, all CDS data sets were partitioned by three codon positions and all intron data sets were not partitioned. The best-fitted models for every data partition were selected with PartitionFinder (Lanfear et al. 2012). In nearly all cases, GTR + GAMMA + I model was the best-fitted model for the first codon partitions and the second codon partitions of the CDS data sets and for the intron data sets, whereas GTR + GAMMA model best fitted the third codon partitions of the CDS data sets (supplementary table S2, Supplementary Material online). Because the usage of P-Invar in combination with Gamma can lead to a ping-pong effect which makes alpha and PInvar cannot be optimized independently from each other (RAxML manual), the authors of RAxML always suggest users to use GTR + GAMMA model instead of GTR + GAMMA + I model. In fact, in our pilot analyses, using the GTR + GAMMA + I model produced nearly identical topologies and branch support as using the GTR + GAMMA model. Therefore, the GTR + GAMMA model was used for all data partitions in both the CDS and intron data sets in our phylogenetic analyses.

ML analyses were performed with RAxML v8.2.2 (Stamatakis 2014) on a two-way high-performance computation station (two E5-2680 CPU, 2.8 GHz, 256 G RAM) using 32 threads. Branch support was estimated with 200 rapid bootstrapping replicates (option -f a). A species tree for each data set was reconstructed using the gene-tree-based coalescent approach implemented in the program ASTRAL v4.7.6 (Mirarab and Warnow 2015). Briefly, for each data set, 200 ML bootstrap trees and the final best-scoring ML tree were estimated for every gene in the data set, using RAxML under the GTR + GAMMA model. These best-scoring ML trees and bootstrapping trees were used as input files for ASTRAL with the option: "–i –b" to calculate the final species tree and branch support.

Results

Data Characteristics

The CDS data set comprised 10,259 genes (89.4% complete for the 22 taxa) and 20,994,285 bp; the intron data set contained 3,638 genes (56.8% complete for the 22 taxa) and a total of 19,055,073 bp of sequence data (table 1). The lengths of CDS genes ranged from 600 to 26,394 bp (median = 1,572 bp) and the length of intron genes ranged from 600 to 62,464 bp (median= 3,451 bp) (fig. 1*A*). The intron data set had lower mean GC content and lower GC-content variation than the CDS data set, both among genes and among species (fig. 1*B*). Because GC-content is positively correlated with the rate of recombination, the low-GC intron data set thus should be less prone to GC-biased gene conversion than the CDS data set. The intron loci had higher average pairwise distance than the CDS loci, consistent with the expectation that noncoding sequences evolve more rapidly than CDS (fig. 1*C*). The average bootstrap support values for gene trees ranged from 11% to 99% (median = 72%) for the CDS data set and 32% to 100% (median = 85%) for the intron data set, indicating that the intron data set had stronger phylogenetic signals than the CDS data set (fig. 1*D*). A summary of data characteristics for each gene including alignment length, taxa occupancy, pairwise distances, GC content, percentage of missing data, and the average bootstrap support values of gene tree is given in the supplementary table (supplementary table S3A and B, Supplementary Material online). To evaluate gene tree heterogeneity, Robinson–Foulds distances among genes were calculated for both data sets. Multidimensional scaling plots of the RF distances among genes (fig. 1*E and F*) showed that the 3,638 intron gene trees were more similar to each other compared with the 10,259 CDS gene trees. The mean among-gene-tree RF distance of the intron data set was 5.028, whereas the mean among-gene-tree RF distance of the CDS data set was 12.69, suggesting that the intron data had more congruent phylogenetic signals than the CDS data (fig. 1*G*).

The CDS Data Set Was Unable to Robustly Resolve the Interordinal Relationships of Laurasiatheria

The maximum likelihood (ML) tree inferred from 10,259 genes of the CDS data set recovered the monophyly of Laurasiatheria and the monophyly of all orders represented by multiple species with 100% bootstrap support (fig. 2). In agreement with most previous studies, the ML tree recovered Eulipotyphla as the first-diverging lineage within Laurasiatheria (BS = 100%). However, this phylogeny provided weak resolution for the interrelationships of Chiroptera, Perissodactyla, Carnivora, and Cetartiodactyla. Chiroptera appeared as the sister group of all remaining Laurasiatheria to the exclusion of Eulipotyphla with low support (BS = 48%), and Perissodactyla was placed as sister to a clade containing Carnivora and Cetartiodactyla (BS = 48%). The species tree analysis (ASTRAL) of the CDS data set produced different relationships among Laurasiatherian orders; Perissodactyla and Cetartiodactyla formed a clade, but again, with weak bootstrap support (BS = 43%; fig. 2). In summary, whether using concatenation or coalescence-based phylogenetic inference, the phylogenetic signals within the CDS data set were not sufficient to fully resolve the interordinal relationships of Laurasiatheria.

The Intron Data Set Yielded Well-Resolved and Congruent Phylogenetic Relationships

Unlike the CDS data set, both ML and species tree analyses of the intron data set produced identical and fully resolved phylogenies for Laurasiatheria (fig. 3). All internodes within this phylogeny received 100% bootstrap support in both analyses. Eulipotyphla was again robustly recovered as the sister group of all other Laurasiatherian mammals. However, the interrelationships of Chiroptera, Perissodactyla, Carnivora, and Cetartiodactyla inferred from the intron data set were somewhat different from that inferred from the CDS data set. The intron phylogeny confirmed the result for CDS that Carnivora is sister to Cetartiodactyla with strong support (BS = 100%). In the intron tree, Perissodactyla is strongly supported as sister group of Chiroptera (BS = 100%).

The Intron Data Set Produced Stable Results under Different Data Resampling Conditions

Resampling of loci within the data set has been widely used as an effective strategy to investigate the consistency and stability of phylogenetic inference in genome-scale data sets (reviewed in Edwards 2016; also see Narechania et al. 2012; Salichos and Rokas 2013; Chen et al. 2015). Therefore, we generated multiple data subsets from the original intron data set according to the GC content of genes, evolutionary rate of genes, different choices of outgroups, missing data of genes, and phylogenetic informativeness of genes to investigate whether the phylogeny inferred from the intron data set was an artifact due to systematic errors, such as compositional bias, long-branch attraction (LBA), and outgroup selection, or due to random errors, such as data completeness and the phylogenetic informativeness of genes. For the sake of comparison, similar data subsets were also generated for the CDS data set according to the same criteria (see Materials and Methods for details). Information for these data subsets is given in table 1.

All data subsets were analyzed using both concatenation and species tree methods. The resulting phylogenies from these data subsets can be found in the Supplementary Material (supplementary figs. S1–S10, Supplementary Material online). Overall, we observed seven unique topologies from these analyses (topologies were colored and labeled F to L; fig. 4*A*) and summarized the results in figure 4.

Table 1
Brief Information of All Data Sets Used for Phylogenetic Inference in This Study

Data Set Names	No. of Species	No. of Genes	Alignment Length (bp)	Parsimony Informative Site	Missing Data (%)	Criteria of Gene Selection	Inferred Topologies [a](RAxML/ASTRAL)
Total10259CDS	22	10,259	20,994,285	5,379,346	10.6	All genes of the CDS data set	Tree G/Tree H
GC_CDS1	22	7,697	16,239,495	4,192,858	8.9	Average GC%t of the third codon position <73.1%	Tree J/Tree H
GC_CDS2	22	5,130	11,216,196	2,850,021	7.8	Average GC% of the third codon position <60.4%	Tree J/Tree H
GC_CDS3	22	2,850	6,460,929	1,583,067	7.1	Average GC% of the third codon position <48%	Tree J/Tree H
Rate_CDS1	22	5,127	10,101,696	2,192,851	10.9	Evolutionary rate smaller than the median rate of all genes	Tree G/Tree H
Rate_CDS2	22	5,132	10,892,589	3,186,495	10.3	Evolutionary rate larger than the median rate of all genes	Tree J/Tree G
Resolution_CDS1	22	5,698	15,118,602	4,084,517	10.5	Average gene tree bootstrap 70 or more	Tree J/Tree G
Resolution_CDS2	22	2,702	9,258,162	2,593,379	10.5	Average gene tree bootstrap 80 or more	Tree K/Tree G
Resolution_CDS3	22	450	2,524,608	745,109	10.8	Average gene tree bootstrap 90 or more	Tree K/Tree G
Completeness_CDS1	22	9,870	20,209,332	5,177,162	9.6	Amount of missing data <30%	Tree G/Tree H
Completeness_CDS2	22	8,808	18,088,938	4,609,905	7.9	Amount of missing data <20%	Tree G/Tree H
Completeness_CDS3	22	6,029	12,686,955	3,196,588	4.7	Amount of missing data <10%	Tree G/Tree H
X+E_out_CDS	19	10,259	20,994,285	4,653,797	10.7	All genes of the CDS data set but remove all Afrotheria sequences	Tree J/-
A+E_out_CDS	20	10,259	20,994,285	5,061,736	9.6	All genes of the CDS data set but remove all Xenarthra sequences	Tree G/-
A+X_out_CDS	20	10,259	20,994,285	4,761,727	11.6	All genes of the CDS data set but remove all Euarchontoglires sequences	Tree G/-
Total3638Intron	22	3,638	19,055,073	6,628,387	43.2	All genes of the Intron data set	Tree F/Tree F
GC_Intron1	22	3,445	18,630,433	6,469,836	43.1	Average GC content <56%	Tree F/Tree F
GC_Intron2	22	3,125	17,631,768	6,121,790	42.7	Average GC content <50%	Tree F/Tree F
GC_Intron3	22	2,728	16,117,169	5,583,015	42.3	Average GC content <44.6%	Tree F/Tree F
Rate_Intron1	22	1,817	7,915,005	2,474,048	43.6	Evolutionary rate smaller than the median rate of all genes	Tree F/Tree F
Rate_Intron2	22	1,821	11,140,068	4,154,339	42.9	Evolutionary rate larger than the median rate of all genes	Tree F/Tree F
Resolution_ Intron1	22	3,279	18,080,775	6,339,204	42.7	Average gene tree bootstrap 70 or more	Tree F/Tree F
Resolution_ Intron2	22	2,537	15,031,270	5,320,830	42.0	Average gene tree bootstrap 80 or more	Tree F/Tree F
Resolution_ Intron3	22	1,068	7,225,682	2,579,450	41.1	Average gene tree bootstrap 90 or more	Tree L/Tree F
Completeness_Intron1	22	2,521	14,575,834	5,456,797	38.8	Amount of missing data <50%	Tree F/Tree F
Completeness_Intron2	22	1,313	7,976,863	3,209,477	33.4	Amount of missing data <40%	Tree F/Tree F
Completeness_Intron3	22	404	2,076,439	914,069	25.9	Amount of missing data <30%	Tree L/Tree F
X+E_out_Intron	19	3,638	19,055,073	5,811,657	39.5	All genes of the Intron data set but remove all Afrotheria sequences	Tree F/-
A+E_out_Intron	20	3,638	19,055,073	6,037,174	41.7	All genes of the Intron data set but remove all Xenarthra sequences	Tree F/-
A+X_out_Intron	20	3,638	19,055,073	5,841,403	43.3	All genes of the Intron data set but remove all Euarchontoglires sequences	Tree F/-

[a] Inferred topologies corresponding to those reported in figure 4.

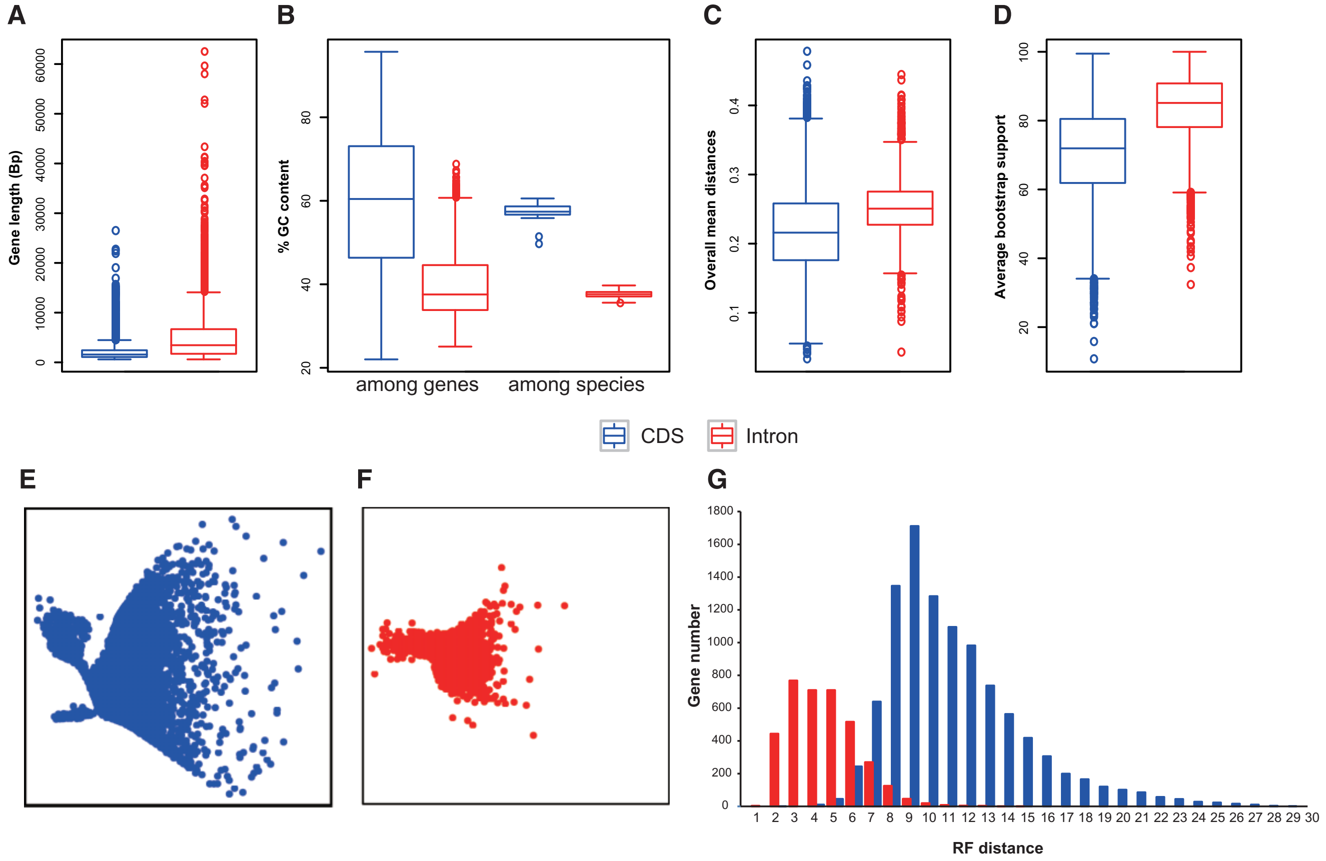


Fig. 1.—Characteristics of CDS and intron data sets (blue = CDS, red = intron). Boxplots show (*A*) variation in gene length, (*B*) GC content of each gene (among genes) and each species (among species), (*C*) relative evolutionary rates of loci (measured by the average pairwise distance for each gene), and (*D*) average bootstrap support values across all estimated gene trees. (*E*) Visualization of ML tree space using multidimensional scaling plot of 10,259 ML gene-trees from the CDS data set; each dot represents a tree inferred from one gene. Distances between dots represent Robinson–Foulds distances between gene trees. (*F*) Multidimensional scaling plot of 3,638 ML gene-trees from the intron data set. (*G*) Histogram of the average RF distance for a gene relative to all other genes, summarized from the CDS data set and the intron data set.

To our surprise, we found that the CDS data set was highly sensitive to data resampling and tree-building methods. Different subsets of the CDS data set produced highly supported but incongruent results. For example, using the concatenated ML inference, all the three CDS data subsets with lower GC content (GC_CDS1, GC_CDS2, and GC_CDS3) highly supported (BS > 90%) topology J (fig. 4*B*). Topology J was once again highly supported when using slowly evolving genes (data subset Rate_CDS2; fig. 4*B*). However, when using genes with high phylogenetic informativeness (data subset Resolution_CDS3), topology K was robustly recovered (fig. 4*B*); when using genes with more data completeness, another result, topology G, was highly supported (data subset Completeness_CDS3; fig. 4*B*). Remarkably, we did not recover the same highly supported topology when using the same data subset in the species tree (ASTRAL) analysis (fig. 4*C*). The species tree analyses highly supported topologies G and H, but the corresponding data subsets were completely different (fig. 4*C*).

In contrast to data subsampling analyses of the CDS data set, we found that phylogenies inferred from the subsets of the Intron data set were highly congruent (fig. 4*D* and *E*). The data subsets of the Intron data set overwhelmingly supported topology F (in 23 out of 25 cases; fig. 4*D* and *E*), in agreement with the result of the original intron data set. Similarly, when we used three different outgroup combinations, we found that the Intron data set consistently supported topology F with maximal support, whereas the CDS data set produced unstable results (fig. 5). These results suggest that the highly supported phylogeny inferred from the Intron data set was unlikely to be caused by systematic or random errors and might reflect the true evolutionary history of Laurasiatherian mammals.

The Major Phylogenetic Signal within the CDS and Intron Data Sets Was Congruent

To further explore the phylogenetic signal within the CDS and Intron data sets, we surveyed gene tree frequency for both

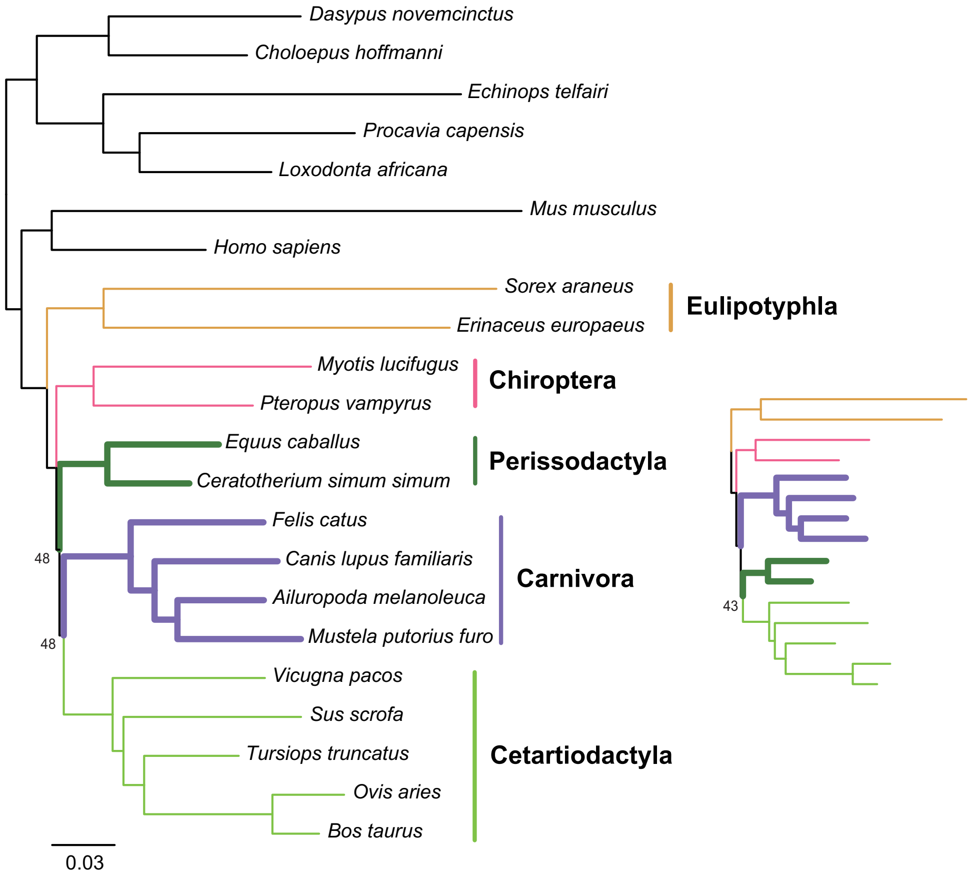


Fig. 2.—Phylogenetic relationships of Laurasiatheria inferred from the CDS data set (10,259 genes; 20,994,285 sites). Phylogeny was inferred by concatenation ML and species tree analysis using the ASTRAL program. The ML phylogeny is shown on the left, and the ASTRAL species tree is shown on the right (outgroup not shown). Values next to branches are bootstrap values. Branches without support values all received a bootstrap value of 100%.

data sets. Because the basal position of Eulipotyphla was rather robust in previous studies and this study, there are a total number of 15 alternative hypotheses for the relative positions of the remaining four orders: Chiroptera, Perissodactyla, Carnivora, and Cetartiodactyla (H1 to H15; fig. 6 and supplementary table S4, Supplementary Material online). After examining the ML tree for each gene, we found that many genes did not support any of the 15 hypotheses because they were unable to recover the expected position of Eulipotyphla or unable to recover the monophyly of Chiroptera, Perissodactyla, Carnivora, or Cetartiodactyla. These genes were recently named "nonmatching" genes. They cannot provide a meaningful answer to the question and are thus thought to have little contribution (Chen et al. 2015). We found that ~72% of genes within the CDS data set were "nonmatching" genes but <41% of intron genes were recognized as "nonmatching" genes (fig. 6). This result was in line with the observation that the Intron data set was more phylogenetically informative in resolving the Laurasiatheria phylogeny than the CDS data set.

We then examined the gene trees of the 2,852 "matching" CDS genes (27.9% of the CDS data) and 2,175 "matching" intron genes (59.8% of the Intron data). These genes could be considered to have contributed major phylogenetic signals to the Laurasiatherian question. For the intron genes, we found that hypothesis H1 received the highest gene-support frequency (9.66%) and the second-place hypothesis H2 received a support frequency of only 7.95% (fig. 6). The relative support-frequency difference between these two hypotheses is large (up to 21.5%). This observation was in line with the fact that the H1 hypothesis (our final result) was robustly supported by the intron data. However, for the CDS genes, the top-three supported hypotheses and their gene-supported frequencies are H1 (8.42%), H13 (8.14%), and H4 (7.58%) (fig. 6). The relative support-frequency differences among these hypotheses ranged from 3.44% to 11.1%. Therefore, it is much more difficult to distinguish among these three hypotheses in the analyses of CDS data. In fact, the concatenation and species-tree analyses of CDS data eventually supported hypothesis H4 (the third best) and hypothesis H13 (the second best), respectively, albeit with low support. The possible cause may be that the CDS data contained a large number of "nonmatching" genes, which diluted the genuine phylogenetic signal for the target

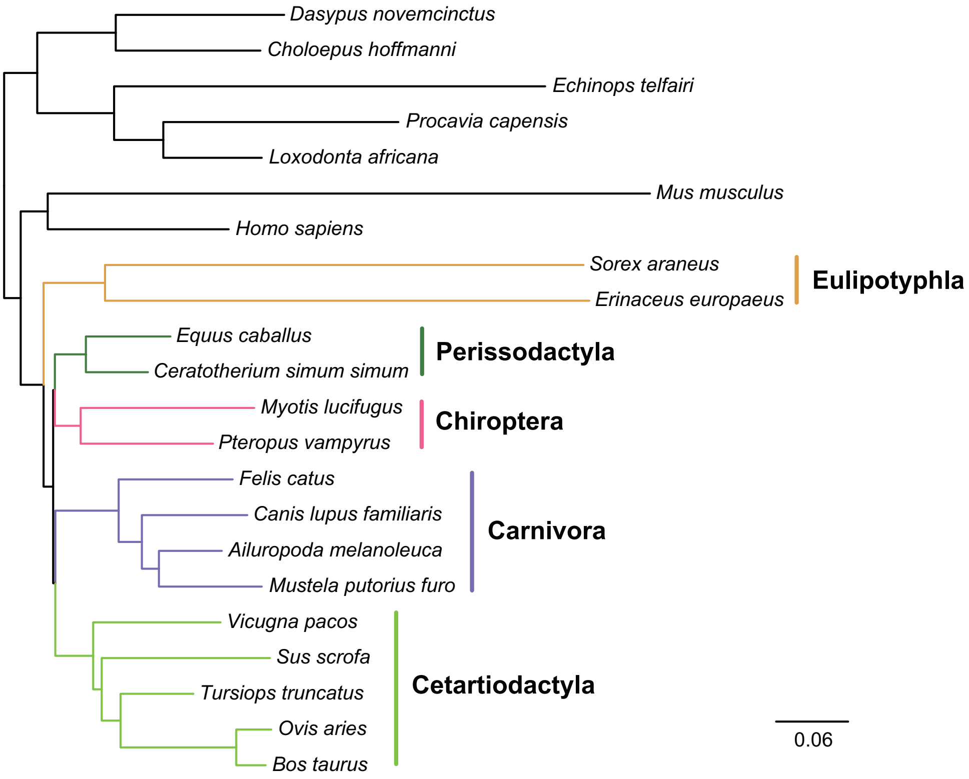


Fig. 3.—Phylogenetic relationships of Laurasiatheria inferred from the Intron data set (3,638 genes; 19,055,073 sites). Phylogeny was inferred by concatenation ML and species tree analysis using the ASTRAL program. Both analyses produced identical phylogenies for the interordinal relationships of Laurasiatherian mammals. All branches have a bootstrap value of 100% in both analyses. Branch lengths are from the ML analysis.

question. Nevertheless, it should be noted that the top supported hypotheses for both the CDS data and the intron data are the same (the H1 hypothesis). This result showed that the major phylogenetic signals of both the CDS and Intron data sets are actually congruent.

Discussion

Laurasiatheria Phylogeny

One of the most controversial problems in mammal phylogenetics is the phylogenetic relationships among the six orders of Laurasiatheria; past studies have produced variable results (reviewed by Hu et al. 2013). Recent phylogenomic studies proposed that Chiroptera is sister to a clade containing Carnivora and Perissodactyla (McCormack et al. 2012), to a clade comprising Carnivora, Cetartiodactyla, and Perissodactyla (Hallström et al. 2011; Song et al. 2012; Zhou et al. 2012; Tsagkogeorga et al. 2013; Hallström et al. 2011), or to Cetartiodactyla (Nery et al. 2012; Romiguier et al. 2013). With respect to the phylogenetic placement of Perissodactyla, several competing hypotheses exist, such as Zooamata (Perissodactyla + Carnivora) (Nery et al. 2012; Song et al. 2012; Romiguier et al. 2013) or Euungulata (Perissodactyla + Cetartiodactyla) (Hou et al. 2009; Zhou et al. 2012; Tsagkogeorga et al. 2013). These controversies are reflected in our analyses of CDS data as highly supported, but conflicting results are often observed when analyzing different data subsets (figs. 4 and 5). Our gene tree statistics show that the CDS data set contains few genes with strong phylogenetic signal bearing on the divergence of laurasiatherian mammal orders (fig. 6). This provides an explanation of why it is difficult to resolve the Laurasiatheria phylogeny with CDS, even when using thousands of genes.

Our study represented the first attempt to resolve Laurasiatheria relationships with genome-scale intronic sequences. Our analyses showed that intron loci are more homogeneous in gene trees than coding loci (fig. 1) and contain more genes with strong phylogenetic signals bearing on the divergence of laurasiatherian mammal orders (fig. 6). In contrast to unstable phylogenies inferred from CDS, our intron data set provided overwhelming evidence for a clade that unites the phenotypically divergent Chiroptera and Perissodactyla and an evolutionary affinity between Cetartiodactyla and Carnivora (bootstrap = 100%; fig. 3). Actually, the topology is not entirely novel. A clade comprising Chiroptera and Perissodactyla was recovered by some of previous studies (McCormack et al. 2012; Zhang et al. 2013; Tarver et al. 2016). It is noteworthy that two of them have

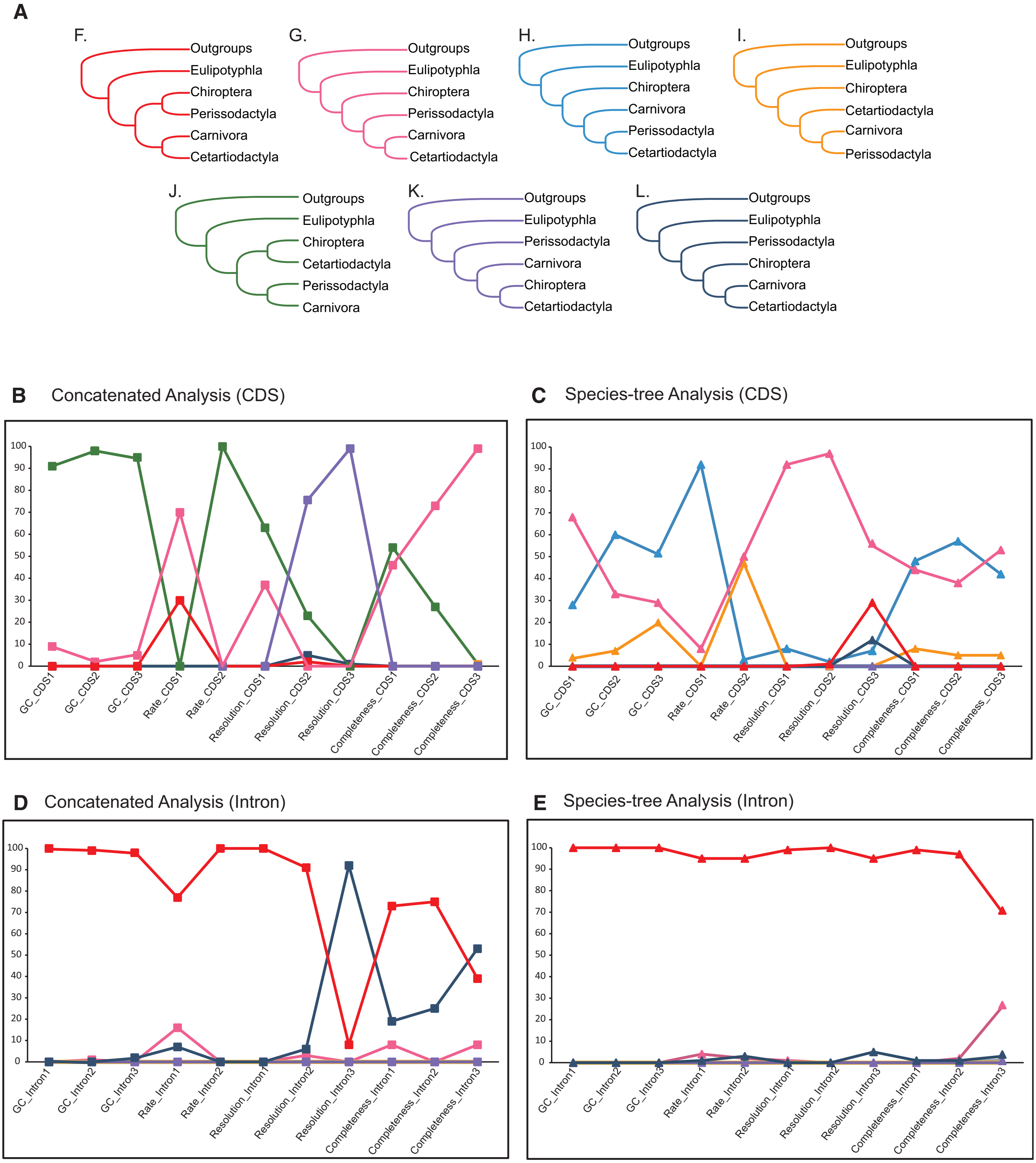
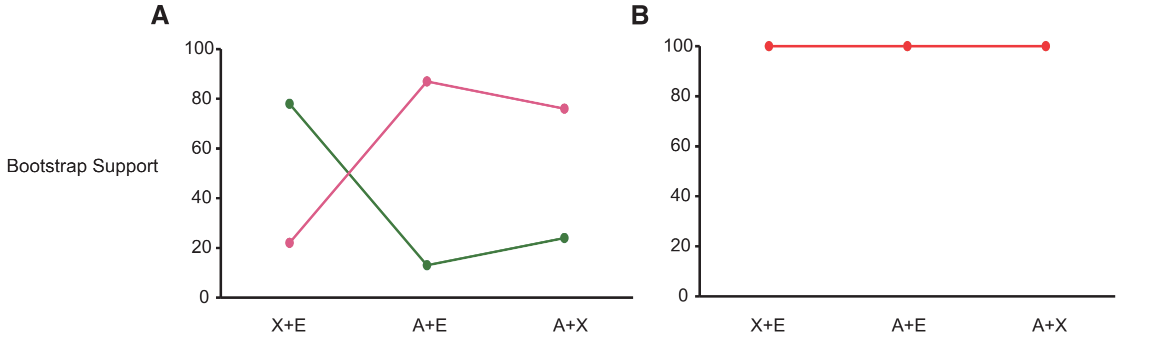


Fig. 4.—Phylogenetic inference robustness for the CDS and Intron data sets, which were resampled into 11 data subsets under different data subsampling criteria (see Materials and Methods for details). These data subsets were analyzed with both concatenated and species-tree inferences. (*A*) There are in total seven topologies found in these phylogenetic analyses (each color represents a specific tree topology). Bootstrap support for certain topologies from different data subsets are shown in charts (*B*) through (*E*).



	X+E	A+E	A+X	X+E	A+E	A+X
● ((Chiro,Certartio),(Perisso,Carni)	78	13	24	0	0	0
● (Chiro,(Perisso,(Carni,Cetartio)))	22	87	76	0	0	0
● ((Chiro,Perisso),(Carni,Cetartio))	0	0	0	100	100	100

Fig. 5.—Effect of outgroup choices on phylogenetic inferences of the CDS data set (*A*) and the Intron data set (*B*). There are three outgroup combination schemes: "X + E" refers to the use of Xenarthra and Euarchontoglires as outgroup; "A + E" refers to the use of Afrotheria and Euarchontoglires as outgroup; "A + X" refers to the use of Afrotheria and Xenarthra as outgroup. Note that the change in outgroup has no effect on the phylogenetic inference of the Intron data set but can influence the phylogenetic inference of the CDS data set. Chiro: Chiroptera, Cetartio: Cetartiodactyla, Perisso: Perissodactyla, Carni: Carnivora.

used noncoding data (683 ultraconserved element loci in McCormack et al. 2012 and 239 noncoding RNA miRNAs in Tarver et al. 2016). Although we have obtained a strongly supported topology, it does not necessarily indicate that the phylogenetic inference is correct. The phylogenetic analysis of genome-scale data sets requires particular attention to potential systematic biases, such as LBA and compositional bias. In particular, an LBA artifact may potentially occur since noncoding sequences generally have fast evolutionary rates. If LBA does influence the phylogenetic inference of the Intron data set, we should obtain different topologies or observe a significant decrease in branch support when removing fast-evolving loci. In fact, both topology and branch support remained stable when we separately analyzed the top half of the intron loci (ranked by evolutionary rate) and the bottom half of intron loci (fig. 4*D*). This demonstrated that LBA should not be responsible for the inferred grouping of Chiroptera and Perissodactyla. In addition, compositional bias did not significantly influence phylogenomic inference with our intron data set because using data subsets with different levels of GC content had almost no effect on the resulting topology and branch support (fig. 4*D*). Therefore, the strongly supported tree recovered from our intron data set cannot be explained by any kind of identifiable systematic bias (LBA and compositional bias) and should constitute the best current hypothesis for the Laurasiatheria phylogeny.

Our results prompt a reinterpretation of morphological data in laurasiatherian phylogeny. In particular, the close proximity between Perissodactyla and Chiroptera but not Cetartiodactyla suggests that extremely quick morphological evolution and extensive morphological homoplasy occurred in the early history of laurasiatherian mammals. Nevertheless, by focusing on genome-scale data, our data sets contained many more loci than species. The taxon sampling for Laurasiatheria in this study is still insufficient; only two species are sampled from each of the two key orders (Chiroptera and Perissodactyla), and data from the order Pholidota is lacking. Poor taxon sampling can interfere phylogenetic inference (Philippe et al. 2011). Therefore, the current hypothesis for Laurasiatheria phylogeny still needs further validation when more genome data for Perissodactyla and Pholidota are available.

Introns Are Promising Genomic Resource to Study Difficult Radiation Problems

With the advance of sequencing technology, whole genome data can be gathered very rapidly. In recent years, genome data have frequently been used to address difficult phylogenetic problems because massive amounts of data often mean stronger phylogenetic resolving power. It is worth noting that the vast majority of phylogenomic studies based on genome data use coding regions as their data resource. CDS have several features suitable for phylogenetic analyses, such as an appropriate level of variation, easy alignment across a large phylogenetic span, and relatively straightforward identification of orthologs.

Compared with CDS, noncoding regions of genomes have received little attention in recent phylogenomic studies. This may be because noncoding sequences are highly variable, which makes it difficult to identify their orthologs across divergent taxonomic groups and accurately align them. Although not easy to use, noncoding regions of genomes have several positive features that make them more suitable

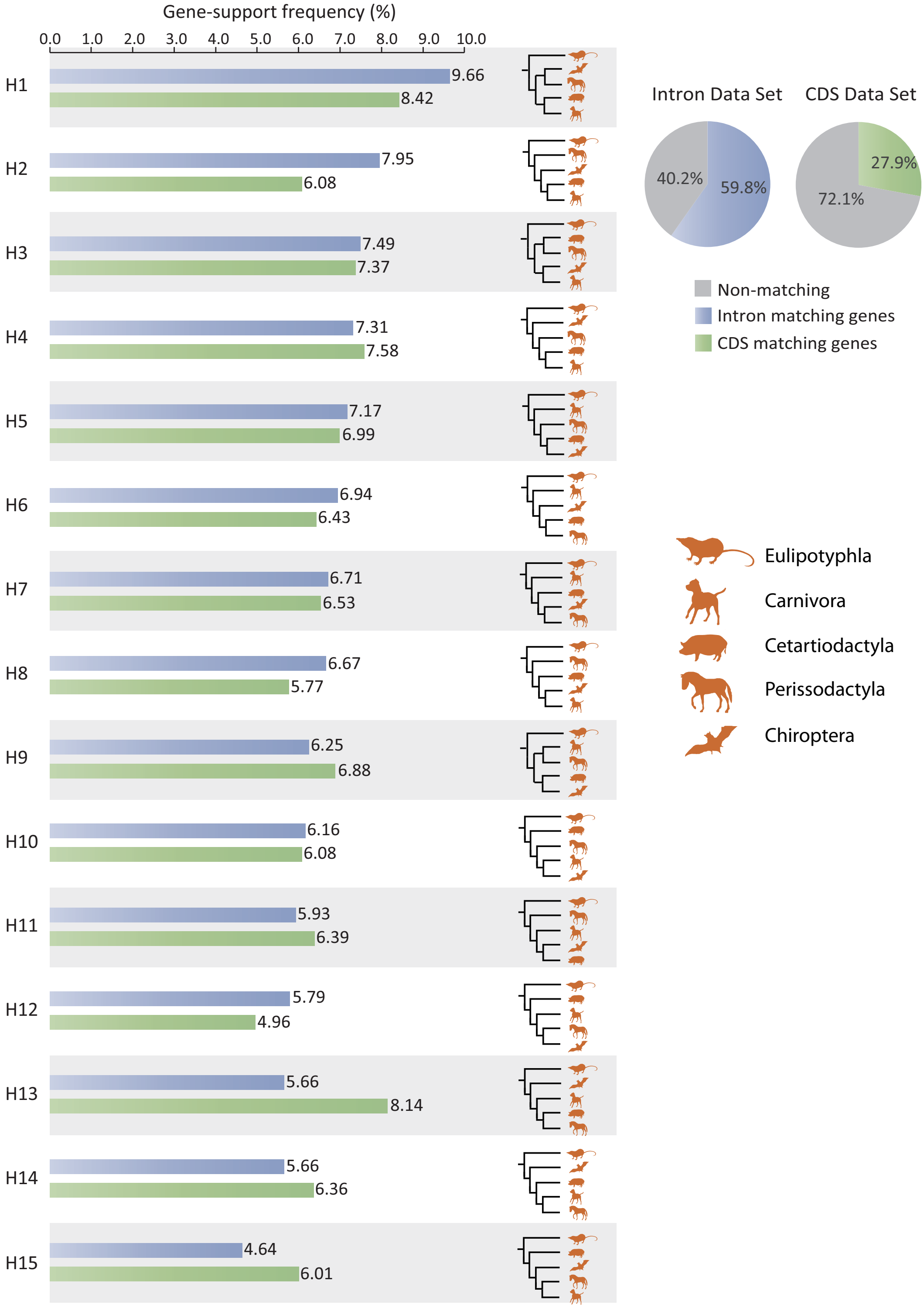


Fig. 6.—Gene-support frequency statistics for the 15 alternative hypotheses (H1–H15) regarding the interrelationships of Chiroptera, Perissodactyla, Carnivora, and Cetartiodactyla. Intron data are displayed in blue, and CDS data are displayed in green. Genes whose gene trees do not support any of the 15 alternative hypotheses are considered "nonmatching." Gene-tree statistics are based on "matching" genes only. The histograms on the left show the proportion of gene trees that support a given hypothesis.

for studying rapid evolutionary events: 1) they generally have fast evolutionary rates and can thus accumulate more mutations in a short time, which is necessary to efficiently resolve closely spaced diversification events; 2) they have a low level of functional constraint, which makes them much less vulnerable to convergent evolution; and 3) they generally have low GC content, which leads to a low frequency of recombination.

Recently, Reddy et al. (https://doi.org/10.1093/sysbio/syx041) compared noncoding versus CDS to resolve the phylogenetic relationships among Neoaves birds. Similar to our findings, they found that data type (coding or noncoding) was the main factor influencing the topological incongruence observed between different data sets. They pointed out that most data-type effects actually reflect poor model fit to the data: patterns of sequence evolution for noncoding regions are less complex than for coding regions so that standard substitution models (i.e., $\mathrm{GTR}+\mathrm{I}+\Gamma$ and its submodels) are likely to fit noncoding regions better than coding regions. Our results of data characteristics showed that CDS have much higher gene-tree heterogeneity than noncoding sequences (fig. 1*E*–*G*). This phenomenon cannot be solely explained by biological factors such as incomplete lineage sorting (ILS) and should partly attribute to poor model fit to the CDS data. Tarver et al. (2016) have recently illustrated that using the CAT-GTR + G site-heterogeneous mixture model (Lartillot et al. 2013) was more advantageous than the standard GTR + G+I model in analyzing heterogeneous mammalian phylogenomic data. However, the CAT-GTR + G model is too computationally intensive for our huge CDS and intron data sets. Therefore, for both the CDS and intron data, we tentatively built two smaller data sets that contain two million sites randomly sampled from both their fastest evolving genes and slowest genes. All the four CAT model analyses produced complete resolved phylogenies with all posterior probabilities equal to one (supplementary fig. S11, Supplementary Material online). However, the incongruences between the two rate subsamples of the CDS data still existed while the two rate subsamples of the intron data produced identical topology to that from the whole intron data set. This result validated the robustness of the intron phylogeny but further showed that the Laurasiatheria CDS data is so complex that even the most sophisticated CAT model is unable to adequately model its evolutionary process. If the models currently in use do offer a better fit to noncoding sequences than to CDS, it is thus more reasonable to use noncoding sequences with standard analytical methods (i.e., those using $\mathrm{GTR}+\mathrm{I}+\Gamma$ and its submodels) to address difficult evolutionary questions.

Among the many types of noncoding sequences, introns may be the most suitable for phylogenetic analyses. Structurally, introns are surrounded by conserved exons on both sides. The orthology of an intron can thus be determined based on the orthology of its two flanking exons, as in this study. The two flanking exon sequences of an intron can also provide anchoring sites, which help to more accurately align the variable intron sequences. Moreover, in recent years, a number of new aligning methods have been developed to efficiently align a large number of highly divergent sequences, such as SATé (Liu et al. 2009, 2012) and PASTA (Nguyen et al. 2015). These further overcome the difficulty of aligning highly variable intron sequences.

In fact, intron data sets with tens of genes have been successfully used to improve the resolution of intractable rapid radiation events (Chojnowski et al. 2008; Yu et al. 2011; Foley et al. 2015; Dool et al. 2016), but few studies have tried to mine intron sequences from genome data for phylogenomic inference (Jarvis et al. 2014). Our study, for the first time, used a genome-scale intron data set to study the Laurasiatheria phylogeny, a relatively ancient rapid radiation event. Our analyses showed that intron data contain substantial and homogeneous phylogenetic signals that are able to robustly resolve the deep relationships of Laurasiatherian mammals, whereas CDS data contain highly heterogeneous signals. This finding is encouraging because it reveals that intron sequences within whole genome data have great potential to resolve difficult phylogenetic problems. Currently, there are still many rapid radiation events that cannot be fully resolved by CDS; we propose that genome-scale intron analyses should be performed to provide new perspectives on this longstanding controversy.

Conclusions

In this study, we reconstructed the interordinal relationships of Laurasiatherian mammals with two genome-scale data sets of CDS and noncoding intron sequences. Our phylogenetic analyses based on the intron data recovered a robust phylogeny: Chiroptera and Perissodactyla formed a well-supported clade that is sister to the clade comprising Cetartiodactyla and Carnivora. Although this phylogeny was not recovered by the CDS data, we found that the major phylogenetic signal of the CDS data is actually congruent with the intron data. By comparing the phylogenetic signal strength and phylogenetic inference robustness, we found that noncoding intron sequences outperform CDS in resolving the Laurasiatheria phylogeny. Our study showed that building genome-scale intron data sets may be an efficient way to resolve challenging short internal nodes in phylogenetic trees.

Note Added in Proof

Although this paper was in review following revision, the pangolin genome (Tan et al. 2016) was released and cleared its 12-month embargo period. This gives us an opportunity to incorporate the pangolin data into our analyses. We therefore reran the concatenated RAxML and coalescent-based ARSTAL analyses for the complete CDS and intron data sets with the newly added pangolin data. However, due to the limited time, we did not redo the data subsampling analyses. The new CDS and intron analyses confirmed the sisterhood between Pholidota and Carnivora with maximal branch support (supplementary fig. S12, Supplementary Material online). Both analyses of the new intron data set still strongly supported the previous intron phylogeny (supplementary fig. S12B, Supplementary Material online) but the new CDS data sets again produced incongruent results in the concatenated and coalescent-based analyses (supplementary fig. S12A, Supplementary Material online). The results of this reanalysis had no material effect on the conclusions of this study.

Data Availability

All data sets, analysis results, and supplementary material are available on FigShare Repository (https://figshare.com/s/a8cea06c05465c939e15).

Supplementary Material

Supplementary figures are available at *Genome Biology and Evolution* online.

Acknowledgments

We thank all members of Zhang's lab and Dr Xing-Xing Shen for helpful discussion. This work was supported by National Natural Science Foundation of China (grants No. 31672266 and 31372172), the National Youth Talent Support Program to P.Z. (W02070133), the National Science Fund for Excellent Young Scholars of China to P.Z. (31322049) and the National Natural Science Foundation of China (No. 31601847), the Natural Science Foundation of Guangdong Province (No. 2015A030313142) to D.L.

Literature Cited

Amemiya CT, et al. 2013. The African coelacanth genome provides insights into tetrapod evolution. Nature 496(7445):311–316.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol. 17(4):540–552.

Chen MY, Liang D, Zhang P. 2015. Selecting question-specific genes to reduce incongruence in phylogenomics: a case study of jawed vertebrate backbone phylogeny. Syst Biol. 64(6):1104–1120.

Chojnowski JL, Kimball RT, Braun EL. 2008. Introns outperform exons in analyses of basal avian phylogeny using clathrin heavy chain genes. Gene 410(1):89–96.

Dool SE, et al. 2016. Nuclear introns outperform mitochondrial DNA in phylogenetic reconstruction: lessons from horseshoe bats (Rhinolophidae: Chiroptera). Mol Phylogenet Evol. 97:196–212.

dos Reis M, et al. 2012. Phylogenomic datasets provide both precision and accuracy in estimating the timescale of placental mammal phylogeny. Proc Biol Sci. 279:3491–3500.

Douzery EJ, et al. 2014. OrthoMaM v8: a database of orthologous exons and coding sequences for comparative genomics in mammals. Mol Biol Evol 31(7):1923–1928.

Edwards SV. 2016. Phylogenomic subsampling: a brief review. Zool Scripta 45:63–74.

Foley NM, et al. 2015. How and why overcome the impediments to resolution: lessons from rhinolophid and hipposiderid bats. Mol Biol Evol. 32(2):313–333.

Gori K, Suchan T, Alvarez N, Goldman N, Dessimoz C. 2016. Clustering genes of common evolutionary history. Mol Biol Evol. 33(6):1590–1605.

Hallström BM, Janke A. 2010. Mammalian evolution may not be strictly bifurcating. Mol Biol Evol. 27(12):2804–2816.

Hallström BM, Schneider A, Zoller S, Janke A. 2011. A genomic approach to examine the complex evolution of laurasiatherian mammals. PLoS One 6(12): e28199.

Hillis DM, Heath TA, St John K. 2005. Analysis and visualization of tree space. Syst Biol. 54(3):471–482.

Hou ZC, Romero R, Wildman DE. 2009. Phylogeny of the Ferungulata (Mammalia: Laurasiatheria) as determined from phylogenomic data. Mol Phylogenet Evol. 52(3):660–664.

Hu J, Zhang YP, Yu L. 2013. Summary of laurasiatheria (mammalia) phylogeny. Zool Res. 33(6):65–74.

Irisarri I, Meyer A. 2016. The identification of the closest living relative(s) of tetrapods: phylogenomic lessons for resolving short ancient internodes. Syst Biol. 65(6):1057–1075.

Jarvis ED, et al. 2014. Whole-genome analyses resolve early branches in the tree of life of modern birds. Science 346(6215):1320–1331.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 30(4):772–780.

Lanfear R, Calcott B, Ho SYW, Guindon S. 2012. PartitionFinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. Mol Biol Evol. 29(6):1695–1701.

Lartillot N, Rodrigue N, Stubbs D, Richer J. 2013. PhyloBayes MPI. Phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. Syst Biol. 62(4):611–615.

Lemmon EM, Lemmon AR. 2013. High-throughput genomic data in systematics and phylogenetics. Annu Rev Ecol Evol Syst. 44(1):99–121.

Liu K, Raghavan S, Nelesen S, Linder CR, Warnow T. 2009. Rapid and accurate large-scale coestimation of sequence alignments and phylogenetic trees. Science 324(5934):1561–1564.

Liu K, et al. 2012. SATe-II: very fast and accurate simultaneous estimation of multiple sequence alignments and phylogenetic trees. Syst Biol. 61(1):90–106.

McCormack JE, et al. 2012. Ultraconserved elements are novel phylogenomic markers that resolve placental mammal phylogeny when combined with species-tree analysis. Genome Res. 22(4):746–754.

Meredith RW, et al. 2011. Impacts of the cretaceous terrestrial revolution and KPg extinction on mammal diversification. Science 334(6055):521–524.

Mirarab S, Warnow T. 2015. ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. Bioinformatics 31(12):44–52.

Morgan CC, et al. 2013. Heterogeneous models place the root of the placental mammal phylogeny. Mol Biol Evol. 30(9):2145–2156.

Murphy WJ, Eizirik E, O'Brien SJ, et al. 2001. Resolution of the early placental mammal radiation using Bayesian phylogenetics. Science 294(5550):2348–2351.

Narechania A, et al. 2012. Random addition concatenation analysis: a novel approach to the exploration of phylogenomic signal reveals strong agreement between core and shell genomic partitions in the cyanobacteria. Genome Biol Evol. 4:30–43.

Nery MF, González DJ, Hoffmann FG, Opazo JC. 2012. Resolution of the laurasiatherian phylogeny: evidence from genomic data. Mol Phylogenet Evol. 64(3):685–689.

Nguyen N, Mirarab S, Kumar K, Warnow T. 2015. Ultra-large alignments using phylogeny-aware profiles. Genome Biol. 22:377–386.

Nikolaev S, et al. 2007. Early history of mammals is elucidated with the ENCODE multiple species sequencing data. PLoS Genet. 3(1): e2.

O'Leary MA, Bloch JI, Flynn JJ, et al. 2013. The placental mammal ancestor and the post-K-Pg radiation of placentals. Science 339:662–667.

Philippe H, et al. 2011. Resolving difficult phylogenetic questions: why more sequences are not enough. PLoS Biol. 9(3): e1000602.

Prasad AB, Allard MW, Green ED, Program NCS. 2008. Confirming the phylogeny of mammals by use of large comparative sequence data sets. Mol Biol Evol. 25(9):1795–1808.

Prum RO, et al. 2015. A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing. Nature 526(7574):569–578.

Reddy S, et al. 2017. Why do phylogenomic datasets yield conflicting trees? Data type influences the avian tree of life more than taxon sampling. Syst Biol. doi: https://doi.org/10.1093/sysbio/syx041.

Robinson D, Foulds LR. 1981. Comparison of phylogenetic trees. Math Biosci. 53(1–2):131–147.

Romiguier J, Ranwez V, Delsuc F, Galtier N, Douzery EJP. 2013. Less is more in mammalian phylogenomics: AT-rich genes minimize tree conflicts and unravel the root of placental mammals. Mol Biol Evol. 30(9):2134–2144.

Salichos L, Rokas A. 2013. Inferring ancient divergences requires genes with strong phylogenetic signals. Nature 497(7449):327–331.

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(37):14942–14947.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30(9):1312–1313.

Tan TK, Tan KY, Hari R, et al. 2016. PGD: a pangolin genome hub for the research community. Database (Oxford). 2016:baw063.

Tarver JE, et al. 2016. The interrelationships of placental mammals and the limits of phylogenetic inference. Genome Biol Evol. 8(2):330–344.

Thomson RC, Wang IJ, Johnson JR. 2010. Genome-enabled development of DNA markers for ecology, evolution and conservation. Mol Ecol. 19(11):2184–2195.

Tsagkogeorga G, Parker J, Stupka E, Cotton JA, Rossiter SJ. 2013. Phylogenomic analyses elucidate the evolutionary relationships of bats. Curr Biol. 23(22):2262–2267.

Yu L, et al. 2011. Phylogenetic utility of nuclear introns in interfamilial relationships of Caniformia (order Carnivora). Syst Biol. 60(2):175–187.

Zhang G, et al. 2013. Comparative analysis of bat genomes provides insight into the evolution of flight and immunity. Science 339(6118):456–460.

Zhou X, et al. 2012. Phylogenomic analysis resolves the interordinal relationships and rapid diversification of the Laurasiatherian mammals. Syst Biol. 61(1):150–164.

Associate editor: Davide Pisani