American Genetic Association

Journal of Heredity, 2018, 297–307
doi:10.1093/jhered/esx089
Original Article
Advance Access publication 25 October 2017

OXFORD

Original Article

Pinniped Diphyly and Bat Triphyly: More Homology Errors Drive Conflicts in the Mammalian Tree

Mark S. Springer and John Gatesy

From the Department of Evolution, Ecology, and Organismal Biology, University of California, Riverside, CA 92521 (Springer); Division of Vertebrate Zoology, American Museum of Natural History, New York, NY 10024 (Gatesy).

Address correspondence to Mark S. Springer or J. Gatesy at the above addresses, or e-mail: springer@ucr.edu (Springer); jgatesy@amnh.org (Gatesy).

Received August 21, 2017; First decision September 24, 2017; Accepted October 7, 2017.

Corresponding editor: Alfred Roca

Abstract

Homology is perhaps the most central concept of phylogenetic biology. At difficult to resolve polytomies that are deep in the Tree of Life, a few homology errors in phylogenomic data can drive spurious phylogenetic results. Feijoo and Parada (2017) assembled three phylogenomic data sets for mammals and reported methodological discrepancies and unexpected results that contradict the monophyly of well-established clades in Pinnipedia and Yangochiroptera. Examination of Feijoo and Parada's (2017) data sets reveals extensive homology errors (paralogous sequences, alignments of different exons to each other) and cross-contamination of sequences from different species. These problems predictably result in distorted estimates of gene trees, species trees, bootstrap support, and branch lengths. Correction of these errors resulted in robust support for conventional relationships in Pinnipedia and Yangochiroptera. Phylogenomic data sets are not immune to the problems of homology errors in sequence alignments. Rather, sequence alignments underlie all inferences in molecular phylogenetics and evolution and should be spot-checked for obvious errors via manual inspection of alignments and gene trees.

Key words: Chiroptera, homology, Mammalia, phylogenomics, Pinnipedia.

Introduction

Phylogenomic data sets have become increasingly popular for addressing questions in higher-level mammalian systematics (dos Reis et al. 2012; McCormack et al. 2012; Shaw et al. 2012; Song et al. 2012; Kumar et al. 2013; Romiguier et al. 2013; Mason et al. 2016; Tarver et al. 2016). By contrast with studies based on one or a few genes that have limited statistical power for resolving short internodes in the Tree of Life, phylogenomic studies have the potential to reduce stochastic errors and provide robust resolution of challenging polytomies (Rokas et al. 2003; Delsuc et al. 2006). However, phylogenomic data have their own attendant problems. Importantly, different data sets and/or different analyses of the same phylogenomic data set sometimes result in well-supported topological conflicts (e.g., Chiari et al. 2012; Simmons and Gatesy 2015). Multiple factors can drive robust incongruence in genome-scale phylogenetic studies. Concatenation (= supermatrix) methods can fail to reconstruct an accurate species tree due to incomplete lineage sorting (ILS) (Degnan and Rosenberg 2006; Rosenberg and Tao 2008), incorrect specification of the substitution model(s) (Sullivan and Swofford 1997), long-branch misplacement (Bergsten 2005), and/or heterotachy (Kolaczkowski and Thornton 2004; Philippe et al. 2005). Coalescence methods, in turn, may fail if underlying

assumptions (e.g., all gene tree heterogeneity results from ILS, no recombination within coalescence genes, no selection) are violated (Patel et al. 2013). Concatenation and coalescence methods can both fail if sequence alignments are sufficiently poor, because such alignments are the basic data that underlie inference of gene trees as well as estimation of the species tree (de Queiroz and Gatesy 2007; Edwards 2009).

In the early days of molecular systematics when data acquisition was based on PCR and Sanger sequencing, it was standard practice to manually screen out paralogous loci and to curate alignments of orthologous gene sequences "by eye" (Philippe et al. 2017). However, with the advent of high-throughput sequencing, researchers commonly trust computer scripts to annotate genes, process genomic sequences, and output huge data sets comprised of putative one-to-one orthologs from genomic assemblies or short-read data from RNA-Seq libraries. This task is especially challenging for protein-coding sequences that are comprised of discontiguous exons that are separated by introns, especially when exons are very short and introns are very long. A reliable automated pipeline for delivering high-quality alignments of protein-coding sequences is a worthwhile goal, but given extensive divergence among genomes, alternate splicing of exons, and incomplete genomic assemblies, an automated approach can yield extensive errors. Cross-contamination, editing errors, paralogous sequences, alignments of different exons to each other, and alignments of exons to introns have been identified in various published phylogenomic data sets composed of protein-coding genes (Springer and Gatesy 2016; Brown and Thomson 2017; Philippe et al. 2017; Shen et al. 2017). These problems can be solved by first searching for correct one-to-one orthologous sequences in genome assemblies or Sequence Read Archive (SRA) databases and then fixing the initial, automated assembly of the phylogenomic data set. Such "homology errors" are not due to local shifts in positional homology, but instead are driven by including segments of sequence that are not orthologous, by excluding segments of orthologous sequence from some species, by not discerning paralogs, or by cross-contamination of sequences from different species.

We have previously called attention to problems with Song et al.'s (2012) phylogenomic data set (447 loci, 36 mammals), including 26 loci for which nonhomologous exons (and sometimes introns) were aligned against each other (Springer and Gatesy 2016). This data set is based on protein-coding sequences that were downloaded from the OrthoMam v6 database (Ranwez et al. 2007) and then aligned with *Gallus gallus* (chicken) out-group sequences (Song et al. 2012). Despite the fact that key relationships and branch support values were impacted by homology problems in this data set (Springer and Gatesy 2016), Edwards et al. (2016) argued that manual curation of alignments is not sustainable in the phylogenomics era and Edwards (2016) has continued to utilize a version of Song et al.'s (2012) data set that is replete with homology errors.

Recently, Feijoo and Parada (2017) built directly on the phylogenomic groundwork of Song et al. (2012), Edwards et al. (2016), and Edwards (2016) by utilizing publicly available sequence data to help resolve difficult polytomies in higher-level mammalian phylogenetics. Feijoo and Parada (2017) employed 113 of 447 loci (protein-coding sequences) from Song et al. (2012) for which 40 mammalian taxa were available in OrthoMam v8 (Douzery et al. 2014). Sequences for 19 additional taxa were extracted from SRA-NCBI public databases. Following steps for filtering, trimming, assembling, and annotating (RNA-Seq reads) or mapping directly to reference genomes (genomic DNA short reads), the newly organized data were assembled into three phylogenomic matrices. The primary data set included 95 of the 113 loci (47 taxa, 214,822 bp) and covered a broad diversity of Mammalia. The second included expanded taxonomic coverage for arctoid carnivorans (35 taxa [7 arctoids], 29 genes, 23,495 bp), and the third focused on Chiroptera (bats) (42 taxa [14 bats], 18 genes, 44,711 bp).

Feijoo and Parada (2017) analyzed the three data sets with both concatenation and coalescence methods. However, their application of coalescence methods to complete protein-coding sequences, which are stitched together across regions of the genome that can exceed the actual size of coalescence genes by several orders of magnitude, is best described as a hybrid method, concatalescence, that ignores the fundamental rationale for employing coalescence methods in the first place (Gatesy and Springer 2013, 2014; Springer and Gatesy 2014, 2016). Specifically, coalescence methods assume that recombination occurs between individual loci but not within loci. However, this assumption is not reasonable for complete coding sequences. A recent empirical study on ILS in placental mammals supports the view that exons from the same gene do not necessarily share the same genealogy and therefore should not be merged in coalescence analyses (Scornavacca and Galtier 2017). This criticism aside, Feijoo and Parada's (2017) resulting species trees are generally similar, but they highlighted several notable exceptions. For the Arctoidea dataset (Figure 1), five arctoid families were sampled: Ursidae (bears), Mustelidae (weasels), Phocidae (seals), Otariidae (sea lions), and Odobenidae (walrus). Within Pinnipedia (Phocidae + Otariidae + Odobenidae), all three coalescence methods (ASTRAL, STAR, MP-EST) recovered the traditional sister-group relationship between Otariidae and Odobenidae (= Otaroidea) with high bootstrap support (96–100%), whereas both analyses of the concatenated data set (IQ-TREE, BEAST2) recovered a robustly supported Otariidae + Phocidae clade (100% bootstrap, 1.0 posterior probability). This latter result that was obtained with concatenation methods (Figure 1A) is overwhelmingly contradicted by previous molecular studies that instead support Odobenidae + Otariidae (Flynn et al. 2005; Arnason et al. 2006; Meredith et al. 2011; Luan et al. 2013; Doronina et al. 2015; Foley et al. 2016), and implies that supermatrix analyses are flawed, while coalescence analyses are not (Feijoo and Parada 2017). For the Chiroptera data set (Figure 2), the three coalescence methods (ASTRAL, STAR, MP-EST) weakly supported a sister-group relationship between *Tadarida* (Molossidae) and *Taphozous* (Emballonuridae) (their Supplementary Figure S1), whereas the ML concatenation analysis (IQ-TREE) favored *Myotis* (Vespertilionidae) sister to *Tadarida* (Molossidae) with 89% bootstrap support. Of these alternatives, an association of Vespertilionidae and Molossidae to the exclusion of Emballonuridae has received overwhelming and consistent support in previous work on bat phylogeny (Eick et al., 2005; Teeling et al. 2005, 2012; Lack et al., 2010; Meredith et al. 2011; Amador et al. 2016; Foley et al. 2016), which suggests that coalescence methods are inadequate for resolving this trichotomy given the number of loci that were sampled. However, Feijoo and Parada's (2017) Bayesian supermatrix analysis (BEAST2) strongly supported (1.0 posterior probability) the unconventional *Tadarida* (Molossidae) + *Taphozous* (Emballonuridae) clade favored by coalescence methods (Figure 2A). This is a striking conflict with Feijoo and Parada's (2017) reported ML supermatrix analysis of the same data set and a host of published work. As for relationships within Pinnipedia, Feijoo and Parada's (2017) conflicting phylogenomic results for bats are perplexing and invite further investigation.

Here, we reexamine Feijoo and Parada (2017) and address two key questions pertaining to data set fidelity/quality. First, are

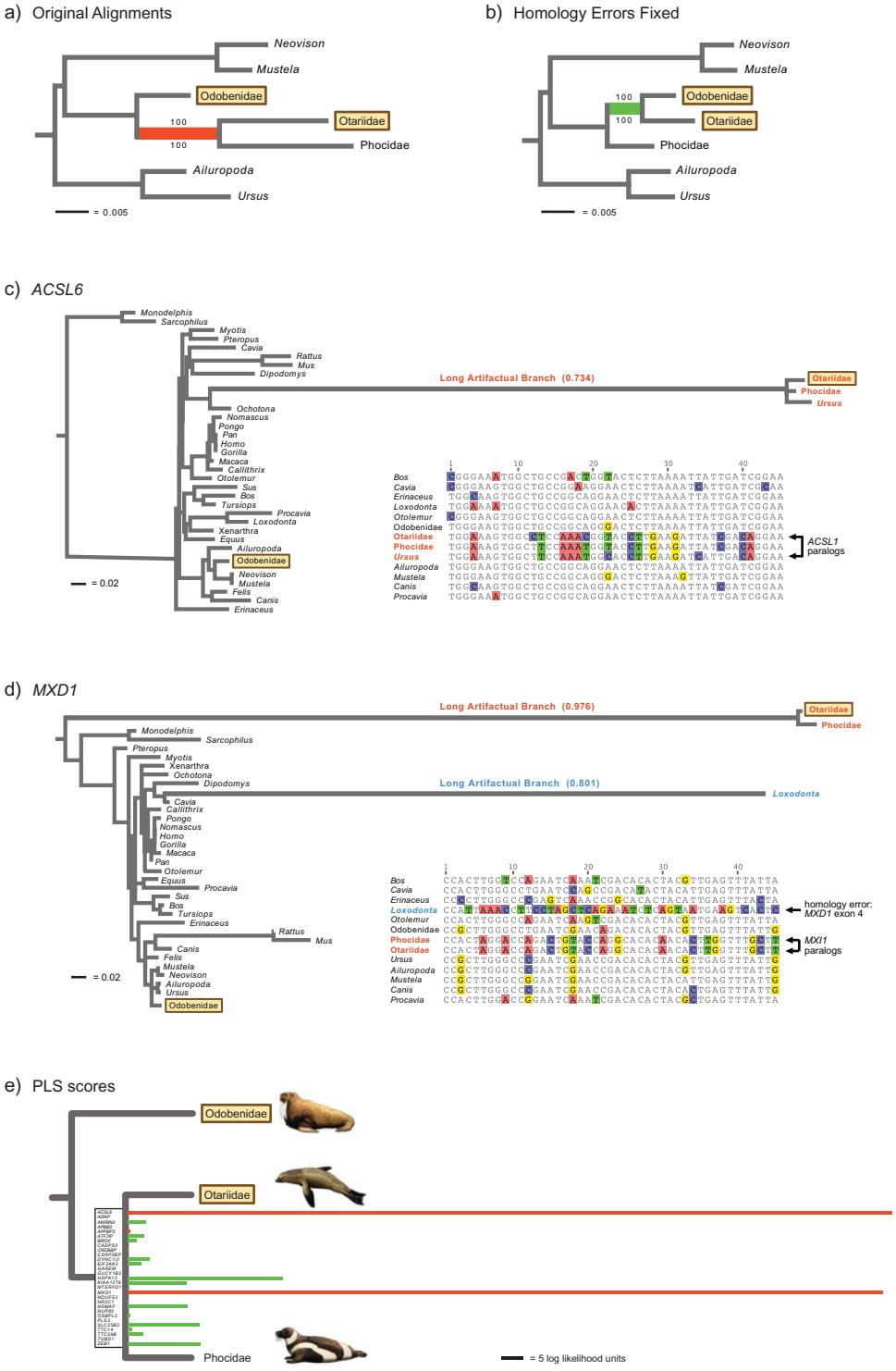

Figure 1. Species trees, homology errors, gene trees, and partitioned likelihood support (PLS) scores for the Arctoidea data set. Maximum likelihood (ML) concatenated analyses of Feijoo and Parada's (2017) (a) original data set and (b) a corrected version of their data set (see Table 1) show contrasting relationships within Pinnipedia and large differences in support and branch lengths. Bootstrap percentages for RAxML (above) and IQ-TREE (below) are shown at the conflicting internodes. In (c) and (d), gene trees and partial sequence alignments are illustrated for two alignments (*ACSL6, MXD1*) that are characterized by paralogy errors. Taxa with paralogous sequences (red type) cluster in each gene tree. For *MXD1*, there also is a homology error in exon 4 of *Loxodonta* (African elephant; blue) that results in a long terminal branch for this species. In (e), partitioned likelihood support for Otariidae + Phocidae shows extreme values for *ACSL6* and *MXD1* (red bars) relative to the more widespread but moderate support for the traditional clade, Otariidae + Odobenidae, among many loci (green bars). Paintings by Carl Buell. See online version for full colors.

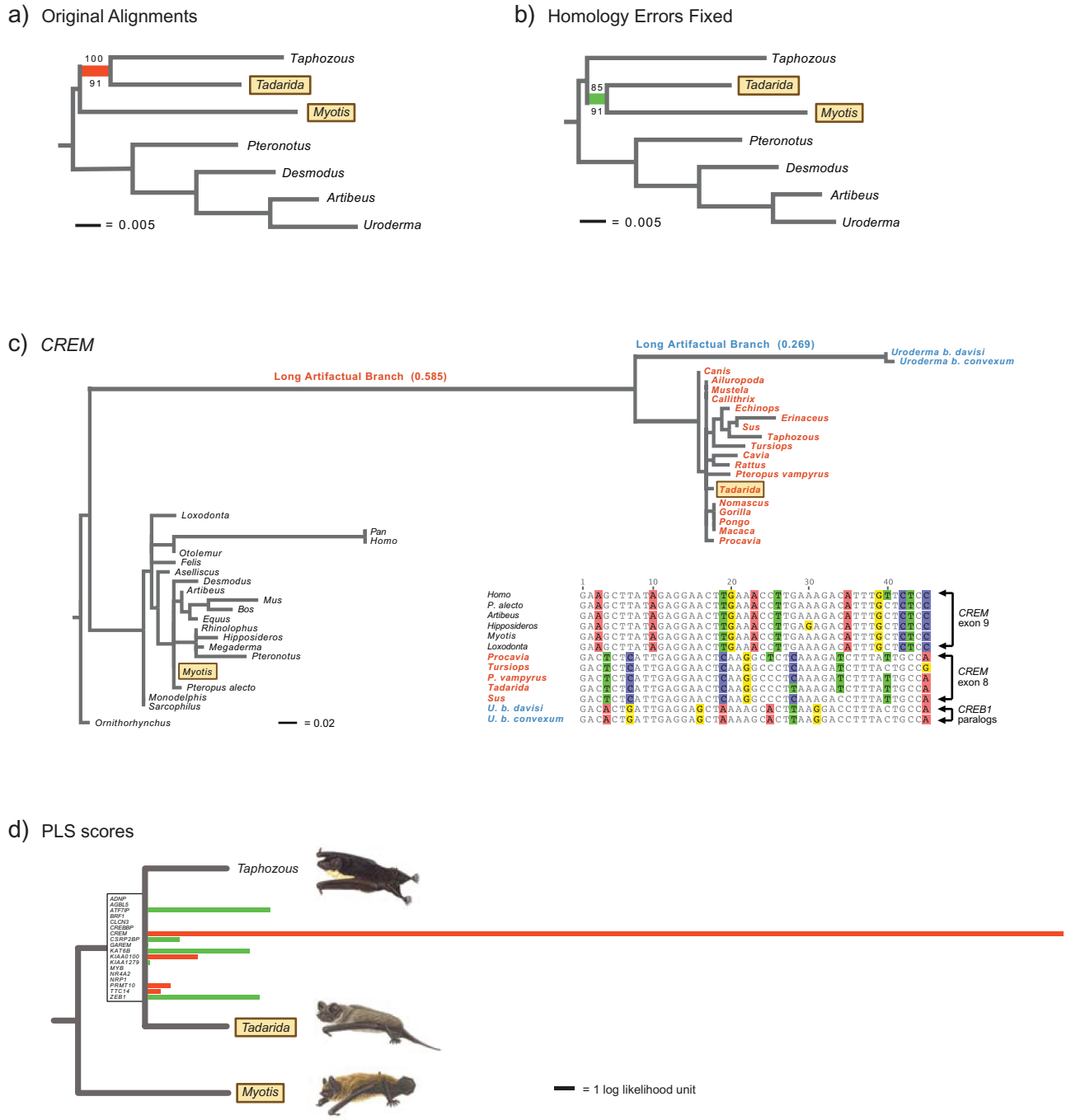

Figure 2. Species trees, homology errors, gene tree, and partitioned likelihood support (PLS) scores for the Chiroptera data set. Maximum likelihood (ML) concatenated analyses of Feijoo and Parada's (2017) (a) original data set and (b) a corrected version of their data set (see Table 1) show contrasting relationships in the Yangochiroptera subtree of the phylogeny. Bootstrap percentages for RAxML (above) and IQ-TREE (below) are shown at the conflicting internodes. Note that our results for IQ-TREE analysis of the original bat data set are shown; for the same concatenated matrix, Feijoo and Parada (2017) reported an IQ-TREE topology that instead supported *Myotis* + *Tadarida* with 89% bootstrap support. In (c), an ML gene tree and partial sequence alignment for the *CREM* gene are shown. Taxa with paralogous *CREB1* sequences (blue type) cluster at the end of a long artifactual branch. For the *CREM* alignment, exon 8 for some taxa (red type) is aligned against exon 9 of other taxa, which results in another long artifactual branch. In (d), partitioned likelihood support scores for *Taphozous* + *Tadarida* (red bars) include an extreme value for *CREM* relative to more moderate support for the traditional clade, *Myotis* + *Tadarida* (green bars). Paintings by Fiona Reid. See online version for full colors.

problems with alignment of clearly nonhomologous elements (e.g., different exons aligned to each other, introns aligned to exons, paralogs aligned to each other, contaminants, editing errors) present and common as in prior studies of mammalian phylogeny that are based on OrthoMaM sequences (Song et al. 2012; Edwards et al. 2016; Edwards 2016), or have such errors been reduced or eliminated? Second, what explains the surprising conflicts among coalescence and concatenation analyses that were observed in this study? Some

phylogenetic relationships contradict extensive published work and garnered strong support in supermatrix trees (Figures 1a and 2a).

Materials and Methods

Alignments and Gene Trees

We visually inspected all sequence alignments in Feijoo and Parada's (2017) Mammalia (95 genes), Arctoidea (29 genes), and Chiroptera (18 genes) data sets as well as all ML gene trees derived from these data. We reconstructed gene trees with the RAxML (Stamatakis 2006) module in Geneious 9.1.7 (Kearse et al. 2012) using the GTR + Γ model and rapid bootstrapping (100 pseudoreplicates) plus a search for the best-scoring ML tree option. Suspected homology problems in the protein-coding sequences were investigated by querying regions of dubious homology against whole genome sequences on NCBI. We used Megablast for queries with highly similar sequences (i.e., same species, genus, or family) and BlastN for queries with more divergent sequences (i.e., different families or orders). When the suspicious sequence region from a particular species mapped to a different position in the reference genome relative to sequences from other species in the alignment, we then searched for the correct (one-to-one orthologous) region in the genome of the species that was misaligned. If the correct segment of sequence was recovered, this evidence was used to conclude that the initial error was due to paralogy, misalignment of introns with exons, or alignment of the wrong exons with each other (i.e., exon 1 from one species aligned with exon 2 from other species, which would require an impossible "jump" in the alignment across intron 1).

Supermatrix Analyses with Maximum Likelihood

We performed partitioned maximum likelihood analyses with RAxML 8.2.10 (Stamatakis 2006) on CIPRES (Miller et al. 2010) and IQ-TREE (Nguyen et al. 2015). RAxML analyses allowed each partition to have its own GTR + Γ model of sequence evolution and employed 500 bootstrap pseudo-replications with a search for the optimal tree in the same run. IQ-TREE analyses were performed with default parameters for partition models (Chernomor et al. 2016), model selection (Kalyaanamoorthy et al. 2017), and the ultrafast approximation for bootstrapping (Minh et al. 2013).

Partitioned Likelihood Support

Partitioned likelihood support scores (Lee and Hugall 2003; Gatesy and Baker 2005; Shen et al. 2017) were determined by optimizing the fit of each gene alignment to the optimal RAxML topology for Feijoo and Parada's (2017) original data (Figure 1a for Arctoidea, Figure 2a Chiroptera) relative to a topology that is identical except for the conflicting (and traditional) clade (Figure 1b for Arctoidea, Figure 2b for Chiroptera). PAUP* 4.0a (Swofford 2002) was used to optimize each data partition onto alternative species trees, and a unique GTR + Γ model was permitted for each gene in the data set.

Results and Discussion

Feijoo and Parada's (2017) Mammalia Data Set

We detected one or more homology errors in ~91% (86/95) of the gene alignments in Feijoo and Parada's (2017) Mammalia data set, and for 35 alignments (37%), these homology errors are associated with topological artifacts, misplacements of the affected taxa, and in some cases, extremely long branches (Supplementary Table S1). For several genes, the homology problems are the result of including entire (*EPYC*) or partial (*ACSL6*, *BRPF1*, *KAT6B*, *OSBPL3*, *SLC4A10*, *TNKS*) sequences from paralogs in the alignments. For the *EPYC* alignment, *Phoca largha* (spotted seal) is represented by the paralog *OGN* instead of *EPYC*. The seal (*P. largha*) is sister to a rodent (*Ictidomys tridecemlineatus*; thirteen-lined ground squirrel) on the corresponding gene tree. Well-established mammalian clades that are disrupted by this specious grouping include Rodentia, Glires, Pinnipedia, Arctoidea, and Carnivora. For *ACSL6*, *P. largha*, *Pteropus alecto* (black flying fox), and *Artibeus jamaicensis* (Jamaican fruit bat) share a segment of the paralog *ACSL1* and form a clade on the associated gene tree. This homology error disrupts support for multiple well-established taxonomic groups (e.g., Pinnipedia, Arctoidea, Carnivora, Chiroptera). Paralogy problems also impact the "*TNKS*" sequence for *P. largha*, which instead is an alternating chimera of eight *TNKS* segments and seven *TNKS2* segments (Supplementary Table S1; Supplementary Figure S1). Further, some of the *TNKS* segments from this "Phocidae" chimera differ from other phocid *TNKS* sequences (*Leptonychotes weddellii*, *Neomonachus schauinslandi*) and are identical to the odobenid (*Odobenus rosmarus*) *TNKS* sequence (Supplementary Figure S1). Even though *P. largha TNKS* is a chimeric sequence with interdigitated contaminated and paralogous segments, the pinnipeds *O. rosmarus* and *P. largha* cluster on the *TNKS* gene tree. In addition to paralogy issues, additional homology problems in the Mammalia data set (wrong exons aligned to each other, introns aligned to exons, contamination of sequences across species) are described in Supplementary Table S1. These errors are commonly associated with gene trees that show conflicts with well-established mammalian clades.

The cross-contamination of *Phoca largha TNKS* with orthologous sequences from *Odobenus rosmarus* is a representative example of a more general contamination problem in the Mammalia dataset where we detected 47 instances of identical sequences that impact three pairs of taxa (Supplementary Table S2). First, the phyllostomid *Artibeus jamaicensis* and the vespertilionid *Myotis lucifugus* (little brown bat) have identical sequences for 11 genes. Second, the pinnipeds *O. rosmarus* and *P. largha* share identical sequences for 14 genes. Finally, two species of *Pteropus* (*P. vampyrus*, *P. alecto*) have identical sequences for 22 genes. All three pairs include at least one species for which Feijoo and Parada (2017) assembled gene sequences from NCBI's Sequence Read Archive (SRA). It is possible that orthologous sequences at some loci are identical for the recently diverged *Pteropus* species, although this is not the case for *ASXL2* that we investigated in detail (Table S2). Identical sequences are increasingly less likely for pinnipeds (*P. largha*, *O. rosmarus*) that diverged ~21 million years ago and for bats (*A. jamaicensis*, *M. lucifugus*) that diverged ~56 million years ago (Meredith et al. 2011). Indeed, these instances of identical sequences appear to have resulted from duplicating *O. rosmarus* sequences for *P. largha*, and *M. lucifugus* sequences for *A. jamaicensis* (Supplementary Table S2).

Overall, homology and contamination errors were so extensive in Feijoo and Parada's (2017) Mammalia data that we made no attempt to correct these various problems. Given the magnitude of errors in this data set (Supplementary Tables S1 and S2), which was constructed to address some of the most challenging parts of the mammalian tree, Feijoo and Parada's (2017) phylogenomic results should be treated with extreme caution.

Feijoo and Parada's (2017) Arctoidea Data Set

Inspection of Feijoo and Parada's (2017) Arctoidea data set also uncovered extensive homology errors. Eleven of 29 gene alignments (~38%) show cross alignment of different exons with each other,

Table 1. Homology and contamination errors in Feijoo and Parada's (2017) Arctoidea data set for 29 genes[a]

Gene	Problematic alignment positions	Description of problem	Phylogenetic artifact on tree	Correction(s) to Feijoo and Parada's (2017) alignment
ACSL6	1–399 (whole segment) for *Phoca*, *Arctocephalus*, and *Ursus*	*Phoca*, *Arctocephalus*, and *Ursus* sequences are from *ACSL1* paralog	*Phoca* + *Arctocephalus* + *Ursus* clade is supported by long internal branch (0.7335 substitutions/site) and is sister to *Ochotona*	Replaced *Ursus maritimus* with sequence for *U. maritimus* from NCBI WGS (AVOR01060252); replaced *Arctocephalus gazella* with assembled sequence from *A. gazella* genome SRA (SRX1338492); replaced Phocidae with sequence from *Leptonychotes weddellii* from NCBI WGS (APMU01150005 [first four exons], APMU01111985 [last exon]).
APPBP2	1–446 (whole segment) for *Dipodomys*	This segment is 100% identical to *Cavia* but is missing from OrthoMaM v8	*Dipodomys* and *Cavia* cluster together and have terminal branch lengths of zero	Replaced *Dipodomys ordii* sequence with correct sequence from NCBI WGS (NW_012267235).
CREBBP	*Neovison* (70–349), *Arctocephalus* (7–401), *Phoca* (864–957), *Callithrix* (1076–1159)	*Neovison* and *Arctocephalus* sequences are from *EP300* paralog; additional homology problems within exon 6 of *Phoca* and *Callithrix*	*Arctocephalus* and *Neovison* are sister taxa; *Phoca* and *Callithrix* have long terminal branches	Replaced exon 6 of *Phoca largha* with exon 6 of another phocid, *Leptonychotes weddellii*, from NCBI WGS (APMU01127148); replaced exon 6 of *Callithrix jacchus* with exon 6 of another platyrrhine (*Saimiri boliviensis*, AGCE01054468); deleted *Neovison vison* and *Arctocephalus gazella* sequences.
DYNC1I2	*Callithrix* (63–106), *Rattus* (227–243), Xenarthra (227–243), *Tursiops* (227–243)	Shared nonhomologous segment in *Rattus*, Xenarthra, and *Tursiops*; additional homology problem in *Callithrix*	*Callithrix* and *Rattus* have long terminal branches	Replaced *Rattus norvegicus* with homologous sequence (AAHX01022593) from *R. norvegicus*; replaced *Tursiops truncatus* with homologous sequence (MRVK01000923) from *T. truncatus*; deleted nonhomologous segment for Xenarthra.
GUCY1B3	*Odobenus* (3–11), *Otolemur* (235–427), *Equus* (432–468)	Various homology problems (exon 1 in *Odobenus*; exon 3 in *Otolemur*; exon 4 in *Equus*)	*Otolemur* sister to Paenungulata	Replaced *Odobenus rosmarus* (3–11) with homologous sequence from *O. rosmarus* (ANOP01005650); replaced *Phoca largha* exon 1 (mostly missing) with sequence from another phocid, *Leptonychotes weddellii* (XM_006727271); replaced *Otolemur garnettii* exon 3 with correct ortholog (AAQR03198278); replaced *Equus caballus* exon 4 with correct homolog (ATDM01050404).
HSPA13	*Loxodonta* (6–24), *Pongo* (985–1415)	Various homology problems (exon 1 in *Loxodonta*; exon 5 in *Pongo*)	*Pongo* sister to other Anthropoidea	Replaced *Loxodonta africana* exon 1 with correct sequence from *L. africana* (AAGU03032957); replaced *Pongo abelii* exon 5 (partial) with assembled sequence from NCBI SRA for *P. abelii* (ERR1407287).
KIAA1279	*Arctocephalus* (348–544)	Homology problems in exons 4 and 5 of *Arctocephalus*	*Arctocephalus* has long terminal branch	Replaced *Arctocephalus gazella* exons 4 and 5 with correct exon sequences that were assembled from *A. gazella* transcriptome SRA (ERX009916).
MXD1	*Phoca* (1–294, whole segment), *Arctocephalus* (11–277), *Loxodonta* (171–282)	*Phoca* and *Arctocephalus* sequences are from *MXI1* paralog; additional homology problem in exon 4 of *Loxodonta*	*Phoca* + *Arctocephalus* clade with long stem branch (0.9759 substitutions/site). *Ornithorhynchus* roots tree between *MXD1* (most taxa) and *MXI1* (*Phoca*, *Arctocephalus*) sequences	Replaced *MXI1* sequence of *Phoca largha* with *MXD1* sequence for *Leptonychotes weddellii* from NCBI WGS (APMU01018803 [exons 1–3], APMU01133638 [exons 4–5]); replaced exon 4 of *Loxodonta africana* with assembled sequence from *L. africana* SRA (SRX339470); deleted *MXI1* sequence of *Arctocephalus gazella*.
OSBPL3	*Phoca* (3–221), *Macaca* (86–164), *Sus* (626–704), *Bos* (705–800)	*Phoca* is *OSBPL7* paralog; additional homology problems in *Macaca* (exon 3), *Sus* (exon 7), and *Bos* (exon 8)	*Phoca* is nested in Marsupialia as sister to *Sarcophilus*; long terminal branches for *Macaca, Sus*, and *Bos*	Replaced *Phoca largha* with sequence for *Leptonychotes weddellii* from NCBI WGS (NW_006386799); replaced exon 3 of *Macaca mulatta* with correct exon 3 sequence from NCBI WGS (MRVA01067729); replaced exon 7 of *Sus scrofa* with correct exon 7 sequence from NCBI WGS (LUXY01028385); replaced exon 8 of *Bos taurus* with correct exon 8 sequence from NCBI WGS (DAAA02010993).

Table 1. *Continued*

Gene	Problematic alignment positions	Description of problem	Phylogenetic artifact on tree	Correction(s) to Feijoo and Parada's (2017) alignment
PSL3	*Phoca* (1–441, whole segment)	*Phoca* sequence is *PSL2* paralog	*Phoca* sister to Marsupialia	Replaced *Phoca largha* sequence with ortholog that was assembled from *P. largha* transcriptome SRA (SRX120902).
TTC39B	*Equus* (1–137), *Macaca* (143–233), *Bos* (471–525)	Various homology problems (exon 1 in *Equus*, exon 2 in *Macaca*, exon 5 in *Bos*)	*Bos* and *Macaca* have long terminal branches	Replaced *Equus caballus* exon 1 sequence with correct *E. caballus* sequence (ATDM01013433); replaced *Macaca mulatta* exon 2 sequence with correct *M. mulatta* sequence (MRVA01172571); replaced *Bos taurus* exon 5, which includes three discontiguous segments from intron 4 and the 3' end of exon 5, with the correct *B. taurus* sequence for exon 5 (AC_000165).

[a]For exon numbering, we assigned exon 1 to the most 5' exon in each of Feijoo and Parada's (2017) alignments and numbered the remaining 3' exons in consecutive order following exon 1. Exon boundaries in Feijoo and Parada's (2017) alignments were determined by BLAST searches with their *Homo sapiens* protein-coding sequences against the *H. sapiens* reference genome. Candidate homology errors in Feijoo and Parada's (2017) Arctoidea data set were initially identified based on visual inspection of alignments in Geneious. When putative homology problems were identified, we then performed BLAST searches to determine if one-to-one orthologous segments are present in available genomic resources for the affected taxon but were missed because of annotation errors/assembly problems. For BLAST searches, we used query sequences from closely related taxa without the homology problem and BLASTed the genome(s) of target taxa with the homology problem. Query sequences from taxa without the homology problem included multiple contiguous exons, including the correct sequence for the exon(s) that was putatively misidentified in the target taxon plus adjacent exons in the same gene. Application of this protocol allowed us to determine if one-to-one orthologs (entire genes or missing exons thereof) are present in the correct genomic position in taxa with candidate homology problems. When candidate one-to-one orthologs were recovered in target taxa with homology problems, these candidate orthologs were then used as BLAST query sequences to determine if the highest BLAST hits were from the expected ortholog. We also used BLAST searches to confirm the identities of all paralogous genes and segments thereof. For example, Feijoo and Parada's (2017) putative sequence for *Phoca largha ACSL6* is 99% similar with 100% query coverage to *ACSL1* in two annotated phocids with genome sequences (*Leptonychotes weddellii, Neomonachus schauinslandi*), but only 72% similar with 92% query coverage to *ACSL6* in these same two phocids. Based on the results of following our protocols for confirming homology problems and finding one-to-one orthologs, we replaced nonhomologous sequences with one-to-one orthologs prior to concatenating these alignments into a supermatrix for partitioned RAxML concatenation analysis and re-estimation of ML gene trees. A detailed summary of our corrections to Feijoo and Parada's (2017) Arctoidea matrix, including accession numbers of sequences that were employed to correct homology errors, is provided above.

alignment of exons to introns, and/or paralogy problems (Table 1). Among these are six genes (*ACSL6, CREBBP, GUCY1B3, MXD1, OSBPL3, PSL3*) with errors that impact one or more pinnipeds. Most notable among these problems are paralogous sequences in the *ACSL6* and *MXD1* alignments (Figure 1c and d). For *ACSL6*, sequences for *Phoca largha* (Phocidae), *Arctocephalus gazella* (Antarctic fur seal, Otariidae), and *Ursus maritimus* (polar bear) are instead from the paralog *ACSL1*. These three taxa cluster together on the "*ACSL6*" gene tree; the paralogy error yields a large set of "pseudo-synapomorphies" for this unconventional clade (Figure 1c). In the *MXD1* alignment, the pinnipeds *P. largha* and *A. gazella* are represented by sequences from the paralog *MXI1*. Not surprisingly, the phocid *P. largha* and the otariid *A. gazella* cluster together on the "*MXD1*" gene tree at the end of another extremely long stem branch (0.9759 substitutions/site) (Figure 1d). These paralogy problems in *ACSL6* and *MXD1*, along with additional homology errors that are detailed in Table 1, provide a possible explanation for the unexpected sister relationship between Phocidae and Otariidae to the exclusion of Odobenidae based on both ML and Bayesian concatenation methods (Figure 1a).

To test this hypothesis, we first calculated partitioned likelihood support (PLS) scores at the controversial Phocidae + Otariidae node (Lee and Hugall 2004; Gatesy and Baker 2005; Shen et al. 2017). The PLS analysis revealed extreme outlier values for the two genes with paralogy problems within Pinnipedia. Support for the unconventional clade is almost wholly concentrated in *ACSL6* and *MXD1* (Figure 2e). We then corrected all of the homology problems that we detected in the Arctoidea data set (Table 1) by either 1) substituting sequences to correct homology errors or 2) deleting nonhomologous sequences when appropriate orthologous regions were not available at online databases. All of our modifications to Feijoo and Parada's (2017) original Arctoidea alignments are listed in Table 1.

Figure 1b shows the RAxML tree that resulted from analyzing the corrected supermatrix for Arctoidea. By contrast with a RAxML analysis of the uncorrected supermatrix (Figure 1A) and Feijoo and Parada's (2017) supermatrix analyses that robustly support Otariidae + Phocidae (100% bootstrap for IQ-TREE and RAxML; 1.0 posterior probability for BEAST2), RAxML analysis of the corrected Arctoidea concatenation provides 100% bootstrap support for a sister-group relationship between Otariidae and Odobenidae (Otaroidea). This result agrees with numerous studies that are based on multigene data sets (Flynn et al. 2005; Arnason et al. 2006; Fulton and Strobeck 2006; Meredith et al. 2011; Emerling et al. 2015; Foley et al. 2016), indels (Luan et al. 2013), and retroposon insertions (Doronina et al. 2015). Feijoo and Parada (2017) attributed the Otariidae + Phocidae result to defects of concatenation, but this misleading result is not the fault of concatenation and instead is a direct consequence of two paralogy errors in the Arctoidea matrix (Figures 1c–d). Indeed, by fixing just two alignments (*ACSL6* and *MXD1*) from Feijoo and Parada's (2017) original data set and rerunning ML concatenated searches, Otaroidea is robustly supported (100% bootstrap). In addition to differences in topology and support, note that some branch lengths are doubled (Otariidae is 2.0× longer) or almost tripled (Phocidae is 2.8× longer) for the ML species tree based on alignments with homology errors (Figure 1a) relative to the ML tree based on the corrected data set (Figure 1b). Branch length

Table 2. Homology, contamination, and missing data problems in Feijoo and Parada's (2017) Chiroptera data set for 18 genes[a]

Gene	Problematic alignment positions	Description of problem	Phylogenetic artifact on tree	Correction(s) to Feijoo and Parada's (2017) alignment
ATF7IP	3393–3809 for *Ornithorhynchus* and *Pongo*	*Ornithorhynchus* and *Pongo* are mostly missing with short regions of nonhomologous sequence relative to other taxa	*Ornithorhynchus* roots Theria (marsupials + placentals) on *Pongo*	Replaced *Ornithorhynchus anatinus* with orthologous exon 14 sequence from *O. anatinus* NCBI RefSeq (NW_001699712); replaced *Pongo abelii* with orthologous exon 14 sequence from *P. abelii* NCBI RefSeq (NC_012603).
BRF1	2097–2287 for *Echinops*	Various homology issues in *Echinops*	*Echinops* is sister to Paenungulata + Rodentia + Primates	Replaced *Echinops telfairi* sequence with orthologous sequence for exon 6 from NCBI RefSeq (NW_004558712).
CREM	889–1035 for all taxa	Exon 8 in 18 taxa (e.g., *Gorilla*) is aligned with exon 9 in 20 other taxa (e.g., *Homo*); two taxa (both *Uroderma bilobatum* subspecies) have paralogous exon from *CREB1*	Tree has three sectors corresponding to three different exons that are aligned with each other; bats are polyphyletic	Expanded *CREM* alignment to include exons 8 and 9. *Uroderma bilobatum CREB1* paralog sequences were deleted from the alignment but not replaced with *CREM* sequences because SRA sequences for *Uroderma* were not accessible as of 14 July 2017. Accession numbers for exons 8 and 9 in different taxa are as follows: *Artibeus jamaicensis* (SRX176203); *Aselliscus stoliczkanus* (SRR2153215); *Bos taurus* (AAFC05001556); *Callithrix jacchus* (BBXK01086289); *Canis lupus* (AOCS01172096); *Cavia porcellus* (AAKN02037523); *Desmodus rotundus* (SRX201174/231/228/167); *Echinops telfairi* (SRR107639); *Equus caballus* (ATDM01064002); *Erinaceus europaeus* (AMDU01096190); *Felis catus* (ACBE01224633); *Gorilla gorilla* (CABD030074496); *Hipposideros pratti* (SRR2153216); *Homo sapiens* (NC_000010); *Ictidomys tridecemlineatus* (AGTP01064214); *Loxodonta africana* (AAGU03083382); *Macaca mulatta* (MRVA01030921); *Megaderma lyra* (AWHB01419657); *Monodelphis domestica* (AAFR03014385); *Mus musculus* (LVXW01058264); *Mustela putorius* (AGTQ01043870); *Myotis lucifugus* (AAPE02013287); *Nomascus leucogenys* (ADFV01091295); *Ornithorhynchus anatinus* (AAPN01000359); *Otolemur garnettii* (AAQR03043815); *Pan troglodytes* (AADA01270532); *Pongo abelii* (ABGA01375434); *Procavia capensis* (ABRQ02138152); *Pteronotus parnellii* (AWGZ01424591); *Pteropus alecto* (ALWS01040371); *Pteropus vampyrus* (ABRP02135855); *Rattus norvegicus* (AAHX01091771); *Rhinolophus ferrumequinum* (AWHA01055806); *Sarcophilus harrisii* (AFEY01455779); *Sus scrofa* (LUXY01079840); *Tadarida teniotis* (SRX1140293); *Taphozous melanopogon* (SRX1140292).
GAREM	*Ornithorhynchus* (273–456)	3rd exon of *Homo sapiens* (264–506) BLASTs to more than six discontiguous regions in *Ornithorhynchus* that have dubious homology	*Ornithorhynchus* roots Theria on *Erinaceus* branch	Deleted last exon of *Ornithorhynchus anatinus* (orthologous replacement not found; closest BLAST results for the deleted exon had query coverage of only 20%).

Table 2. *Continued*

Gene	Problematic alignment positions	Description of problem	Phylogenetic artifact on tree	Correction(s) to Feijoo and Parada's (2017) alignment
KAT6B	*Nomascus* (5–543), *Ornithorhynchus* (3144–3204), *Loxodonta* (4530–4881)	Unrelated homology problems (exon 1 in *Nomascus*; exon 14 in *Ornithorhynchus*; exon 16 in *Loxodonta*)	*Nomascus* nested inside of Hominidae	Replaced *Nomascus leucogenys* (exon 1) with orthologous sequence from *N. leucogenys* that was assembled from NCBI SRA (SRR408501, SRR408502); replaced dubious region in *Ornithorhynchus* exon 14 with Ns; replaced *Loxodonta africana* exon 16 (4530–4881) with orthologous sequence from *L. africana* that was assembled from NCBI SRA (SRR408501, SRR408502).
KIAA0100	*Loxodonta* (2218–2947), *Tursiops* (6427–6528)	Unrelated homology problems (exon 16 in *Loxodonta*; exon 36 in *Tursiops*)	*Loxodonta* is sister to *Megaderma*; *Tursiops* is sister to *Erinaceus*	Replaced *Loxodonta africana* exon 16 with orthologous sequence from *L. africana* (AAGU03032957) that was assembled from NCBI SRA (SRR958467); replaced *Tursiops truncatus* exon 36 with orthologous sequence from *T. truncatus* (MRVK01001488).
MYB	300–526 (*Aselliscus* and *Hipposideros*)	*Aselliscus* and *Hipposideros* sequences are from *MYBL1* paralog	*Aselliscus* + *Hipposideros* clade is sister to *Cavia*, which breaks up monophyly of bats	Deleted *Aselliscus stoliczkanus* and *Hipposideeros pratti*. Unable to assemble complete *MYB* segments for *A. stoliczkanus* and *H. pratti* based on NCBI SRA brain transcriptomes for these species (*A. stoliczkanus* [SRX1140287], *H. pratti* [SRX1140288]).
NRP1	1–201 (*Sarcophilus*)	*NRP1* alignment is 203 bp in length and *Sarcophilus* is coded as Ns for 201 of the 203 bp	*Sarcophilus* and *Monodelphis* (Marsupialia) do not cluster together and are nested inside of Placentalia	Filled in extensive missing data in *Sarcophilus harrisii* with sequence from NCBI RefSeq (NW_003843688).
PRMT10	1328–1676 (*Sarcophilus*)	*Sarcophilus* sequence is not homologous with other sequences	*Sarcophilus* + *Echinops* clade	Replaced *Sarcophilus harrisii* sequence with orthologous sequence from *S. harrisii* NCBI RefSeq (NW_003846797).
TTC14	1-2302 (*Hipposideros*)	*Hipposideros* sequence is contaminated with *Homo* sequence, including the only region of alignment (1709–2302) with more than five taxa	*Hipposideros* is sister to one of the two *Homo* sequences, which makes bats polyphyletic	Deleted *Hipposideros pratti* (*Homo sapiens* contaminant). Unable to assemble complete *TTC14* segment based on NCBI SRA brain transcriptome for *H. pratti* (SRX1140288).

[a]For exon numbering, we assigned exon 1 to the 5′ exon in each of Feijoo and Parada's (2017) alignments and numbered the remaining 3' exons in consecutive order following exon 1. Exon boundaries in Feijoo and Parada's (2017) alignments were determined by BLAST searches with their *Homo sapiens* protein-coding sequences against the *H. sapiens* reference genome. See Table 1 for procedures that were used to detect and validate homology errors and to find one-to-one orthologs. Based on the results of following this protocol, we replaced nonorthologous sequences with one-to-one orthologs prior to concatenating these alignments into a supermatrix for partitioned RAxML concatenation analysis and re-estimation of ML gene trees. A detailed summary of our corrections to Feijoo and Parada's (2017) Chiroptera matrix, including accession numbers of sequences that were employed to correct homology errors, is provided above.

distortions due to homology errors impact both concatenation and parametric coalescence methods, such as *BEAST, that base divergence times between species on the amount of divergence between assumed orthologous sequences (Heled and Drummond 2010).

Feijoo and Parada's (2017) Chiroptera Data Set

Like the Mammalia and Arctoidea data sets, Feijoo and Parada's (2017) Chiroptera data set for 42 taxa is replete with homology errors (Table 2). Ten of the 18 genes (56%) show homology problems including paralogous sequences, alignments of different exons to each other, cross-contamination of sequences between species, and extensive missing data (~74.5% of the Chiroptera data set is coded as missing for at least 37 taxa). For each of the ten problematic alignments, taxa impacted by homology errors are misplaced in associated gene trees, including multiple cases where the monophyly of Chiroptera is disrupted (Table 2; Figure 2c). The alignment for *CREM* is especially problematic. The only region with more than four taxa (positions 889–1035) aligns exon 8 of *CREM* (three bat species), exon 9 of *CREM* (nine bat species), and paralogous sequences from *CREB1* (two bats) against each other. Likelihood analysis predictably yields a gene tree with polyphyly of Chiroptera and several extremely long internal branches that correspond with the homology errors (Figure 2c). Partitioned likelihood support scores record uneven support among the 18 genes in the Chiroptera data set for the controversial *Taphozous* + *Tadarida* clade, with an extreme outlier PLS score from the tainted *CREM* alignment (Figure 2d). By fixing the *CREM* alignment in Feijoo and Parada's (2017) Chiroptera data set and rerunning ML concatenated searches, the traditional association of *Tadarida* + *Myotis* is supported.

Following correction of all of the homology errors that we detected in the Chiroptera data set (see footnote of Table 2), we executed concatenated ML analyses (IQ-TREE and RAxML) and recovered the *Tadarida* + *Myotis* clade with 85–91% bootstrap support (Figure 2b). Although Feijoo and Parada (2017) reported

that IQ-TREE analysis of their original bat data set supported the *Tadarida* + *Myotis* clade (89% bootstrap), we could not replicate this result. Instead, our IQ-TREE reanalysis of Feijoo and Parada's (2017) original bat data set yielded 91% bootstrap support for *Tadarida* + *Taphozous* (Figure 2a), which closely matches the results of our RAxML analysis for this data set. We therefore reinterpret the conflicts that Feijoo and Parada (2017) noted among different concatenation methods at this node as due to investigator error, not defects of concatenation analysis.

Conclusions

In summary, we have documented extensive homology errors, cross-contamination problems, and irreproducible analytical results in Feijoo and Parada's (2017) phylogenomic study of deep mammalian phylogeny. These errors are reminiscent of the many problems that Springer and Gatesy (2016) detected in an earlier phylogenomic study of mammals that also incorporated protein-coding sequences extracted from an earlier version of OrthoMaM (Song et al. 2012). OrthoMaM is a valuable resource for compiling individual exons and complete protein-coding sequences from various mammalian species, but systematists should employ strict filtering steps to minimize homology problems in complete protein-coding sequences that are stored in the OrthoMaM database (e.g., see Mason et al. 2016).

Pipeline scripts for extracting sequences from genomic and transcriptomic databases can be useful for assembling very large phylogenomic data sets, but the shear magnitude of data does not excuse systematists from quality control and data filtering. To ensure that homology errors are reduced prior to downstream phylogenetic/evolutionary analyses, visual checking of alignments and gene trees (or at least spot-checking of a sample of alignments and gene trees) is critical to detect glitches in pipeline procedures. Programs such as RF Distances Filter (Simmons et al., 2016) are especially useful for identifying problematic alignments and gene trees that may plague a phylogenomic data set. Prior work has demonstrated convincingly that solutions to difficult phylogenetic problems can be influenced profoundly by the inclusion or exclusion of a single gene in a large data set (Gatesy et al. 1999, 2017; Brown and Thomson 2017; Shen et al. 2017), and that in some cases, "outlier" genes that exert a large influence in phylogenomic analyses are characterized by homology errors (e.g., Brown and Thomson 2017; Shen et al. 2017).

The artifactual Otariidae + Phocidae clade (Figure 1a) demonstrates that even "easy to resolve" clades such as Otaroidea can be overturned by just a few homology problems. After correcting 11 of the 29 alignments in the Arctoidea data set (Table 1), 100% bootstrap support for Otariidae + Phocidae (Figure 1a) was converted to 100% bootstrap support for Otariidae + Odobenidae (Figure 1b). Correction of paralogy errors in just two genes, *ACSL6* and *MXD1* (Figure 1c–d), likewise yielded traditional relationships among pinniped families. A similar pattern was observed for the Chiroptera data set, where a very high PLS score identified an outlier gene (*CREM*) at the *Tadarida* + *Taphozous* node (Figure 2d). Correction of homology problems for this one gene yielded a more conventional phylogenetic result, as did more thorough corrections of the homology errors in the Chiroptera data set (Figure 2b). Big data are not immune to homology problems that can distort analyses of challenging phylogenetic problems. Systematic biologists should strive for both accountability and responsibility by scrutinizing sequence alignments, the primary hypotheses of homology that underlie all phylogenetic/evolutionary inferences in the field of molecular systematics.

Supplementary Material

Supplementary material can be found at https://academic.oup.com/jhered/.

Funding

This work was supported by National Science Foundation grant DEB-1457735 to J.G. and M.S.S.

Acknowledgments

We thank M. Feijoo for providing alignments. C. Buell and F. Reid provided images of mammals. We thank two reviewers and A. Roca for helpful comments on an earlier version of this manuscript.

Data Availability

Corrected versions of Feijoo and Parada's (2017) Arctoidea and Chiroptera data sets are provided in Supplementary Material.

References

Amador LI, Arévalo RLM, Almeida FC, Catalano SA, Giannini NP. 2016. Bat systematics in the light of unconstrained analyses of a comprehensive molecular supermatrix. *J Mammal Evol.* DOI:10.1007/s10914-016-9363-8.

Arnason U, Gullberg A, Janke A, Kullberg M, Lehman N, Petrov EA, Väinölä R. 2006. Pinniped phylogeny and a new hypothesis for their origin and dispersal. *Mol Phylogenet Evol.* 41:345–354.

Bergsten J. 2005. A review of long-branch attraction. *Cladistics.* 21:163–193.

Brown JM, Thomson RC. 2017. Bayes factors unmask highly variable information content, bias, and extreme influence in phylogenomic analyses. *Syst Biol.* 66:517–530.

Chiari Y, Cahais V, Galtier N, Delsuc F. 2012. Phylogenomic analyses support the position of turtles as the sister group of birds and crocodiles (Archosauria). *BMC Biol.* 10:65.

Chernomor O, von Haeseler A, Minh BQ. 2016. Terrace aware data structure for phylogenomic inference from supermatrices. *Syst Biol.* 65:997–1008.

Degnan JH, Rosenberg NA. 2006. Discordance of species trees with their most likely gene trees. *PLoS Genet.* 2:e68.

Delsuc F, Brinkmann H, Chourrout D, Philippe H. 2006. Tunicates and not cephalochordates are the closest living relatives of vertebrates. *Nature.* 439:965–968.

de Queiroz A, Gatesy J. 2007. The supermatrix approach to systematics. *Trends Ecol Evol.* 22:34–41.

Doronina L, Churakov G, Shi J, Brosius J, Baertsch R, Clawson H, Schmitz J. 2015. Exploring massive incomplete lineage sorting in arctoids (Laurasiatheria, Carnivora). *Mol Biol Evol.* 32:3194–3204.

dos Reis M, Inoue J, Hasegawa M, Asher RJ, Donoghue PC, Yang Z. 2012. Phylogenomic datasets provide both precision and accuracy in estimating the timescale of placental mammal phylogeny. *Proc Biol Sci.* 279:3491–3500.

Douzery EJ, Scornavacca C, Romiguier J, Belkhir K, Galtier N, Delsuc F, Ranwez V. 2014. OrthoMaM v8: a database of orthologous exons and coding sequences for comparative genomics in mammals. *Mol Biol Evol.* 31:1923–1928.

Edwards SV. 2009. Is a new and general theory of molecular systematics emerging? *Evolution.* 63:1–19.

Edwards SV. 2016. Phylogenomic subsampling: a brief review. *Zool Scr.* 45(suppl 1):63–74.

Edwards SV, Xi Z, Janke A, Faircloth BC, McCormack JE, Glenn TC, Zhong B, Wu S, Lemmon EM, Lemmon AR, *et al.* 2016. Implementing and testing the multispecies coalescent model: A valuable paradigm for phylogenomics. *Mol Phylogenet Evol*. 94:447–462.

Eick GN, Jacobs DS, Matthee CA. 2005. A nuclear DNA phylogenetic perspective on the evolution of echolocation and historical biogeography of extant bats (Chiroptera). *Mol Biol Evol*. 22:1869–1886.

Emerling CA, Huynh HT, Nguyen MA, Meredith RW, Springer MS. 2015. Spectral shifts of mammalian ultraviolet-sensitive pigments (short wavelength-sensitive opsin 1) are associated with eye length and photic niche evolution. *Proc R Soc B*. 282:20151817.

Feijoo M, Parada A. 2017. Macrosystematics of eutherian mammals combining HTS data to expand taxon coverage. *Mol Phylogenet Evol*. 113:76–83.

Flynn JJ, Finarelli JA, Zehr S, Hsu J, Nedbal MA. 2005. Molecular phylogeny of the Carnivora (Mammalia): assessing the impact of increased sampling on resolving enigmatic relationships. *Syst Biol*. 54:317–337.

Foley NM, Springer MS, Teeling EC. 2016. Mammal madness: Is the mammal tree of life not yet resolved. *Phil Trans R Soc B*. 371:20150140.

Fulton TL, Strobeck C. 2006. Molecular phylogeny of the Arctoidea (Carnivora): effect of missing data on supertree and supermatrix analyses of multiple gene data sets. *Mol Phylogenet Evol*. 41:165–181.

Gatesy J, Baker RH. 2005. Hidden likelihood support in genomic data: can forty-five wrongs make a right? *Syst Biol*. 54:483–492.

Gatesy J, Springer MS. 2013. Concatenation versus coalescence versus "concatalescence". *Proc Natl Acad Sci USA*. 110:E1179.

Gatesy J, Springer MS. 2014. Phylogenetic analysis at deep timescales: unreliable gene trees, bypassed hidden support, and the coalescence/concatalescence conundrum. *Mol Phylogenet Evol*. 80:231–266.

Gatesy J, O'Grady P, Baker R. 1999. Corroboration among data sets in simultaneous analysis: hidden support for phylogenetic relationships among higher level artiodactyl taxa. *Cladistics*. 15:271–313.

Gatesy J, Meredith RW, Janecka JE, Simmons MP, Murphy WJ, Springer MS. 2017. Resolution of a concatenation/coalescence kerfuffle: partitioned coalescence support and a robust family-level tree for Mammalia. *Cladistics*. 33:295–332.

Heled J, Drummond AJ. 2010. Bayesian inference of species trees from multilocus data. *Mol Biol Evol*. 27:570–580.

Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods*. 14:587–589.

Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C, *et al.* 2012. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*. 28:1647–1649.

Kolaczkowski B, Thornton JW. 2004. Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. *Nature*. 431:980–984.

Kumar V, Hallström BM, Janke A. 2013. Coalescent-based genome analyses resolve the early branches of the Euarchontoglires. *PLoS One*. 8:e60019.

Lack JB, Roehrs ZP, Stanley CE Jr, Ruedi M, Van Den Bussche RA. 2010. Molecular phylogenetics of *Myotis* indicate familial-level divergence for the genus *Cistugo* (Chiroptera). *J Mammal*. 91:976–992.

Lee MS, Hugall AF. 2003. Partitioned likelihood support and the evaluation of data set conflict. *Syst Biol*. 52:15–22.

Luan PT, Ryder OA, Davis H, Zhang YP, Yu L. 2013. Incorporating indels as phylogenetic characters: impact for interfamilial relationships within Arctoidea (Mammalia: Carnivora). *Mol Phylogenet Evol*. 66:748–756.

Mason VC, Li G, Minx P, Schmitz J, Churakov G, Doronina L, Melin AD, Dominy NJ, Lim NT, Springer MS, *et al.* 2016. Genomic analysis reveals hidden biodiversity within colugos, the sister group to primates. *Sci Adv*. 2:e1600633.

McCormack JE, Faircloth BC, Crawford NG, Gowaty PA, Brumfield RT, Glenn TC. 2012. Ultraconserved elements are novel phylogenomic markers that resolve placental mammal phylogeny when combined with species-tree analysis. *Genome Res*. 22:746–754.

Meredith RW, Janečka JE, Gatesy J, Ryder OA, Fisher CA, Teeling EC, Goodbla A, Eizirik E, Simão TL, Stadler T, *et al.* 2011. Impacts of the Cretaceous Terrestrial Revolution and KPg extinction on mammal diversification. *Science*. 334:521–524.

Miller MA, Pfeiffer W, Schwartz T. 2010. Creating the CIPRES Science Gateway for inference of large phylogenetic trees. In Proceedings of the Gateway Computing Environments Workshop (GCE), 14 Nov. 2010, New Orleans, LA. p. 1–8.

Minh BQ, Nguyen MA, von Haeseler A. 2013. Ultrafast approximation for phylogenetic bootstrap. *Mol Biol Evol*. 30:1188–1195.

Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol*. 32:268–274.

Patel S, Kimball RT, Braun EL. 2013. Error in phylogenetic estimation for bushes in the tree of life. *Phylogenet Evol Biol*. 1:110.

Philippe H, Zhou Y, Brinkmann H, Rodrigue N, Delsuc F. 2005. Heterotachy and long-branch attraction in phylogenetics. *BMC Evol Biol*. 5:50.

Philippe H, de Vienne DM, Ranwez V, Roure B, Baurain D, Delsuc F. 2017. Pitfalls in supermatrix phylogenomics. *Eur J Taxon*. 283:1–25.

Ranwez V, Delsuc F, Ranwez S, Belkhir K, Tilak MK, Douzery EJ. 2007. OrthoMaM: a database of orthologous genomic markers for placental mammal phylogenetics. *BMC Evol Biol*. 7:241.

Rokas A, Williams BL, King N, Carroll SB. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature*. 425:798–804.

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

Rosenberg NA, Tao R. 2008. Discordance of species trees with their most likely gene trees: the case of five taxa. *Syst Biol*. 57:131–140.

Scornavacca C, Galtier N. 2017. Incomplete lineage sorting in mammalian phylogenomics. *Syst Biol*. 66:112–120.

Shaw TI, Srivastava A, Chou WC, Liu L, Hawkinson A, Glenn TC, Adams R, Schountz T. 2012. Transcriptome sequencing and annotation for the Jamaican fruit bat (*Artibeus jamaicensis*). *PLoS One*. 7:e48472.

Shen XX, Hittinger CT, Rokas A. 2017. Contentious relationships in phylogenomic studies can be driven by a handful of genes. *Nat Ecol Evol*. 1:126.

Simmons MP, Gatesy J. 2015. Coalescence vs. concatenation: Sophisticated analyses vs. first principles applied to rooting the angiosperms. *Mol Phylogenet Evol*. 91:98–122.

Simmons MP, Sloan DB, Gatesy J. 2016. The effects of subsampling gene trees on coalescent methods applied to ancient divergences. *Mol Phylogenet Evol*. 97:76–89.

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 USA*. 109:14942–14947.

Springer MS, Gatesy J. 2014. Land plant origins and coalescence confusion. *Trends Plant Sci*. 19:267–269.

Springer MS, Gatesy J. 2016. The gene tree delusion. *Mol Phylogenet Evol*. 94:1–33.

Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*. 22:2688–2690.

Sullivan J, Swofford DL. 1997. Are guinea pigs rodents? The importance of adequate models in molecular phylogenetics. *J Mammal Evol*. 4: 77–86.

Swofford DL. 2002. *PAUP*. Phylogenetic analysis using parsimony (* and other methods)*. Sunderland (MA): Sinauer Associates.

Tarver JE, Dos Reis M, Mirarab S, Moran RJ, Parker S, O'Reilly JE, King BL, O'Connell MJ, Asher RJ, Warnow T, *et al.* 2016. The interrelationships of placental mammals and the limits of phylogenetic inference. *Genome Biol Evol*. 8:330–344.

Teeling EC, Springer MS, Madsen O, Bates P, O'Brien SJ, Murphy WJ. 2005. A molecular phylogeny for bats illuminates biogeography and the fossil record. *Science*. 307:580–584.

Teeling EC, Dool S, Springer MS. 2012. Phylogenies, fossils and functional genes: The evolution of echolocation in bats. In: Gunnell GF, Simmons NB, editors. *Evolutionary history of bats: fossils, molecules and morphology.* New York: Cambridge University Press. p. 1–22.