Syst. Biol. 65(6):1057–1075, 2016

DOI:10.1093/sysbio/syw057
Advance Access publication July 17, 2016

The Identification of the Closest Living Relative(s) of Tetrapods: Phylogenomic Lessons for Resolving Short Ancient Internodes

IKER IRISARRI AND AXEL MEYER*

Laboratory for Zoology and Evolutionary Biology, Department of Biology, University of Konstanz, 78464 Konstanz, Germany
**Correspondence to be sent to: Laboratory for Zoology and Evolutionary Biology, Department of Biology, University of Konstanz, 78464 Konstanz, Germany; E-mail: axel.meyer@uni-konstanz.de.*

Received 23 November 2015; reviews returned 7 June 2016; accepted 8 June 2016
Associate Editor: Thomas Near

Abstract.—Identifying the closest living relative(s) of tetrapods is an important, yet still contested question in vertebrate phylogenetics. Three hypotheses are possible and ruling out alternatives has proven difficult even with large molecular data sets due to weak phylogenetic signal coupled nonphylogenetic noise resulting from relatively rapid speciation events that occurred a long time ago (>400 Ma). Here, we revisit the identity of the closest living relative of land vertebrates from a phylogenomic perspective and include new genomic data for all extant lungfish genera. RNA-seq proves to be a great alternative to genomic sequencing, which currently is technically not feasible in lungfishes due to their huge (50–130 Gb) and repetitive genomes. We examined the most important sources of systematic error, namely long-branch attraction (LBA), compositional heterogeneity and distribution of missing data and applied different correction techniques. A multispecies coalescent approach is used to account for deep coalescence that might come from the short and deep internodes separating early sarcopterygian splits. Concatenation methods favored lungfishes as the closest living relatives of tetrapods with strong statistical support. Amino acid profile mixture models can unambiguously resolve this difficult internode thanks to their ability to avoid systematic error. We assessed the performance of different site-heterogeneous models and data partitioning and compared the ability of different strategies designed to overcome LBA, including taxon manipulation, reduction of among-lineage rate heterogeneity and removal of fast-evolving or compositionally heterogeneous positions. The identification of lungfish as sister group of tetrapods is robust regarding the effects of nonstationary composition and distribution of missing data. The multispecies coalescent method reconstructed strongly supported topologies that were congruent with concatenation, despite pervasive gene tree heterogeneity. We reject alternative topologies for early sarcopterygian relationships by increasing the signal-to-noise ratio in our alignments. The analytical pipeline outlined here combines probabilistic phylogenomic inference with methods for evaluating data quality, model adequacy, and assessing systematic error, and thus is likely to help resolve similarly difficult internodes in the tree of life. [Coalescence; coelacanth; compositional heterogeneity; gene tree; long-branch attraction; lungfish; missing data; model misspecification; phylogenomic; species tree; systematic error.]

The conquest of land was a major event in vertebrate evolution (Carroll 1988; Dial et al. 2015) and it has been extensively studied from paleontological, morphological, physiological, behavioral, and molecular perspectives. The terrestrial lifestyle brought new selective forces and required many key innovations and adaptations, such as limbs with digits, modified musculoskeletal and nervous systems, improved hearing and smell, as well as changes in physiology and behavior (Clack 2002). Tetrapods originated from lobe-finned fishes in the Devonian, as supported by the strong paleontological record that associates them with fossil tetrapodomorphs such as *Panderichthys*, *Tiktaalik*, and *Elpisostege* (Ahlberg and Johanson 1998; Clack 2002; Daeschler et al. 2006). Out of water, tetrapods diversified into amphibians, sauropsidans (including birds), and mammals and occupied most terrestrial and aerial niches. In contrast to the ~30,000 extant tetrapod species (Sahney et al. 2010), very few members of the early branching (nontetrapod) sarcopterygian lineages have survived until today: two species of coelacanths (*Latimeria*), and the Australian (*Neoceratodus forsteri*), South American (*Lepidosiren paradoxa*), and African (*Protopterus*; four species) lungfishes (Nelson 2006). The phylogenetic relations among coelacanth, lungfish, and tetrapods have been debated for decades, up to the point to be considered an "irresolvable trichotomy" (Takezaki et al. 2004). Three alternative hypothetical resolutions are possible: lungfish as closest relative of tetrapods (T1), coelacanth as sister group of tetrapods (T2) and a sister group relationship between lungfish and coelacanth, both being equally distant to tetrapods (T3) (Fig. 1). For three taxa plus an outgroup, three alternative rooted trees exist and selecting the right one is thus a problem of correctly identifying the root (Rota-Stabelli and Telford 2008). Despite disagreements among different studies and data sets, the majority of morphological, paleontological, and molecular studies favored lungfish as the closest living relative of tetrapods (T1; Panchen and Smithson 1987; Meyer and Wilson 1990; Hedges et al. 1993; Zardoya and Meyer 1996; Zardoya et al. 1998; Venkatesh et al. 2001; Meyer and Zardoya 2003; Brinkmann et al. 2004a, 2004b). Some paleontological analyses provided support for coelacanth to be the sister group of tetrapods (T2; Fritzsch 1987; Zhu and Schultze 2001), whereas morphological, paleontological, and molecular phylogenetic analyses have also supported in some instances the clade composed of lungfish and coelacanth as the sister group of tetrapods (T3; Northcutt 1986; Chang 1991; Forey et al. 1991; Zardoya and Meyer 1996; Zardoya et al. 1998; Shan and Gras 2011). Noticeably, molecular analyses that favored the lungfish + tetrapod hypothesis could not statistically reject the latter hypothesis of a lungfish +

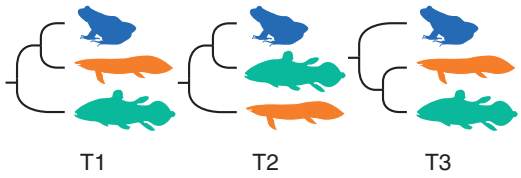

FIGURE 1. Possible phylogenetic hypotheses for early sarcopterygian evolution. Lungfish (T1) or coelacanth (T2) are alternatively the closest relatives of tetrapods, or lungfish and coelacanth form a clade that is equally close to tetrapods (T3).

coelacanth clade (Zardoya and Meyer 1996; Zardoya et al. 1998; Brinkmann et al. 2004a; Takezaki et al. 2004). From a molecular perspective, ruling out the relative phylogenetic position of both lungfish and coelacanth with respect to tetrapods has been difficult because the split among these lineages occurred rapidly (in about 10 myr or less) and a long time ago (>400 Ma) (Blair and Hedges 2005; Müller and Reisz 2005). This implies that the phylogenetic information available for resolving these nodes is scarce, and thus it can easily be confounded by nonphylogenetic signal in the data.

Systematic Error and Long-Branch Attraction

Phylogenomic analyses, thanks to their high power of resolution, represent an unprecedented opportunity to revisit the long-standing question of the closest living relative(s) of tetrapods. The sequencing of the coelacanth genome (Amemiya et al. 2013) revealed important aspects of sarcopterygian genome evolution and reported a phylogenomic tree that strongly supported lungfishes to be the closest living relatives of tetrapods, also confirmed by Liang et al. (2013) and Chen et al. (2015). Despite the ability of phylogenomic analyses to produce highly supported trees due to the reduction of sampling error, this gain in precision does not guarantee more accurate results (i.e., finding the *true* tree) if model assumptions are violated (Kumar et al. 2012). Indeed, phylogenomic data sets are known to exacerbate systematic error (e.g., Phillips et al. 2004; Jeffroy et al. 2006; Rodríguez-Ezpeleta et al. 2007; Liu et al. 2014), which in a probabilistic framework is produced by departures of the real data from the assumptions of evolutionary models. Thus, examining and reducing the effect of systematic error and increasing the robustness to violations of model assumptions are as important as reducing stochastic sampling error (Yang and Rannala 2012). Systematic error is often reflected as long-branch attraction (LBA) artifacts (Felsenstein 1978) that produce clustering of long branches irrespective of their genuine phylogenetic position. Probabilistic methods, that is, maximum likelihood (ML) and Bayesian inference (BI), are less prone to LBA artifacts than maximum parsimony or distance methods but are not immune to them (Pol and Siddall 2001; Felsenstein 2004; Susko 2015). LBA can simply occur between lineages that evolve fast or diverged a long time ago, which accumulate by chance high levels of homoplasy that obscure *bona fide* phylogenetic signal. Moreover, LBA can also result from model misspecifications, such as shifts in sequence composition for which the model does not account for (see below).

Modeling Heterogeneous Evolution

Adequate modeling of the heterogeneities in the evolutionary processes across different loci and/ or sites as well as across lineages is a key aspect of phylogenomic analysis. Commonly used models are site-homogeneous, that is, they describe a single evolutionary process that treats all sites equally. Previous research has shown that modeling of site-specific variability greatly improves the statistical fit between the model and the data and subsequent phylogenetic reconstruction. The evolutionary rate is well known to vary among sites, a phenomenon routinely modeled with a discrete gamma distribution (Yang 1996). Additionally, substitution rates and nucleotide or amino acid frequencies are known to vary among sites due to a number of factors, including solvent accessibility, secondary and tertiary structure or biological function (Le et al. 2008b). Several approaches have been used to model this heterogeneity. A common approach is data partitioning, that is, grouping genes or sites with similar evolutionary features and applying different (usually site-homogeneous) models to these subsets, best done using a statistical criterion (Lanfear et al. 2014). This approach usually relies on *a priori* defined meaningful subsets (e.g., codon positions, genes, stems and loops in secondary structure, evolutionary rate classes). Site-heterogeneous models propose a conceptually different approach where site-specific properties are accounted for individually but without requiring prior knowledge of the pattern of evolution across sites (Pagel and Meade 2005). One of the most commonly used approaches is the profile mixture (Lartillot and Philippe 2004; Le et al. 2008a), where sites are assumed to belong to different classes (profiles) that differ in their nucleotide or amino acid equilibrium frequencies. For example, the CAT model (Lartillot and Philippe 2004) uses a Dirichlet process prior to estimate the total number of profiles, as well as the affiliation of each site to a given profile. An alternative approach for amino acid data are LG4M and LG4X models (Le et al. 2012), where sites are categorized depending on their evolutionary rate, and four different replacement matrices (with different frequencies and exchangeabilities) are used for each site category. Previous research has shown that site-heterogeneous models usually have a much better fit (given enough data) and are less sensitive to LBA problems than site-homogeneous models (Baurain et al. 2007; Lartillot et al. 2007; Le et al. 2012).

Sequence Composition and Missing Data

Compositional heterogeneity among lineages (i.e., nonstationary composition) is another common

source of systematic error and has been shown to even force the clustering of sequences with similar composition (Lockhart et al. 1992; Mooers and Holmes 2000; Phillips and Penny 2003; Hassanin et al. 2005). Commonly used models assume that the process of sequence evolution is not only stationary, but also reversible and globally homogeneous (SRH; Bryant et al. 2005; but see e.g., Jayaswal et al. 2014). This implies that the marginal probabilities of each nucleotide or amino acid (stationarity) and the substitution rates (homogeneity) are constant throughout the tree and that the direction of evolution can be ignored (reversibility) (for a detailed discussion see e.g., Jermiin et al. 2008). The presence of among-lineage compositional heterogeneity supposes that both the stationarity and reversibility assumptions are violated. Depending on the magnitude of this heterogeneity, models not accounting for this effect might estimate very long branches and produce an LBA artifact. An additional important source of error in phylogenomics is the presence of missing data. Even though probabilistic methods can in theory accommodate missing data (Felsenstein 2004), several previous studies found that nonrandom missing data can negatively impact phylogenetic inference (Hartmann and Vision 2008; Lemmon et al. 2009; Dell'Ampio et al. 2013; Roure et al. 2013). No agreement exists on whether and how highly incomplete taxa might affect phylogenetic results (e.g., Wiens 2003; Dwivedi and Gadagkar 2009; Hejnol et al. 2009; Wiens and Morrill 2011; Dell'Ampio et al. 2013) and their impact on large-scale analyses is still not well understood (Yang and Rannala 2012). Nevertheless, there is often extensive missing data in phylogenomic matrices (up to >90%; e.g., Driskell et al. 2004; Dunn et al. 2008; Hejnol et al. 2009; Streicher et al. 2015) and this has the potential to intensify systematic errors (Roure et al. 2013).

Concatenation Versus Species Tree Methods

Data concatenation methods assume that the same phylogenetic history underlies all genes, but this assumption might be violated in the presence of phenomena such as deep coalescence (Degnan and Rosenberg 2006). Concatenation methods have the ability to capture emergent informative positions that can go unnoticed when loci are separately analyzed (Gatesy and Baker 2005; Townsend et al. 2011). This approach is particularly important when phylogenetic information is limited, such as among the earliest sarcopterygian branching events. However, concatenation methods can fail to recover the species tree in the presence of high levels of discordance among gene trees (Mossel and Vigoda 2005; Kubatko and Degnan 2007; Salichos and Rokas 2013). The combination of short internal and long external branches has been shown to mislead concatenation methods, either because of the presence of deep coalescence (Xi et al. 2014) and/or the confounding effect of homoplasy of fast-evolving sites in the long branches (Townsend et al. 2012; Chen et al. 2015). Several previous studies have advocated for the superiority of coalescent methods over concatenation, even for ancient speciation events (Song et al. 2012; Kumar et al. 2013; Zhong et al. 2013; Xi et al. 2014).

Here, we use available vertebrate genomes together with newly generated lungfish sequence data to revisit the controversial problem of finding the closest living relative(s) of land vertebrates. Our study is the first to include genomic data for all three extant lungfish lineages, of which the Australian lungfish (*Neoceratodus*) is particularly relevant as it diverged early from all other lungfishes (~180 Ma; Heinicke et al. 2009). We take advantage of the power of genome-scale data to resolve internodes at the base of sarcopterygians and carefully examine whether the support for alternative resolutions is produced by the misleading effect of systematic error, incomplete lineage sorting or infrequent amino acid replacements. In the genomic era, having enough data to resolve phylogenetic questions is rarely the problem, but rather how to best analyze it, dissect phylogenetic signal and examine systematic sources of error to assess the robustness of the obtained estimates. For recalcitrant phylogenetic problems, notably old and fast speciation events, such analyses become even more important because the genuine but faint phylogenetic signal can easily be overcome by nonphylogenetic noise (Philippe et al. 2011). For this reason, the analytical pipeline outlined in this study could help resolve other similarly recalcitrant nodes in the tree of life.

Materials and Methods

Transcriptome Sequencing and Assembly

Because lungfishes have huge (50–130 Gb) and repetitive genomes that greatly complicate their sequencing and assembly, we used RNA-seq to obtain exonic data. A single specimen of *L. paradoxa* (from the pet trade) was euthanized with MS-222 and the following tissues dissected and stored in RNAlater (Ambion, Austin, TX, USA): brain, caudal fin, head, heart, gut, gonads, liver, and muscle. An appropriately stored (RNAlater) fin clip from a living *N. forsteri* was obtained from the Berlin Zoo (Germany). Total RNA was extracted with Trizol (Invitrogen, Carlsbad, CA, USA) according to the manufacturer's recommendations, treated with DNAse I and purified in spin columns (Qiagen, Hilden, Germany). Quantification and integrity was assessed using a Ribogreen assay and Bioanalyzer 2100 (Agilent Technologies, Waldbronn, Germany). All samples had RNA integrity values above 8.0. Individual libraries were prepared using the Illumina TruSeq RNA Sample Preparation Kit after poly-T selection, according to the manufacturer's instructions. All barcoded libraries were pooled into one Illumina lane and sequenced using HiSeq2000 2 × 100 bp technology. Transcriptomic data from *Protopterus annectens* (brain, liver, and kidney; Amemiya et al. 2013) was downloaded from NCBI's SRA (SRR505723–SRR505725). SeqPrep (St John 2013)

was used to remove any remaining adapter sequence and merge overlapping read pairs, and prinseq (Schmieder and Edwards 2011) to filter and trim reads by quality (settings: -exact_only -min_len 40 -min_qual_mean 20 -lc_method dust -lc_threshold 32 -trim_qual_right 30 -trim_ qual_window 1 -trim_qual_step 1).

Because assemblies are known to vary significantly depending on the choice of software and kmer (e.g., Bradnam et al. 2013; Yang and Smith 2013), we used different algorithms and parameterizations to reconstruct the highest possible number of full-length transcripts as a first premise to obtain long informative alignments. Trinity r20131110 (Grabherr et al. 2011) used a fixed kmer size of 25 and three algorithms (default, "cufffly" and "pasafly"), of which the latter two maximize transcript length by merging compatible transcripts. Oases v.2.0.8 (Schulz et al. 2012) was used with different kmer values (21, 31, 41, 51, 61, and 71), and we further created a nonredundant merged assembly from previous single-kmer assemblies with kmer = 27 (Schulz et al. 2012). We assessed the performance of the different assembly strategies by comparing the (i) number of transcripts, (ii) sequence coverage with respect to the coelacanth proteome v.1.73 (Haas et al. 2013), and (iii) number of assembly chimeras (Yang and Smith 2013; see Supplementary Figs. S1–3; available on Dryad at http://dx.doi.org/10.5061/dryad.gd74v). Newly generated raw reads are available in NCBI's SRA (SRR3632078–SRR3632086).

Phylogenomic Data Set Construction

Figure 2 summarizes the main steps of our analytical pipeline. Single-copy genes were searched in orthoDB v.8 (Kriventseva et al. 2015) and we required their presence in >80% of the following 20 species that represent main vertebrate lineages: armadillo (*Dasypus novemcinctus*), chicken (*Gallus gallus*), clawed frog (*Silurana tropicalis*), coelacanth (*Latimeria chalumnae*), dog (*Canis familiaris*), elephant (*Loxodonta africana*), human (*Homo sapiens*), lamprey (*Petromyzon marinus*), lizard (*Anolis carolinensis*), mouse (*Mus musculus*), opossum (*Monodelphis domestica*), platypus (*Ornithorhynchus anatinus*), pufferfish (*Takifugu rubripes*), spotted gar (*Lepisosteus oculatus*), tammar wallaby (*Macropus eugenii*), tilapia (*Oreochromis niloticus*), turkey (*Meleagris gallopavo*), turtle (*Pelodiscus sinensis*), zebrafinch (*Taeniopygia guttata*), and zebrafish (*Danio rerio*). This filtering rendered 5071 orthogroups, whose sequences were downloaded from ENSEMBL v.75 (Cunningham et al. 2015). For each orthogroup, homologous data was retrieved from the Elephant shark proteome (*Callorhinchus milii*; Venkatesh et al. 2014) (BLASTP) and lungfish transcriptomes (TBLASTN) using a reciprocal best BLAST hit procedure that successfully identified >89% putative ortholog hits. To maximize the recovery or homologs for lungfishes, TBLASTN searches were performed against the collection of transcripts assembled by the different software and parameterizations. Elephant shark was selected as an appropriate outgroup for our phylogeny because of its slow-evolving genome (Venkatesh et al. 2014) that reduces the probability of LBA. Multiple sequence alignment was performed with MAFFT v.7.158 (Katoh and Standley 2013) using an iterative refinement algorithm (L-INS-i). Poorly aligned positions were removed using the "-strict" method in trimAl v.1.4 (Capella-Gutiérre et al. 2009). Individual gene alignments were visualized in Seaview v.4.4.3 (Gouy et al. 2010) to identify and remove problematic sequences that were too divergent (putative instances of paralogy, misalignments, misannotations, or reading frame shifts). A second quality control aimed to remove putative lungfish paralogs, by first identifying instances of nonmonophyletic sarcopterygians or lungfishes in individual gene trees, followed by careful visual examination of alignments not passing the above test (see Chen et al. 2015 for a similar strategy). This visual examination allowed the discrimination of likely paralogs (which were eliminated) from cases where monophyly failed due to limited signal in single genes alignments. For this analysis, gene trees were estimated in RAxML v.8.1.16 (Stamatakis 2014) by 100 independent ML searches under LG+G and monophyly tests were implemented in a custom Perl script. We further applied a stringent taxonomy filter (only alignments containing all three lungfishes, coelacanth, tetrapods, and nonsarcopterygian outgroups were retained) and short alignments (<100 amino acids) were discarded. The resulting 2960 gene alignments were concatenated using FASconCAT-G (Kuck and Longo 2014) to generate the 2960 data set.

A second matrix was built by complementing the data set of Amemiya et al. (2013) with the newly generated lungfish transcriptomes (*Lepidosiren* and *Neoceratodus*), using the reciprocal BLAST procedure mentioned above. Gene alignments were estimated with MAFFT, and trimmed to conform to the length of the original gene alignments (trimAl "-gt 0.9"). Following the strategy outlined above, putative paralogs were removed after visualization of single gene alignments and monophyly test on ML gene trees. Two hundred and fifty-one gene alignments were concatenated into the 251 data set, containing a total of 24 taxa and 100,593 aligned amino acid positions and being 92.8% complete (nonambiguous amino acids other than gaps and missing data).

The software MARE v.0.1.2 (Meyer et al. 2011) was used to create an optimized subset of most-informative taxa and genes. Briefly, MARE calculates the information content as tree-likeness using quartet mapping (Nieselt-Struwe and von Haeseler 2001) and reduces the original matrix by iteratively dropping the least informative genes and sequences according to an optimality criterion. We used increasing weighting for information content ($\alpha = 3, 4, 5$) to create shorter and more informative matrices, both using (i) the concatenated 2960 data set and (ii) the 251 data set (Supplementary Figs. S4 and S5 available on Dryad). For

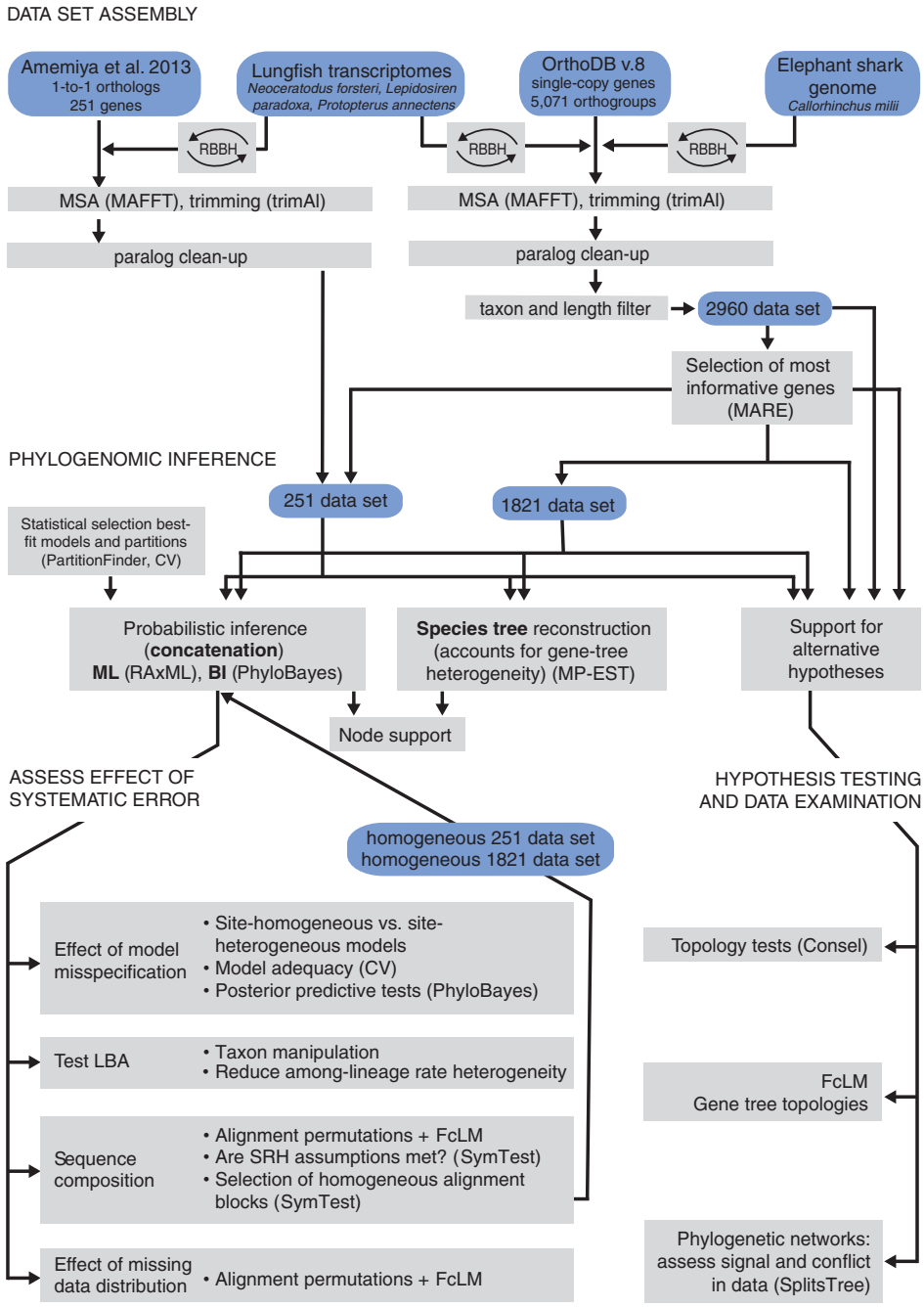

FIGURE 2. Analytical pipeline showing main steps of data set assembly, phylogenomic analysis, assessment of systematic error, hypothesis testing, and additional data examination techniques. Rectangles denote processes and round-corner rectangles, data sets. See main text for full details. BI = Bayesian inference; CV = cross-validation; FcLM = four-cluster likelihood mapping; ML = maximum likelihood; MSA = multiple sequence alignment; RBBH = reciprocal best BLAST hit procedure; SRH = stationary, reversible and homogeneous.

further phylogenetic analyses, we selected the shortest and most informative ($\alpha = 5$) sub-matrix derived from the initial 2960 genes. This matrix contains 1821 genes, 733,057 aligned amino acid positions, and 23 taxa (lamprey was dropped due to low information content) and it is 89.3% complete. Hereafter, we refer to this optimized matrix as the 1821 data set.

Testing the Effect of Compositional Heterogeneity and Missing Data

Sequence composition was studied using the three matched pairs tests of symmetry (see Supplementary Information available on Dryad) implemented in SymTest v.2.0.37 (available from L.S. Jermiin upon

request). Given the nonhomogeneity of both the 251 and 1821 data sets shown by matched pairs tests (see Supplementary Fig. S6 available on Dryad), we used a sliding window approach to identify sequence blocks that are consistent with evolution under stationary, reversible, and homogeneous (SRH) conditions (see introduction and Supplementary Information available on Dryad). Using a window size of 3000 and step size of 300, we selected a total of 191 nonoverlapping blocks and 573,000 positions (homogeneous 1821 data set) and 27 blocks and 79,593 positions (homogeneous 251 data set) that were concatenated into two matrices, respectively. To gain insight into the effect of discarding alignment regions that strongly violated homogeneity assumptions, the new matrices of concatenated homogeneous blocks were separately subjected to ML tree inference and branch support (see below) using partitioned best-fit JTT+G models selected according to the Akaike Information Criterion (AIC) in ProtTest v.3.2 (Abascal et al. 2005; Darriba et al. 2011). To compare jackknife proportions to those obtained from the original 1821 data set (see below), 100 gene jackknife replicates of comparable size (~40,000 amino acids; 13 blocks) were generated and analyzed under ML. The heterogeneous blocks discarded from both original 1821 and 251 data sets were also concatenated and subjected to ML analyses.

We used a second approach to assess the effect of compositional heterogeneity and missing data specifically for the early sarcopterygian branching events. This approach is based on four-cluster likelihood mapping (FcLM; Strimmer and von Haeseler 1997) and follows Misof et al. (2014). Briefly, we permuted both 1821 and 251 data sets by eliminating phylogenetic signal but retaining compositional noise or missing data distribution and compared the support of permuted alignments for alternative hypotheses (Fig. 1) to that obtained from the original alignments. This approach aimed to explore whether the presence of compositional heterogeneity, distribution of missing data, or both in synergy might be affecting the resolution of early sarcopterygian relationships (see Supplementary Information available on Dryad for details).

Phylogenomic Analyses

BI was performed with PhyloBayes MPI v.1.5. (Lartillot et al. 2013) without constant sites ("-dc" option), running two independent MCMC chains until convergence, sampling every cycle. The 1821 data set was analyzed using a gene jackknifing approach (e.g., see Delsuc et al. 2008) by generating 50 alignment replicates, each consisting of 100 genes sampled without replacement. The 50 gene jackknife replicates (two chains each) were run under best-fit CAT-GTR+G model (see below) and topology and branch lengths summarized over the 100 MCMC chains. To test the effect of the assumed evolutionary model under BI, the 251 data set was analyzed under site-homogeneous (LG+G) and site-heterogeneous (CAT+G, CAT-GTR+G) models. Convergence of analyses was checked *a posteriori* using the tools implemented in PhyloBayes (maxdiff <0.1, maximum discrepancy <0.1, and effective size >100; Supplementary Table S1 available on Dryad). ML reconstruction relied on the rapid hill-climbing algorithm implemented in RAxML v.8.1.16, starting from 100 maximum parsimony trees. We used best-fit models and partitions statistically selected with PartitionFinder with aid of the AIC (Akaike 1973) and the "rcluster" clustering method (Lanfear et al. 2014). The 251 data set was additionally analyzed under unpartitioned LG+G (Le and Gascuel 2008) and unpartitioned LG4X (Le et al. 2012). For comparison with BI, we ran additional ML analyses on 100 gene jackknife replicates derived from the larger 1821 data set, using best-fit models ("–m PROTGAMMAAUTO" option in RAxML) and summarized the resulting 100 trees by majority-rule consensus. The model LG was preferred for analyzing unpartitioned alignments because it has been shown to consistently outperform models like WAG or JTT, particularly in the presence of high among-site rate heterogeneity (Le and Gascuel 2008).

The performance of LG+G, CAT+G and CAT-GTR+G models was assessed using a 10-fold cross-validation in PhyloBayes 3.3e (Lartillot et al. 2009). For computational tractability, we used subsamples of 10,000 nonconstant positions randomly drawn from the original matrices. For both data sets, model cross-validation clearly favored CAT-GTR > CAT > LG (see Supplementary Table S2 available on Dryad). To further investigate the overall model adequacy, we used two posterior predictive tests implemented in PhyloBayes aimed to understand how well each model can account for site-specific biochemical patterns ("ppred –div") and anticipate homoplasy ("ppred –sat"). In both cases, the observed amino acid diversity and homoplasy values are compared against a posterior predictive distribution generated from post-burnin MCMC states (Lartillot et al. 2007). In the ML framework, node support was assessed by 500 pseudoreplicates of nonparametric bootstrapping (Felsenstein 1985) and SH-like aLRT support (SHS; Guindon et al. 2010). Although based on different principles, both bootstrapping and SHS have been shown to produce comparable results, SHS values above 0.8–0.9 being considered strong support (Guindon et al. 2010). For gene jackknife replicates, we calculated the proportion of times a given bipartition was recovered by each replicate. For ML searches, the number of times each bipartition is recovered among the 100 independent ML searches was also recorded to identify possible plateaus in the likelihood surface.

To assess and correct the LBA artifact found in the 251 data set (see "Results" section), additional ML analyses were performed after (i) eliminating fast-evolving actinopterygians, (ii) breaking up the long actinopterygian branch by the addition of the spotted gar, (iii) eliminating genes showing signs of high among-lineage rate heterogeneity, and (iv) removing fast-evolving positions. For approach (ii), orthologous sequences from the spotted gar

were identified as specified above for the Elephant shark. For approach (iii), genes were ranked according to the extent of among-lineage rate heterogeneity, estimated as the largest difference of mean ML distances among actinopterygians, lungfishes + coelacanth and tetrapods. The genes with the highest among-lineage rate heterogeneity were excluded progressively (by groups of 10%) to generate nine submatrices of decreasing size, which were analyzed by ML under best-fit models. For approach (iv), the alignment positions of the 251 data set were divided into 20 bins by their evolutionary rate using TIGER v.1.02 (Cummins and McInerney 2011) and the fastest evolving bins were excluded progressively to generate 10 submatrices of decreasing size that were analyzed under ML.

Topology and Relative-Rate Tests

The three competing phylogenetic hypotheses (Fig. 1) were tested using the well-accepted backbone phylogeny of vertebrates (Fig. 3) and the following alternative sister groups: lungfish + tetrapods (T1), coelacanth + tetrapods (T2), lungfish + coelacanth (T3). Site-wise log-likelihoods were estimated with RAxML under LG+G. Consel v.0.1i (Shimodaira and Hasegawa 2001) was used to perform KH (Kishino and Hasegawa 1989), SH (Shimodaira and Hasegawa 1999), and AU (Shimodaira 2002) tests with one million multiscale bootstrap replicates. Given that the test hypotheses are defined *a priori*, the KH test is most appropriate. Note also that the SH and AU tests might be biased if other similarly "plausible" topologies that were not included in the test also exist (Goldman et al. 2000). Topology tests were performed on nine data sets: (i) the initial 2960 matrix, (ii) reduced matrices after applying MARE with $\alpha = 3$, (iii) $\alpha = 4$, (iv) $\alpha = 5$, (v) the 251 data set, (vi) reduced matrices after applying MARE with $\alpha = 3$, (vii) $\alpha = 4$, (viii) $\alpha = 5$, and (ix) the original concatenated matrix from Amemiya et al. (2013).

The faster evolutionary rates of actinopterygians and tetrapods with respect to lungfishes and coelacanth were tested for both 1821 and 251 data sets using relative-rate tests. We used RRTree (Robinson-Rechavi and Huchon 2000), with cartilaginous fishes as outgroup and correcting for phylogenetic relatedness with the BI topologies.

Gene Tree Heterogeneity and Multispecies Coalescent Species Trees

For both 1821 and 251 data sets, individual gene trees were estimated using RAxML under best-fit models. To assess the topological variation among estimated gene trees, we quantified the prevalence of topologies relevant for different early sarcopterygian branching orders (Fig. 1) using a custom Perl script. For both 1821 and 251 data sets, we inferred species trees using the pseudo-likelihood multispecies coalescent method implemented in MP-EST v.1.4 (Liu et al. 2010) using all previously estimated gene trees as input. MP-EST runs were initialized with random trees (default), as well as with all three competing topologies (Fig. 1). To discriminate between real conflict among gene trees and random stochastic error, we ran additional MP-EST analyses on subsets of genes that had the power to recover robust, yet conflicting relationships among the main sarcopterygian lineages. In practice, we estimated SHS for individual gene trees with RAxML, and for each tree, we extracted the support value for the most recent common ancestor of lungfishes and coelacanth. Trees receiving strong support for this node were selected (>0.85; analogous to $>75\%$ bootstrap support; Guindon et al. 2010). A total of 412 and 52 gene trees had high support for this node in the 1821 and 251 data sets, respectively. Robustness of species trees was assessed with 100 pseudoreplicates of multilocus bootstrap (Seo 2008), using Phybase v.1.3 (Liu and Yu 2010) and RAxML to estimate gene trees under best-fit models.

Phylogenetic Networks and Alignment Site Patterns

To visualize the tree-likeness of the data sets as well as to identify conflicting signals, we computed neighbor-nets for both 1821 and 251 data sets using SplitsTree v.4.13.1 (Huson and Bryant 2006). We further analyzed alignment site patterns in the 1821 and 251 data sets to investigate the types of molecular evidence supporting each of the three competing hypotheses under study (Fig. 1). To do this, we first eliminated invariable and undetermined positions to avoid ambiguities in site patterns and used a custom Perl script implementing the method of Wägele and Rödding (1998) to identify alignment positions that alternatively support the following sister groups: lungfishes + tetrapods, coelacanth + tetrapods, or lungfishes + coelacanth. Site patterns were then classified into (i) symmetrical or binary patterns that define two clades both with uniform but different character states, (ii) asymmetrical, where the clade of interest has a uniform state different from those in the remaining species that are variable, and (iii) noisy positions, where the state in the clade of interest is uniform and can occur in the remaining species up to a frequency of 0.25. For each hypothesis, we estimated the proportions of each amino acid supporting the clade of interest and compared it against the overall amino acid proportions observed in each data set. We also compared the proportion of replacements that support each hypothesis by correlating them against the replacement rates in the best-fit JTT model (selected by AIC in ProtTest). For the 1821 data set, these comparisons were performed at the level of each amino acid and replacement, as well as after grouping them by biochemical properties into hydrophilic (D, E, K, N, P, R), neutral (G, H, Q, S, T), hydrophobic (A, C, Y) and very hydrophobic (F, I, L, M, V, W), following Le and Gascuel (2010). Due to the relatively low number of replacements found, comparisons for the 251 data set were only performed at the level of

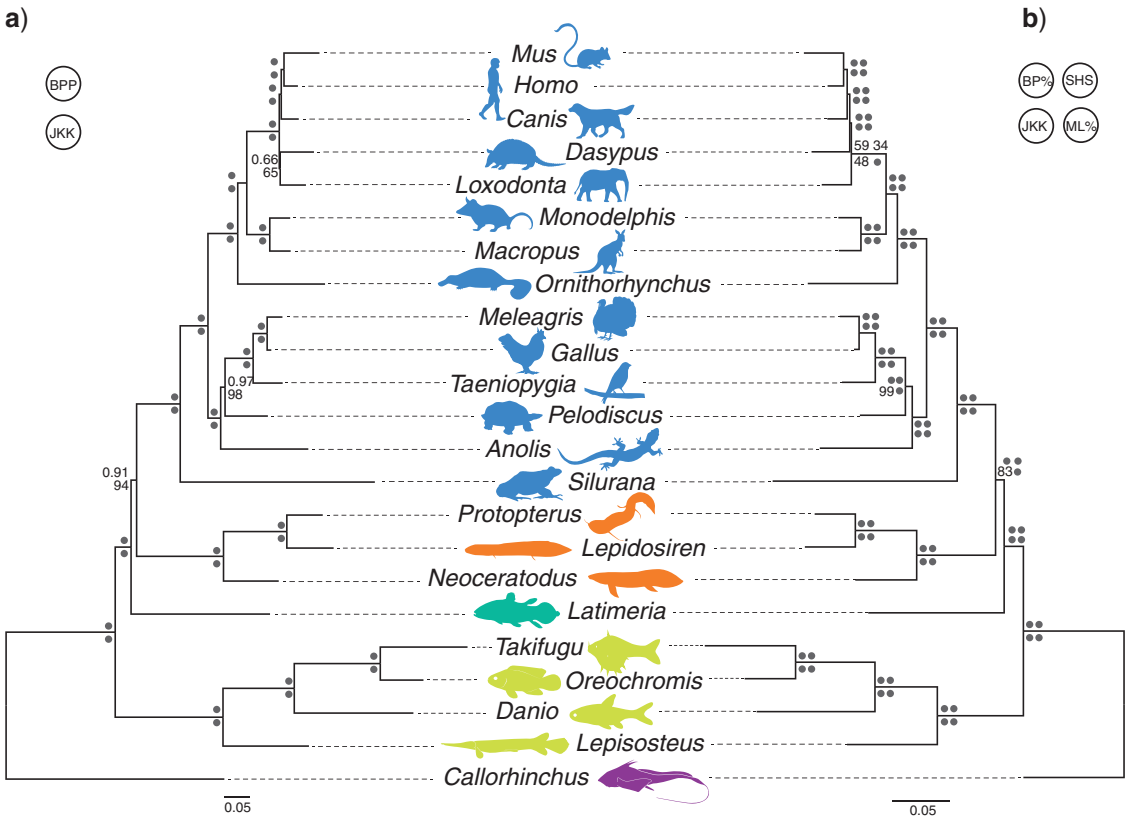

FIGURE 3. Phylogenetic tree reconstructed from the 1821 data set by (a) PhyloBayes with CAT-GTR model and (b) partitioned RAxML with best-fit partition scheme and models. Branch support is shown as Bayesian posterior probabilities (BPP), nonparametric bootstrap proportions (BP%), and SH-like support (SHS). The proportion of jackknife replicates (JKK) and the proportion of bipartitions obtained from 100 ML searches (ML%) are also shown. Filled circles represent maximal support. Scale bar is in substitutions per site.

amino acid biochemical groups. All statistical tests were performed in R (R Development Core Team 2009).

RESULTS

Gene Concatenation Favors Lungfish as Closest Living Relative to Tetrapods

All concatenated analyses produced highly supported and congruent topologies, showing differences only in the relative position of lungfishes and coelacanth and the root of placental mammals (Figs. 3 and 4). The larger 1821 data set (1821 genes) favored lungfishes as closest to tetrapods in both BI and ML analyses, as did the BI analysis of the smaller 251 data set (251 genes). All these relationships received strong statistical support from Bayesian posterior probabilities (BPP $>$ 0.90), nonparametric bootstrap proportions (BP $=$ 100%) and SH-like support (SHS $=$ 1.00), as well as congruence between gene jackknife replicates of the larger data set (94% for BI and 83% for ML) (Figs. 3 and 4a). In contrast, the ML analysis of the smaller 251 data set under best-fit models and partitions (Fig. 4b) favored the alternative sister group of lungfishes and coelacanth, although with low node support (BP $=$ 40%, SHS $=$ 0.10). Note that the internodes at the origin of sarcopterygians are very short in all analyses (Figs. 3 and 4).

The incongruence between BI and ML for the 251 data set is likely the result of an LBA artifact between faster evolving actinopterygians and tetrapods with respect to coelacanth and lungfishes (relative-rate test $p < 1 \times 10^{-6}$). This LBA is also visible in the phylogenetic network as a strong reticulation among actinopterygians and tetrapods (amphibians), but not for the 1821 data set (Supplementary Figs. S7 and S8 available on Dryad). CAT site-heterogeneous models (CAT, CAT-GTR) congruently supported hypothesis T1 with maximal BPP support (Figs. 4a and 5a), whereas the LG4X site-heterogeneous model recovered the alternative T3 hypothesis (Fig. 5b), as did site-homogeneous models (LG) analyzed in both ML and BI frameworks (Fig. 5c). The per-site amino acid diversity and homoplasy predicted by both CAT and CAT-GTR models did not significantly differ from those observed in either the 1821 or 251 data sets ($p > 0.72$), whereas LG significantly overestimated amino acid diversity ($p = 0$) and underestimated homoplasy ($p = 0$). The number of substitutions predicted by all three models did not differ from the observed one ($p > 0.26$) in either of the tested data sets (see Supplementary Table S3 available on Dryad for full results). ML partitioned analyses after (i) removing fast-evolving actinopterygians, (ii) breaking up the long

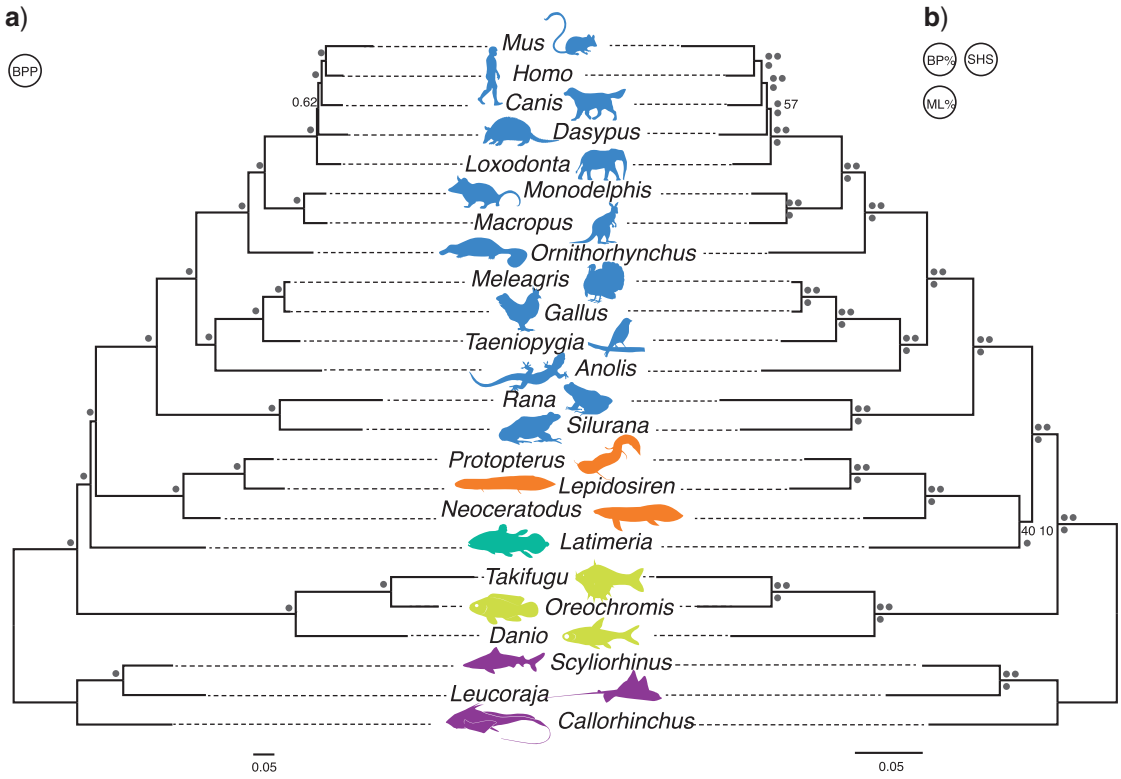

FIGURE 4. Phylogenetic tree reconstructed from the 251 data set by (a) PhyloBayes with CAT-GTR model and (b) partitioned RAxML with best-fit partition scheme and models. Branch support is shown as Bayesian posterior probabilities (BPP), nonparametric bootstrap proportions (BP%), and SH-like support (SHS). The proportion of bipartitions obtained from 100 ML searches are also indicated (ML%). Filled circles represent maximal support. Scale bar is in substitutions per site.

actinopterygian branch by adding the spotted gar, or (iii) eliminating genes displaying high among-lineage rate heterogeneity correctly eliminated the LBA artifact and recovered T1 with moderate to high support (Fig. 5d–f). Removing only 10% of the genes showing the strongest among-lineage rate heterogeneity was enough to recover T1 (albeit with low support, SHS = 55) and the highest support for T1 was obtained after the exclusion of 70% of such genes (Supplementary Fig. S9 available on Dryad). The ML trees inferred after the elimination of fast-evolving sites in the first, second, and third fastest categories favored T3 with rather low support (Fig. 5g) and further removal of site categories produced an overall lack of resolution for the whole vertebrate phylogeny (not shown). Even though FcLM analyses did not reveal a strong effect of sequence composition in favoring any particular hypothesis (see below), an ML reconstruction after eliminating compositionally deviant alignment regions (homogeneous 251 matrix; selected with SymTest) favored the T1 hypothesis with moderate support (Fig. 5h). Moreover, an ML analysis of the excluded heterogeneous regions strongly favored the T3 hypothesis (BP = 86%; Supplementary Figs. S10 and S11 available on Dryad). In the case of the 1821 data set, the exclusion of compositionally most deviant alignment regions did not alter the ML recovery of T1, but slightly increased the agreement among gene jackknife replicates (from 83% to 86%; Supplementary Fig. S12 available on Dryad). An ML analysis of the heterogeneous regions excluded from the 1821 data set recovered T1 with high support (Supplementary Fig. S13 available on Dryad). Branch lengths estimated by ML on trees derived from the homogeneous submatrices were consistently shorter than those estimated on the full data set (in >95% of the branches), and on average trees were ~7.4% and ~9.3% shorter, respectively, for the homogeneous 1821 and homogeneous 251 data sets.

The study of alignment site patterns identified a small proportion of binary, asymmetrical, and noisy sites supporting any of the three hypotheses under consideration. In both the 1821 and 251 data sets, positions supporting T3 (0.90% and 1.21%, respectively) outnumbered those in favor of T1 (0.47% and 0.35%) and T2 (0.22% and 0.09%). In all comparisons, noisy positions were most frequent, followed by binary and asymmetrical positions. Overall, the amino acids alternatively supporting each of the three tested clades were not significantly different from their overall proportions in the data sets (chi-square test $p > 0.05$). The only significant difference was found for the 1821 data set, where no hydrophobic amino acids (A, C, Y) were present in the asymmetrical class supporting T1 (chi-square test $p < 0.01$). In the 1821 data set, T1 is supported by a higher proportion of hydrophilic

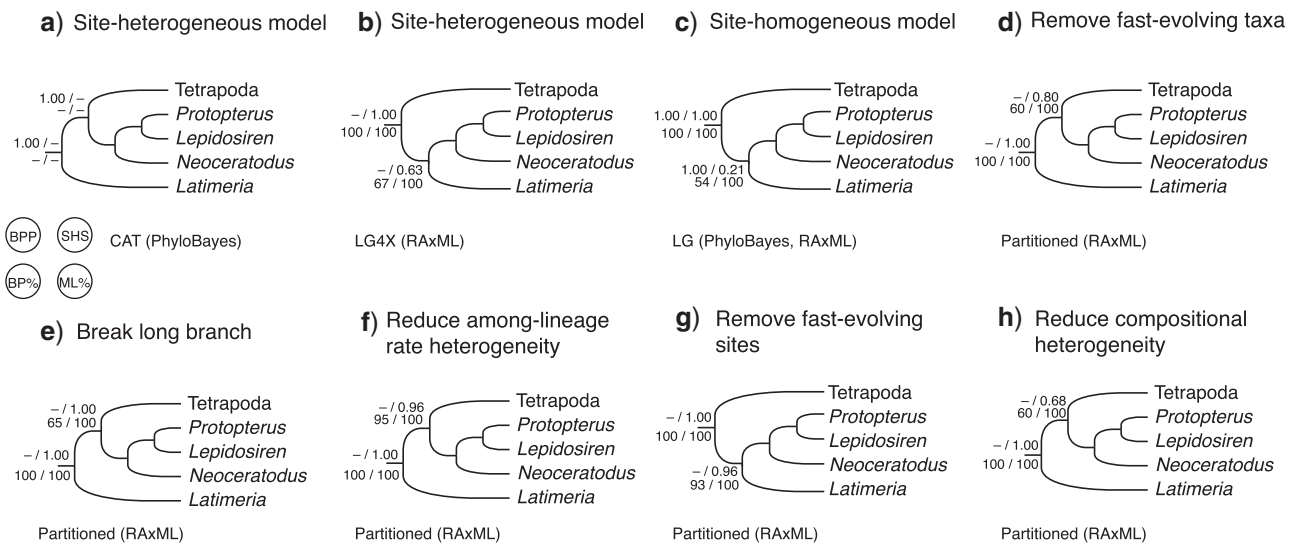

FIGURE 5. Effect of model misspecification and strategies to overcome LBA in the 251 data set to resolve the relationships among coelacanth (*Latimeria*), lungfishes (*Neoceratodus, Lepidosiren, Protopterus*), and tetrapods. (a–c) Accounting for site heterogeneity. (d) Removal of fast-evolving species. (e) Addition of slow-evolving and early branching species. (f) Exclusion of genes with high among-lineage rate heterogeneity. (g) Removal of fast-evolving sites. (h) Reduction of compositional heterogeneity. Figure shows only relevant subtrees, as all other nodes agree with Figure 4 (full trees are available in Supplementary Figs. S10, S14–S21 available on Dryad). For each analysis, the applied models (and software) are shown. Numbers at nodes represent (clockwise) support from Bayesian posterior probabilities (BPP), SH-like support (SHS), nonparametric bootstrap proportions (BP%), and proportion of bipartitions recovered by 100 ML searches (ML%).

positions than T3 (respectively, 39.48% vs. 28.80%) and lower proportion of very hydrophobic ones (29.47% vs. 38.60%; Supplementary Fig. S22 available on Dryad), although these differences were not significant (chi-square test $p > 0.05$). In the 1821 data set, the proportion of amino acid changes supporting all hypotheses were highly correlated with the JTT replacement rates, with slightly higher correlations of T1 and T3 over T2 ($r = 0.89$ and 0.90 vs. 0.79 for the rates among the four amino acid classes). In the 251 data set, the proportion of amino acid replacements that support hypotheses T1 and T3 were significantly correlated with the JTT replacement rates (Pearson's correlation $p < 0.05$; $r = 0.91$ and 0.90, respectively), but not replacements supporting T2 (Pearson's correlation $p > 0.05$; $r = 0.33$). In this case, T2 is supported by only six amino acid changes, five of which are hydrophilic (D, E, K, N, P, R).

The Multispecies Coalescent Favors the Lungfish + Tetrapod Hypothesis

Individual gene trees reconstructed by ML showed pervasive heterogeneity in their topologies. The monophyly of two well-accepted clades such as tetrapods and sarcopterygians was recovered by a relatively low number of genes (32% and 17% of the genes from the 1821 set, and 57% and 25% from the 251 set, respectively) and <15% of them simultaneously recovered the monophyly of sarcopterygians, lungfishes, and tetrapods (Fig. 6a). Among this latter group of gene trees, the majority favored T1, followed by T3 and T2. In the larger collection of gene trees (1821 data set), support for these hypotheses was, respectively, 5.7%, 3.6%, and 3.5%, whereas in the smaller set of gene trees (251 data set) support was 10.0%, 9.1%, and 4.0% (Fig. 6a). For the larger data set, 412 out of 1821 genes (23%) reconstructed the most recent common ancestor of lungfishes and coelacanth with high support (SHS $\geqslant 0.85$), whereas for the smaller data set 52 out of 251 genes (21%) received high support.

Multispecies coalescent analyses reconstructed topologies that were congruent with the concatenated ML analyses. For the 1821 data set, MP-EST supported T1 with high support from multilocus bootstrap (BP = 99%; Fig. 6b) and this result was robust to the use of different starting trees. For the smaller 251 data set, the T3 alternative was favored with substantial support (BP = 76%; Supplementary Fig. S23 available on Dryad). For both data sets, MP-EST analyses on the subsets of genes with high SHS support confirmed the topologies obtained respectively by their corresponding larger gene sets (Supplementary Figs. S24 and S25 available on Dryad). These sets of 52 and 412 genes had on average higher substitution rates (significant only for the set of 412 genes, Wilcoxon Rank test $p < 0.05$), and were not enriched in any particular GO term (assessed with Fisher's exact test with FDR < 0.05 in Blast2GO; Conesa et al. 2005). Note that the internodes at the origin of sarcopterygians are very short in all coalescent trees, although they are not directly comparable to those estimated by concatenation because they also depend on effective population sizes.

Topology Tests Congruently Reject Alternative Resolutions for Basal Sarcopterygian Relationships

All tested data sets strongly rejected T2 with all three performed topology tests. The topology tests were also

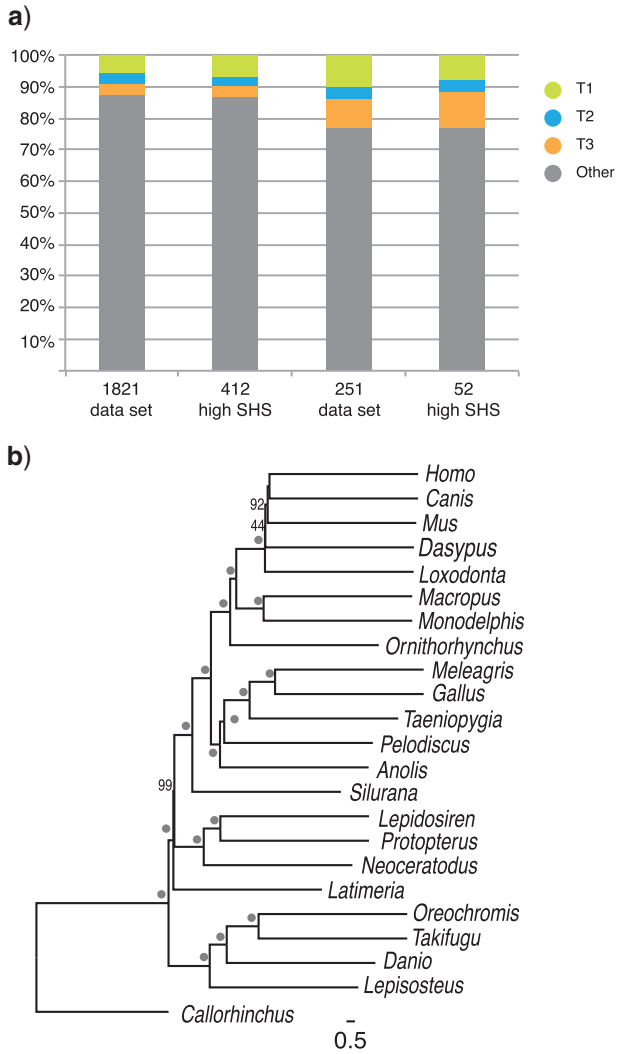

FIGURE 6. Gene tree heterogeneity and multispecies coalescent tree reconstruction. (a) Histogram shows the proportion of gene trees with fully resolved sarcopterygian relationships that support T1, T2, or T3 (*sensu* Fig. 1) for four collections of trees: gene trees from the 1821 and 251 data sets, as well as subsets of highly supported gene trees derived from them, respectively, the 412 high SHS and 52 high SHS sets. Gene trees that failed to simultaneously recover the monophyly of all sarcopterygians, tetrapods, and lungfishes are also shown in gray. (b) MP-EST tree estimated from the 1821 data set. Numbers at nodes are proportion of multilocus nonparametric bootstrapping and filled circles represent 100%. Scale bar is in coalescent units (note that branch lengths estimated by MP-EST in this case are only meaningful for internal branches).

congruent in the rejection of the T3 on the large data sets resulting from the matrix reduction under different levels of stringency, which includes the 1821 data set (Table 1). Neither the full 2960 data set, the 251 data set, nor the original data of Amemiya et al. (2013) could reject T3. The use of MARE produced a significant increase in the phylogenetic information content (P) of matrices, e.g., from $P=0.60$ in the 2960 data set to $P=0.70$ in the 1821 data set allowing the rejection of both T2 and T3. In the case of the 251 data set, MARE increased

TABLE 1. Results of KH, SH, and AU topology tests for six data sets

Data set	T1			T2			T3		
	KH	SH	AU	KH	SH	AU	KH	SH	AU
2960 data set	0.918	0.974	0.918	**0**	**0**	**2e-07**	0.082	0.125	0.082
after MARE ($\alpha=3$)	1.000	1.000	1.000	**0**	**0**	**5e-08**	**0**	**0**	**3e-07**
after MARE ($\alpha=4$)	1.000	1.000	1.000	**0**	**0**	**1e-06**	**0**	**0**	**1e-13**
after MARE ($\alpha=5$) (=1821 data set)	1.000	1.000	1.000	**0**	**0**	**2e-56**	**0**	**0**	**6e-77**
251 data set	0.523	0.685	0.533	**0.008**	**0.018**	**0.003**	0.477	0.629	0.483
Amemiya et al. (2013)	0.549	0.710	0.569	**0.017**	**0.035**	**0.007**	0.451	0.604	0.464

Notes: Significant values are highlighted in bold. Note that KH might be most useful for *a priori* defined topologies.

phylogenetic information but it was insufficient to reject T3 (Supplementary Table S4 available on Dryad).

DISCUSSION

Lungfishes and not Coelacanths are the Closest Living Relatives of Tetrapods

Phylogenomic concatenation, coalescent analyses, and topology tests congruently favored lungfishes as the closest living relatives of tetrapods (T1). Both BI and ML methods favored the same topology in the larger 1821 data set with high statistical support (Fig. 3). The congruence among gene jackknife replicates further shows that these results are robust to gene sampling. The same association was also favored with high support by the smaller 251 data set analyzed with Bayesian mixture models and partitioned ML analyses after reducing among-lineage rate heterogeneity or discarding compositionally heterogeneous alignment regions (Figs. 4a and 5). Despite the presence of compositional heterogeneity in both matrices, FcLM did not support that either compositional heterogeneity or distribution of missing data compromised the recovery of T1 (Supplementary Figs. S25 and S26 available on Dryad). Multispecies coalescent methods consistently supported hypothesis T1 with high support when using the largest data set (Fig. 6b). Remarkably, the two alternative topologies (T2 and T3) were strongly rejected by topology tests using the larger 1821 data set (Table 1).

Overall, the obtained trees agree with the current understanding of the vertebrate phylogeny, recovering all major clades with high support: monophyly of actinopterygians, sarcopterygians, and tetrapods; amphibians as sister group of amniotes, which include the sister group of lepidosaurians and turtles + archosaurians (represented here by birds) and both as sister of mammals, where platypus is the sister group of therian (marsupial and placental) mammals (e.g., Cotton and Page 2002; Meyer and Zardoya 2003; Fong and Fujita 2011; Crottini et al. 2012; Amemiya et al. 2013; Chen et al. 2015). A notable exception is the root of placental mammals, which has remained controversial (e.g., Song et al. 2012; Morgan et al. 2013; Romiguier

et al. 2013; Chen et al. 2015), but taxon and gene sampling in the present study was not designed to address such a specific question (see e.g., Romiguier et al. 2013; Chen et al. 2015). The recovery of lungfishes as the closest living relatives of tetrapods agrees with most previous morphological, paleontological; and molecular systematic studies (Panchen and Smithson 1987; Hedges et al. 1993; Zardoya and Meyer 1996; Zardoya et al. 1998; Venkatesh et al. 2001; Meyer and Zardoya 2003; Brinkmann et al. 2004a, 2004b; Amemiya et al. 2013) and contributes to the correct polarity interpretation of paedomorphic neural characters in sarcopterygians (Northcutt 1986).

Effect of Assumed Evolutionary Model

Previous studies have demonstrated the importance of using correct evolutionary models in phylogenomic analysis; for the same data, different models can produce highly supported but contradicting topologies (e.g., Jeffroy et al. 2006; Kumar et al. 2012; Xi et al. 2012; Morgan et al. 2013). We present a clear case of this phenomenon: the BI tree reconstructed under site-homogeneous (LG) and site-heterogeneous models (CAT, CAT-GTR) contradict each other in the relative position of lungfish and coelacanth, and yet these nodes receive maximal support in both cases (Fig. 5a,c). In such a situation, *a priori* examination of the data and studying the relative performance of the different models becomes crucial. Here, we show that vertebrate sequences are unlikely to have evolved under the globally SRH conditions assumed by evolutionary models (see below). This is the case for most real sequence data, particularly among species that diverged millions of years ago (e.g., Lockhart et al. 1992; Phillips et al. 2004; Misof et al. 2014). Sequence evolution is highly complex and besides genuine historical signal, alignments also reflect a variety of nonphylogenetic signals that most evolutionary models currently available are unable to describe adequately (Jermiin et al. 2008). The complexity of sequence data is also the result of different sites evolving under different conditions depending on factors such as solvent accessibility, protein structure or function, among others (Le et al. 2012). We show that accounting for the heterogeneity in amino acid equilibrium frequencies and/ or substitution rates across sites in concatenated alignments is crucial. We illustrate this effect empirically by analyzing the 251 data set under different models and further demonstrate it by testing model fit in a Bayesian framework.

Model cross-validation clearly determined that amino acid profile mixture models fit both data sets significantly better than the site-homogeneous LG model and that using empirically estimated substitution rates (CAT-GTR) rather than a uniform rate distribution (CAT) increases model fit. Posterior predictive tests demonstrated that the better fit of CAT-GTR and CAT over LG is achieved by their ability to correctly model the observed site-specific amino acid diversity and homoplasy. In fact, this ability to better identify homoplasious changes comes from their capacity of accurately modeling amino acid diversity per position, which consequently makes mixture models less prone to LBA artifacts (Lartillot et al. 2007). In contrast, site-homogeneous models overestimate the expected amino acid diversity and underestimate homoplasy (Supplementary Table S3 available on Dryad), being thus more sensitive to LBA (Fig. 5c). Our phylogenetic analyses demonstrate that the LG4X model is unable to overcome the LBA artifact in the 251 data set (Fig. 5b), suggesting that more complex (two-level mixture) models such as CAT are required to adequately model among-site heterogeneities (and support T1) in this case. The superiority of CAT over LG4X comes from the use of reduced sets of amino acids in each profile (compared to the 20 states expected by LG4X), which allows a more efficient identification of homoplasy (Lartillot et al. 2007). This result contradicts Le et al. (2012) in their suggestion that LG4X has a performance comparable to two-level mixture models like CAT (Le and Gascuel 2008, 2010). Partitioned ML analyses of the 251 data set also recover T3 instead of T1 (with low support), suggesting that data partitioning also fails to adequately model among-site heterogeneities in this case. Notably the fit of data partitioning and LG4X could not be compared against CAT with cross-validation because these models are not implemented in PhyloBayes (Lartillot et al. 2013), but the failure to recover T1 suggests that both methods are inferior to two-level mixture models. Note that the ability of CAT to account for complex evolutionary processes, as compared to LG4X and data partitioning, can be attributed to the model itself and not to the use of BI, as demonstrated by posterior predictive tests and a BI analysis under a site-homogeneous model (LG).

When phylogenetic signal is strong, even poorly fitting models can recover the genuine phylogenetic tree, but the sophistication of the model becomes a key issue when the question at hand is difficult (Lartillot et al. 2007; Philippe and Roure 2011; Philippe et al. 2011). This is clearly the case for disentangling early sarcopterygian relationships: molecular synapomorphic changes could accumulate only during a short time period and many of them have been masked by saturation during millions of years. For this reason, the faint phylogenetic signal that remains can easily be obscured by heterogeneous evolutionary processes that most models do not account for (note that for the rest of the tree there are no problems of node resolution and support regardless of the phylogenetic method and data set).

Effect of Compositional Heterogeneity and Missing Data

Matched pairs tests of symmetry showed that both studied data sets are compositionally heterogeneous and that sequences are unlikely to have evolved under globally SRH conditions (for details, see Supplementary Information and Fig. S6 available

on Dryad). Nevertheless, results of FcLM on permutated data sets suggest that neither compositional heterogeneity nor the distribution of missing data, alone or in combination, compromised the resolution of early sarcopterygian relationships in any of the two analyzed data sets (see Supplementary Information and Fig. S26 available on Dryad). However, excluding compositionally heterogeneous regions from the 251 data set does improve phylogenetic reconstruction, as it allows the LBA artifact that compromises the correct recovery of early sarcopterygian relationships under ML to be overcome. In fact, we further show that these excluded heterogeneous regions favor the topology affected by LBA (T3) (Supplementary Fig. S11 available on Dryad). Taken together, these results suggest that compositional heterogeneity does not clearly support any of the tree possible topologies, but it affects the resolution of early sarcopterygian splits due to the weakness of the genuine phylogenetic signal. Our analyses demonstrate that removing highly heterogeneous alignment regions does improve phylogenetic inference (Misof et al. 2014; Doyle et al. 2015) and advocate for the use of analytical tools to understand and correct the negative effects of compositional heterogeneity.

Effect of Among-Lineage and Among-Site Rate Heterogeneity

Among-lineage rate heterogeneity is known to negatively impact phylogenetic reconstruction, the most extreme case being represented by LBA, which is a well-characterized phylogenetic artifact (e.g., Felsenstein 1978; Bergsten 2005; Wägele and Mayer 2007; Susko 2015) that becomes crucial when genome-scale data are analyzed (Philippe et al. 2011). We show that the topology T3 reconstructed by ML from the 251 matrix (Fig. 4b) suffers from LBA between fast-evolving actinopterygians and tetrapods. We empirically demonstrate this LBA artifact by recovering the T1 hypothesis after removing fast-evolving actinopterygians (*sensu* Rodríguez-Ezpeleta et al. 2007) from the 251 data set (Fig. 5d). We also show how LBA artifacts can be assessed visually by inspecting reticulation patterns in phylogenetic networks (Supplementary Fig. S7 available on Dryad). Besides the exclusion of fast-evolving species, several alternative approaches can be used to reduce among-lineage rate heterogeneity and overcome LBA artifacts. We have already shown the robustness of amino acid profile mixture models against LBA. The addition of the spotted gar (Braasch et al. 2016) to the 251 data set allows inference of the correct sarcopterygian branching order, demonstrating that the addition of earlier branching and slower-evolving species to fast-evolving lineages (i.e., actinopterygians) reduces the probability for LBA, in agreement with several previous studies (Graybeal 1998; Wägele and Mayer 2007; Townsend and López-Giráldez 2010; Prum et al. 2015). Note that the larger 1821 data set already included the spotted gar and we find no evidence of LBA in this case, even if stronger LBA effects might be expected for an alignment that is >7 times longer. We also observe that the 1821 data set has an overall lower evolutionary rate compared to the 251 data set (U Mann–Whitney test $p < 0.05$; Supplementary Figs. S27 and S28 available on Dryad), which together with the presence of the spotted gar and the turtle (Wang et al. 2013) might explain the absence of LBA in this data set. In fact, an ML analysis of the 1821 data set after the exclusion of the spotted gar (using best-fit models and partitions) recovered T1 with full support. Besides taxon manipulation, we demonstrate that the exclusion of genes with extensive among-lineage rate variation does reduce LBA. Removing the 10% of the genes with strongest among-lineage rate heterogeneity proved enough to overcome LBA, even though at least 30% need to be eliminated to obtain high support (SHS > 0.85; Supplementary Fig. S8 available on Dryad). In this case, the best result was obtained by excluding up to 70% of the genes with strongest among-lineage rate variation, allowing the recovery of T1 with BP = 95% despite using only 33,433 amino acids (33% of the original 251 data set).

In all our analyses, we have accounted for among-lineage rate heterogeneity using the discrete gamma distribution (Yang 1996), as it is routinely done in phylogenetic inference. However, modeling evolutionary rate variation among sites is sometimes insufficient and particular sites or genes with, for example, fast rates or compositional heterogeneity need to be excluded prior to phylogenetic inference (e.g., Doyle et al. 2015). Fast-evolving positions are likely to be saturated for divergent sequences and tend to produce complex character patterns that evolutionary models might fail to adequately describe. Several studies have shown that eliminating fast-evolving positions can avoid LBA (e.g., Brinkmann and Philippe 1999; Kostka et al. 2008; Irisarri et al. 2010; Cummins and McInerney 2011). However, the removal of fast-evolving positions in our case study did not overcome LBA: neither eliminating the sites in the fastest, two fastest or three fastest categories (with 74,104 amino acids remaining; 74% of the original data) allowed recovery of T1, and further elimination of sites produced an overall lack of resolution at the backbone of the vertebrate phylogeny. This result suggests that among-lineage rate heterogeneity has a more adverse effect than among-site rate heterogeneity in the reconstruction of early sarcopterygian splits. An alternative explanation might be that our strategy to exclude rate bins was too severe, but this is unlikely simply because after removing sites in the three fastest categories 74% of the data still remained, whereas the alignment with reduced among-lineage rate heterogeneity recovered T1 with only 33% of the data. Interestingly, the removal of compositionally heterogeneous regions allows overcoming the LBA artifact, even if compositional biases are generally expected from fast-evolving positions (Rodríguez-Ezpeleta et al. 2007).

Less is More: Increasing the Signal-to-Noise Ratio

Increasing the number of genes is generally expected to improve phylogenetic inference and contentious nodes were once expected to be unambiguously resolved with genome-scale data (e.g., Rokas et al. 2003). Nevertheless, it soon became apparent that simply adding more data does not necessarily solve a phylogenetic problem but instead can lead to a wrong answer due to systematic error (Jeffroy et al. 2006). Here, we used MARE to reduce a data set composed of 2960 genes into a smaller set of genes that are most informative. MARE produced an increase of ~10% of information content between the initial and most stringent final matrices (1821 data set). Phylogenetic signal in all tested matrices is enough to reject the alternative topology T2, but the alternative T3 could only be rejected when subsets of most informative genes were selected with MARE (Table 1). The use of MARE on the smaller 251 data set appreciably increased its phylogenetic signal as judged by lower p values than with the original matrix, even though it was insufficient to statistically reject T2 (Supplementary Table S4 available on Dryad). This result illustrates the difficulty of rejecting alternative resolutions of early sarcopterygian splits despite using about 3000 genes. Likewise, we show that removing compositionally heterogeneous regions or genes with a strong among-lineage rate variation from the 251 data set significantly improves both the topology and statistical support of reconstructed trees (Fig. 5) despite removing 20% and 67% of the original positions, respectively. These results demonstrate that the signal-to-noise ratio is more important than the size of the data set, a conclusion advocated by some previous phylogenomic studies (e.g., Salichos and Rokas 2013; Chen et al. 2015).

Multispecies Coalescent Methods Congruently Reconstruct Early Sarcopterygian Relationships despite Pervasive Gene Tree Heterogeneity

Both data sets displayed extensive gene tree heterogeneity, a pattern recognized by several previous investigations (e.g., Cranston et al. 2009; Pabijan et al. 2012; Song et al. 2012; Chen et al. 2015; Doyle et al. 2015). This heterogeneity could originate from biological processes such as incomplete lineage sorting or hybridization (Degnan and Rosenberg 2009), but can also result from stochastic error due to limited phylogenetic information in single gene alignments (Than et al. 2007; Doyle et al. 2015). The fact that the monophyly of well-established clades such as tetrapods was not recovered in an appreciable number of gene trees (44–68%) suggests that stochastic error might be a major confounding factor. Nevertheless, deep coalescence in the short internodes separating coelacanth, lungfishes, and tetrapods can be a further source of noise generating real incongruence among gene trees. Incomplete lineage sorting is typically expected for shallow divergences, but discordance between gene trees can also remain after long time periods because lineage sorting only depends on the length of the internode and the effective population size and not in the depth of that internode (Avise 2000; Edwards et al. 2005, 2016; Degnan and Rosenberg 2009). This effect might be exacerbated if population sizes in the ancestor of sarcopterygians were small, as might be the case currently for some coelacanth and lungfishes (Frentiu et al. 2001; Nikaido et al. 2013).

Multispecies coalescent analyses performed on the 1821 data set were congruent in pointing to lungfishes as the closest living relatives of tetrapods, regardless of whether all gene trees or just subsets of highly supported gene trees were used. The reconstructed topology was almost identical to that favored by concatenation methods. Most nodes in the species tree estimated from the 1821 data set received maximal multilocus bootstrap support and the T1 hypothesis was supported with 99% (Fig. 6b). The multispecies coalescent tree reconstructed from the 251 data set instead favored the T3 hypothesis, mirroring the concatenated ML phylogeny. The topology T3 is also more prevalent among gene trees in the 251 data set compared to the 1821 data set (Fig. 6a). These results mean that gene trees in the smaller 251 data set are more severely affected by LBA, hindering also the recovery of T1 by MP-EST. We demonstrate the good performance of MP-EST in reconstructing robust species trees for the 1821 data set despite widespread heterogeneity in gene tree topologies, even when strongly supported but contradicting gene trees are used (the subset of 412 highly supported gene trees), as shown by previous studies (e.g., Song et al. 2012; Liu et al. 2015a). However, the inability of MP-EST to overcome the LBA artifact in the 251 data set suggests that the robustness of species tree methods to taxon and gene sampling or inclusion of fast-evolving sites (Song et al. 2012; Xi et al. 2014; Liu et al. 2015b; see also Edwards et al. 2016; Springer and Gatesy 2016) is not universal. It should be noted that in all cases, MP-EST estimates extremely short internodes separating coelacanth, lungfishes, and tetrapods, which could reflect the combined effect of short speciation periods and small effective population sizes.

Lessons to Resolve Ancient and Short Internodes

Here, we have proposed an analytical pipeline to resolve early splits within sarcopterygians, which occurred rapidly and a long time ago. In the tree of life, other contentious internodes also await resolution, such as the relationships among lamprey, hagfish, and jawed vertebrates (Takezaki et al. 2003; Near 2009), the root of placental mammals (e.g., Song et al. 2012; Morgan et al. 2013; Romiguier et al. 2013; Chen et al. 2015), the neoavian radiation (e.g., Jarvis et al. 2014; Prum et al. 2015), and the relationships among lineages at the origin of land plants (e.g., Turmel et al. 2009; Laurin-Lemay et al. 2012; Zhong et al. 2014). The proposed pipeline uses available exploratory

tools to assess assumptions of models and phylogenetic inference methods and to increase the signal-to-noise ratio in phylogenomic data sets. We assess the presence of compositional heterogeneity, strength of phylogenetic signal and conflict and incomplete lineage sorting. Selecting more informative genes, removing compositionally heterogeneous alignment regions and reducing among-lineage rate heterogeneity is shown to improve subsequent phylogeny reconstruction. Large and more importantly informative data sets are always desirable because they increase the probability of generating accurate phylogenies, even using worse-fitting models. However, when we are limited by the size or low phylogenetic signal in the data, using models that are robust against compositional noise and LBA (i.e., amino acid profile mixture models) becomes crucial. Our pipeline includes a number of possible *a posteriori* checks to confirm the robustness of the obtained results, including topology tests, assessment of evolutionary model adequacy, testing LBA artifacts, and evaluating the effect of compositional heterogeneity and missing data distribution (Fig. 2). The selection of more informative genes and species and the *a posteriori* assessment of the robustness of results is a key issue that should be addressed in future phylogenomic studies, particularly when dealing with recalcitrant nodes such as those in early sarcopterygian evolution.

Supplementary Material

Data available from the Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.gd74v.

Funding

This work was supported by the Alexander von Humboldt Foundation; postdoctoral fellowships from the Alexander von Humboldt [application 1150725; to I.I.]; and European Molecular Biology Organization (EMBO) [ALTF 440-2013]. Further support came from grants of the Deutsche Forschungsgemeinschaft (DFG) [to A.M.] and the University of Konstanz [to I.I. and A.M.].

Acknowledgments

Marco Hasselmann (Zoo Berlin) provided access to the *Neoceratodus* tissue. Daniel Monné and Lénia Beck helped with wet lab protocols. We are grateful to Bernhard Misof, Karen Meusemann, and Arndt von Haeseler for assistance with permutation and FcLM analyses; and to Nicolas Lartillot for support and discussions about PhyloBayes. We are grateful to Frank E. Anderson, Thomas Near, Rafael Zardoya, Diego San Mauro, and two anonymous reviewers for insightful comments on the manuscript and to C. Darrin Hulsey for proofreading. The following people and institutions are acknowledged for access to HPC clusters: Paolo Franchini and Andreas Kautt (Meyer Lab server), Karsten Schäfer (HPC2 server at the University of Konstanz), and Jesús E. Marco and Luis J. Cabellos (Altamira supercomputer at the IFCA-CSIC). I.I. acknowledges the support from the "Angel Cabrera" award of the Dept. of Biodiversity and Evolutionary Biology from the Museo Nacional de Ciencias Naturales (CSIC), Madrid, Spain.

References

Abascal F., Zardoya R., Posada D. 2005. ProtTest: Selection of best-fit models of protein evolution. Bioinformatics 21:2104–2105.

Ahlberg P.E., Johanson Z. 1998. Osteolepiforms and the ancestry of tetrapods. Nature 395:792–794.

Akaike H. 1973. Information theory and an extension of the maximum likelihood principle. In: Petrov B.N., Csaki F., editors. Second international symposium of information theory. Budapest: Akademiai Kiado. p. 267–281.

Amemiya C.T., Alfoldi J., Lee A.P., Fan S., Philippe H., MacCallum I., Braasch I., Manousaki T., Schneider I., Rohner N., Organ C., Chalopin D., Smith J.J., Robinson M., Dorrington R.A., Gerdol M., Aken B., Biscotti M.A., Barucca M., Baurain D., Berlin A.M., Blatch G.L., Buonocore F., Burmester T., Campbell M.S., Canapa A., Cannon J.P., Christoffels A., De Moro G., Edkins A.L., Fan L., Fausto A.M., Feiner N., Forconi M., Gamieldien J., Gnerre S., Gnirke A., Goldstone J.V., Haerty W., Hahn M.E., Hesse U., Hoffmann S., Johnson J., Karchner S.I., Kuraku S., Lara M., Levin J.Z., Litman G.W., Mauceli E., Miyake T., Mueller M.G., Nelson D.R., Nitsche A., Olmo E., Ota T., Pallavicini A., Panji S., Picone B., Ponting C.P., Prohaska S.J., Przybylski D., Saha N.R., Ravi V., Ribeiro F.J., Sauka-Spengler T., Scapigliati G., Searle S.M.J., Sharpe T., Simakov O., Stadler P.F., Stegeman J.J., Sumiyama K., Tabbaa D., Tafer H., Turner-Maier J., van Heusden P., White S., Williams L., Yandell M., Brinkmann H., Volff J.-N., Tabin C.J., Shubin N., Schartl M., Jaffe D.B., Postlethwait J.H., Venkatesh B., Di Palma F., Lander E.S., Meyer A., Lindblad-Toh K. 2013. The African coelacanth genome provides insights into tetrapod evolution. Nature 496:311–316.

Avise J.C. 2000. Phylogeography. Cambridge (MA): Harvard University Press.

Baurain D., Brinkmann H., Philippe H. 2007. Lack of resolution in the animal phylogeny: Closely spaced cladogeneses or undetected systematic errors? Mol. Biol. Evol. 24:6–9.

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

Blair J.E., Hedges S.B. 2005. Molecular phylogeny and divergence times of deuterostome animals. Mol. Biol. Evol. 22:2275–2284.

Braasch I., Gehrke A.R., Smith J.J., Kawasaki K., Manousaki T., Pasquier J., Amores A., Desvignes T., Batzel P., Catchen J., Berlin A.M., Campbell M.S., Barrell D., Martin K.J., Mulley J.F., Ravi V., Lee A.P., Nakamura T., Chalopin D., Fan S., Wcisel D., Canestro C., Sydes J., Beaudry F.E.G., Sun Y., Hertel J., Beam M.J., Fasold M., Ishiyama M., Johnson J., Kehr S., Lara M., Letaw J.H., Litman G.W., Litman R.T., Mikami M., Ota T., Saha N.R., Williams L., Stadler P.F., Wang H., Taylor J.S., Fontenot Q., Ferrara A., Searle S.M.J., Aken B., Yandell M., Schneider I., Yoder J.A., Volff J.-N., Meyer A., Amemiya C.T., Venkatesh B., Holland P.W.H., Guiguen Y., Bobe J., Shubin N.H., Di Palma F., Alfoldi J., Lindblad-Toh K., Postlethwait J.H. 2016. The spotted gar genome illuminates vertebrate evolution and facilitates human-teleost comparisons. Nat. Genet. 48:427–437.

Bradnam K., Fass J., Alexandrov A., Baranay P., Bechner M., Birol I., Boisvert S., Chapman J., Chapuis G., Chikhi R., Chitsaz H., Chou W.-C., Corbeil J., Del Fabbro C., Docking T., Durbin R., Earl D., Emrich S., Fedotov P., Fonseca N., Ganapathy G., Gibbs R., Gnerre S., Godzaridis E., Goldstein S., Haimel M., Hall G., Haussler D., Hiatt J., Ho I. 2013. Assemblathon 2: Evaluating de novo methods of genome assembly in three vertebrate species. GigaScience 2:10.

Brinkmann H., Denk A., Zitzler J., Joss J.J., Meyer A. 2004a. Complete mitochondrial genome sequences of the south american and the

australian lungfish: testing of the phylogenetic performance of mitochondrial data sets for phylogenetic problems in tetrapod relationships. J. Mol. Evol. 59:834–848.
Brinkmann H., Philippe H. 1999. Archaea sister group of Bacteria? Indications from tree reconstruction artifacts in ancient phylogenies. Mol. Biol. Evol. 16:817–825.
Brinkmann H., Venkatesh B., Brenner B.R., Meyer A. 2004b. Nuclear protein-coding genes support lungfish and not the coelacanth as the closest living relatives of land vertebrates. Proc. Natl Acad. Sci. USA 101:4900–4905.
Bryant D., Galtier N., Poursat M.-A. 2005. Likelihood calculation in molecular phylogenetics. In: Gascuel O., editor. Mathematics of evolution and phylogeny. Oxford, New York: Oxford University Press. p. 33–58.
Capella-Gutiérrez S., Silla-Martínez J.M., Gabaldón T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25:1972–1973.
Carroll R.L. 1988. Vertebrate paleontology and evolution. New York: W.H. Freeman & Co.
Chang M.M. 1991. "Rhipidistians," dipnoans, and tetrapods. In: Schultze H.-P., Trueb L., editors. Origins of the higher groups of tetrapods: controversy and consensus. Ithaca (NY): Cornell University Press. p. 3–28.
Chen M.Y., Liang D., Zhang P. 2015. Selecting question-specific genes to reduce incongruence in phylogenomics: A case study of jawed vertebrate backbone phylogeny. Syst. Biol. 64:1104–1120.
Clack J.A. 2002. Gaining ground: the origin and early evolution of tetrapods. Bloomington (IN): Indiana University Press.
Conesa A., Götz S., García-Gómez J.M., Terol J., Talón M., Robles M. 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21:3674–3676.
Cotton J.A., Page R.D.M. 2002. Going nuclear: gene family evolution and vertebrate phylogeny reconciled. Proc. R. Soc. B 269: 1555–1561.
Cranston K.A., Hurwitz B., Ware D., Stein L., Wing R.A. 2009. Species trees from highly incongruent gene trees in rice. Syst. Biol. 58: 489–500.
Crottini A., Madsen O., Poux C., Strauß A., Vieites D.R., Vences M. 2012. Vertebrate time-tree elucidates the biogeographic pattern of a major biotic change around the K-T boundary in Madagascar. Proc. Natl Acad. Sci. USA 109:5358–5363.
Cummins C.A., McInerney J.O. 2011. A method for inferring the rate of evolution of homologous characters that can potentially improve phylogenetic inference, resolve deep divergence and correct systematic biases. Syst. Biol. 60:833–844.
Cunningham F., Amode M.R., Barrell D., Beal K., Billis K., Brent S., Carvalho-Silva D., Clapham P., Coates G., Fitzgerald S., Gil L., Girón C.G., Gordon L., Hourlier T., Hunt S.E., Janacek S.H., Johnson N., Juettemann T., Kähäri A.K., Keenan S., Martin F.J., Maurel T., McLaren W., Murphy D.N., Nag R., Overduin B., Parker A., Patricio M., Perry E., Pignatelli M., Riat H.S., Sheppard D., Taylor K., Thormann A., Vullo A., Wilder S.P., Zadissa A., Aken B.L., Birney E., Harrow J., Kinsella R., Muffato M., Ruffier M., Searle S.M.J., Spudich G., Trevanion S.J., Yates A., Zerbino D.R., Flicek P. 2015. Ensembl 2015. Nucleic Acids Res. 43:D662–D669.
Daeschler E.B., Shubin N.H., Jenkins F.A. 2006. A Devonian tetrapod-like fish and the evolution of the tetrapod body plan. Nature 440:757–763.
Darriba D., Taboada G.L., Doallo R., Posada D. 2011. ProtTest 3: Fast selection of best-fit models of protein evolution. Bioinformatics 27:1164–1165.
Degnan J.H., Rosenberg N.A. 2006. Discordance of species trees with their most likely gene trees. PLoS Genet. 2:e68.
Degnan J.H., Rosenberg N.A. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends Ecol. Evol. 24:332–340.
Dell'Ampio E., Meusemann K., Szucsich N.U., Peters R.S., Meyer B., Borner J., Petersen M., Aberer A.J., Stamatakis A., Walzl M.G., Minh B.Q., von Haeseler A., Ebersberger I., Pass G., Misof B. 2013. Decisive data sets in phylogenomics: Lessons from studies on the phylogenetic relationships of primarily wingless insects. Mol. Biol. Evol. 31:239–249.
Delsuc F., Tsagkogeorga G., Lartillot N., Philippe H. 2008. Additional molecular support for the new chordate phylogeny. Genesis 46: 592–604.
Dial K.P., Shubin N., Brainerd E.L. 2015. Great transformations in vertebrate evolution. Chicago (IL), London: The University of Chicago Press. p. 424.
Doyle V.P., Young R.E., Naylor G.J.P., Brown J.M. 2015. Can we identify genes with increased phylogenetic reliability? Syst. Biol. 64: 824–837.
Driskell A.C., Ané C., Burleigh J.G., McMahon M.M., O'Meara B.C., Sanderson M.J. 2004. Prospects for building the tree of life from large sequence databases. Science 306:1172–1174.
Dunn C.W., Hejnol A., Matus D.Q., Pang K., Browne W.E., Smith S.A., Seaver E., Rouse G.W., Obst M., Edgecombe G.D., Sorensen M.V., Haddock S.H.D., Schmidt-Rhaesa A., Okusu A., Kristensen R.M., Wheeler W.C., Martindale M.Q., Giribet G. 2008. Broad phylogenomic sampling improves resolution of the animal tree of life. Nature 452:745–749.
Dwivedi B., Gadagkar S.R. 2009. Phylogenetic inference under varying proportions of indel-induced alignment gaps. BMC Evol. Biol. 9:211.
Edwards S.V., Bryan Jennings W., Shedlock A.M. 2005. Phylogenetics of modern birds in the era of genomics. Proc. R. Soc. B 272: 979–992.
Edwards S.V., Xi Z., Janke A., Faircloth B.C., McCormack J.E., Glenn T.C., Zhong B., Wu S., Lemmon E.M., Lemmon A.R., Leache A.D., Liu L., Davis C.C. 2016. Implementing and testing the multispecies coalescent model: A valuable paradigm for phylogenomics. Mol. Phylogenet. Evol. 94:447–462.
Felsenstein J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. Syst. Zool. 27:401–410.
Felsenstein J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39:783–791.
Felsenstein J. 2004. Inferring phylogenies. Sunderland (MA): Sinauer Associates.
Fong J.J., Fujita M.K. 2011. Evaluating phylogenetic informativeness and data-type usage for new protein-coding genes across Vertebrata. Mol. Phylogenet. Evol. 61:300–307.
Forey P.L., Gardiner B.G., Patterson C. 1991. The lungfish, the coelacanth and the cow revisited. In: Schultze H.-P., Trueb L., editors. Origins of the higher groups of tetrapods: controversy and consensus. Ithaca (NY): Cornell University Press. p. 145–172.
Frentiu F., Ovenden J., Street R. 2001. Australian lungfish (*Neoceratodus forsteri*: Dipnoi) have low genetic variation at allozyme and mitochondrial DNA loci: A conservation alert? Conserv. Genet. 2:63–67.
Fritzsch B. 1987. Inner ear of the coelacanth fish *Latimeria* has tetrapod affinities. Nature 327:153–154.
Gatesy J., Baker R.H. 2005. Hidden likelihood support in genomic data: can forty-five wrongs make a right? Syst. Biol. 54:483–492.
Goldman N., Anderson J.P., Rodrigo A.G. 2000. Likelihood-based tests of topologies in phylogenetics. Syst. Biol. 49:652–670.
Gouy M., Guindon S., Gascuel O. 2010. SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol. Biol. Evol. 27:221–224.
Grabherr M.G., Haas B.J., Yassour M., Levin J.Z., Thompson D.A., Amit I., Adiconis X., Fan L., Raychowdhury R., Zeng Q., Chen Z., Mauceli E., Hacohen N., Gnirke A., Rhind N., di Palma F., Birren B.W., Nusbaum C., Lindblad-Toh K., Friedman N., Regev A. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. 29:644–652.
Graybeal A. 1998. Is it better to add taxa or characters to a difficult phylogenetic problem? Syst. Biol. 47:9–17.
Guindon S., Dufayard J.-F., Lefort V., Anisimova M., Hordijk W., Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. Syst. Biol. 59:307–321.
Haas B.J., Papanicolaou A., Yassour M., Grabherr M., Blood P.D., Bowden J., Couger M.B., Eccles D., Li B, Lieber M., MacManes M.D., Ott M., Orvis J., Pochet N., Strozzi F., Weeks N., Westerman R., William T., Dewey C.N., Henschel R., LeDuc R.D., Friedman N., Regev A. 2013. *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat. Protoc. 8:1494–1512.

Hartmann S., Vision T.J. 2008. Using ESTs for phylogenomics: Can one accurately infer a phylogenetic tree from a gappy alignment? BMC Evol. Biol. 8:345.
Hassanin A., Léger N., Deutsch J. 2005. Evidence for multiple reversals of asymmetric mutational constraints during the evolution of the mtochondrial genome of Metazoa, and consequences for phylogenetic inferences. Syst. Biol. 54:277–298.
Hedges S.B., Hass C.A., Maxson L.R. 1993. Relations of fish and tetrapods. Nature 363:501–502.
Heinicke M.P., Sanders J.M., Hedges S.B. 2009. Lungfishes (Dipnoi). In: Hedges S.B., Kumar S., editors. The timetree of life. New York: Oxford University Press. p. 348–350.
Hejnol A., Obst M., Stamatakis A., Ott M., Rouse G.W., Edgecombe G.D., Martinez P., Baguñá J., Bailly X., Jondelius U., Wiens M., Müller W.E.G., Seaver E., Wheeler W.C., Martindale M.Q., Giribet G., Dunn C.W. 2009. Assessing the root of bilaterian animals with scalable phylogenomic methods. Proc. R. Soc. B 276:4261–4270.
Huson D.H., Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. Mol. Biol. Evol. 23:254–267.
Irisarri I., San Mauro D., Green D.M., Zardoya R. 2010. The complete mitochondrial genome of the relict frog *Leiopelma archeyi*: Insights into the root of the frog tree of life. Mitochondr. DNA 21:173–182.
Jarvis E.D., Mirarab S., Aberer A.J., Li B., Houde P., Li C., Ho S.Y.W., Faircloth B.C., Nabholz B., Howard J.T., Suh A., Weber C.C., da Fonseca R.R., Li J., Zhang F., Li H., Zhou L., Narula N., Liu L., Ganapathy G., Boussau B., Bayzid M.S., Zavidovych V., Subramanian S., Gabaldón T., Capella-Gutiérrez S., Huerta-Cepas J., Rekepalli B., Munch K., Schierup M., Lindow B., Warren W.C., Ray D., Green R.E., Bruford M.W., Zhan X., Dixon A., Li S., Li N., Huang Y., Derryberry E.P., Bertelsen M.F., Sheldon F.H., Brumfield R.T., Mello C.V., Lovell P.V., Wirthlin M., Schneider M.P.C., Prosdocimi F., Samaniego J.A., Velazquez A.M.V., Alfaro-Núñez A., Campos P.F., Petersen B., Sicheritz-Ponten T., Pas A., Bailey T., Scofield P., Bunce M., Lambert D.M., Zhou Q., Perelman P., Driskell A.C., Shapiro B., Xiong Z., Zeng Y., Liu S., Li Z., Liu B., Wu K., Xiao J., Yinqi X., Zheng Q., Zhang Y., Yang H., Wang J., Smeds L., Rheindt F.E., Braun M., Fjeldsa J., Orlando L., Barker F.K., Jønsson K.A., Johnson W., Koepfli K.-P., O'Brien S., Haussler D., Ryder O.A., Rahbek C., Willerslev E., Graves G.R., Glenn T.C., McCormack J., Burt D., Ellegren H., Alström P., Edwards S.V., Stamatakis A., Mindell D.P., Cracraft J., Braun E.L., Warnow T., Jun W., Gilbert M.T.P., Zhang G. 2014. Whole-genome analyses resolve early branches in the tree of life of modern birds. Science 346:1320–1331.
Jayaswal V., Wong T.K., Robinson J., Poladian L., Jermiin L.S. 2014. Mixture models of nucleotide sequence evolution that account for heterogeneity in the substitution process across sites and across lineages. Syst. Biol. 63:726–742.
Jeffroy O., Brinkmann H., Delsuc F., Hervé P. 2006. Phylogenomics: the beginning of incongruence? Trends Genet. 22:225–231.
Jermiin L.S., Jayaswal V., Ababneh F., Robinson J. 2008. Phylogenetic model evaluation. In: Keith J.M., editor. Bioinformatics, volume I: Data, Sequence Analysis, and Evolution, vol. 452. Totowa (NJ): Springer. p. 331–364.
Katoh K., Standley D.M. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. 30:772–780.
Kishino H., Hasegawa M. 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. J. Mol. Evol. 29: 170–179.
Kostka M., Uzlikova M., Cepicka I., Flegr J. 2008. SlowFaster, a user-friendly program for slow-fast analysis and its application on phylogeny of Blastocystis. BMC Bioinform. 9:1–6.
Kriventseva E.V., Tegenfeldt F., Petty T.J., Waterhouse R.M., Simão F.A., Pozdnyakov I.A., Ioannidis P., Zdobnov E.M. 2015. OrthoDB v8: update of the hierarchical catalog of orthologs and the underlying free software. Nucleic Acids Res. 43:D250–D256.
Kubatko L.S., Degnan J.H. 2007. Inconsistency of phylogenetic estimates from concatenated data under coalescence. Syst. Biol. 56:17–24.
Kuck P., Longo G. 2014. FASconCAT-G: extensive functions for multiple sequence alignment preparations concerning phylogenetic studies. Front. Zool. 11:81.
Kumar S., Filipski A.J., Battistuzzi F.U., Kosakovsky Pond S.L., Tamura K. 2012. Statistics and truth in phylogenomics. Mol. Biol. Evol. 29:457–472.
Kumar V., Hallström B.M., Janke A. 2013. Coalescent-based genome analyses resolve the early branches of the Euarchontoglires. PLoS One 8:e60019.
Lanfear R., Calcott B., Kainer D., Mayer C., Stamatakis A. 2014. Selecting optimal partitioning schemes for phylogenomic datasets. BMC Evol. Biol. 14:82.
Lartillot N., Brinkmann H., Philippe H. 2007. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. BMC Evol. Biol. 7:S4–S4.
Lartillot N., Lepage T., Blanquart S. 2009. PhyloBayes 3: A Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics 25:2286–2288.
Lartillot N., Philippe H. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol. Biol. Evol. 21:1095–1109.
Lartillot N., Rodrigue N., Stubbs D., Richer J. 2013. PhyloBayes MPI: Phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. Syst. Biol. 62:611–615.
Laurin-Lemay S., Brinkmann H., Philippe H. 2012. Origin of land plants revisited in the light of sequence contamination and missing data. Curr. Biol. 22:R593–R594.
Le S.Q., Dang C.C., Gascuel O. 2012. Modeling protein evolution with several amino acid replacement matrices depending on site rates. Mol. Biol. Evol. 29:2921–2936.
Le S.Q., Gascuel O. 2008. An improved general amino acid replacement matrix. Mol. Biol. Evol. 25:1307–1320.
Le S.Q., Gascuel O. 2010. Accounting for solvent accessibility and secondary structure in protein phylogenetics is clearly beneficial. Syst. Biol. 59:277–287.
Le S.Q., Gascuel O., Lartillot N. 2008a. Empirical profile mixture models for phylogenetic reconstruction. Bioinformatics 24: 2317–2323.
Le S.Q., Lartillot N., Gascuel O. 2008b. Phylogenetic mixture models for proteins. Phil. Trans. R. Soc. B 363:3965–3976.
Lemmon A.R., Brown J.M., Stanger-Hall K., Lemmon E.M. 2009. The effect of ambiguous data on phylogenetic estimates obtained by maximum likelihood and Bayesian inference. Syst. Biol. 58:130–145.
Liang D., Shen X.X., Zhang P. 2013. One thousand two hundred ninety nuclear genes from a genome-wide survey support lungfishes as the sister group of tetrapods. Mol. Biol. Evol. 30:1803–1807.
Liu L., Wu S., Yu L. 2015a. Coalescent methods for estimating species trees from phylogenomic data. J. Syst. Evol. 53:380–390.
Liu L., Xi Z., Wu S., Davis C.C., Edwards S.V. 2015b. Estimating phylogenetic trees from genome-scale data. Ann. NY Acad. Sci. 1360:569–573.
Liu L., Yu L. 2010. Phybase: an R package for species tree analysis. Bioinformatics 26:962–963.
Liu L., Yu L, Edwards S. 2010. A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. BMC Evol. Biol. 10:302.
Liu Y., Cox C.J., Wang W., Goffinet B. 2014. Mitochondrial phylogenomics of early land plants: Mitigating the effects of saturation, compositional heterogeneity, and codon-usage bias. Syst. Biol. 63:862–878.
Lockhart P.J., Howe C.J., Bryant D.A., Beanland T.J., Larkum A.W.D. 1992. Substitutional bias confounds inference of cyanelle origins from sequence data. J. Mol. Evol. 34:153–162.
Meyer A., Wilson A.C. 1990. Origin of tetrapods inferred from their mitochondrial DNA affiliation to lungfish. J. Mol. Evol. 31:359–364.
Meyer A., Zardoya R. 2003. Recent advances in the (molecular) phylogeny of vertebrates. Ann. Rev. Ecol. Evol. Syst. 34:311–338.
Meyer B., Meusemann K., Misof B. 2011. MARE: MAtrix REduction - a tool to select optimized data subsets from supermatrices for phylogenetic inference. Version 0.1.2-rc. https://www.zfmk.de/en/research/research-centres-and-groups/mare.
Misof B., Liu S., Meusemann K., Peters R.S., Donath A., Mayer C., Frandsen P.B., Ware J., Flouri T., Beutel R.G., Niehuis O., Petersen M., Izquierdo-Carrasco F., Wappler T., Rust J., Aberer A.J., Aspöck U., Aspöck H., Bartel D., Blanke A., Berger S., Böhm A., Buckley T.R., Calcott B., Chen J., Friedrich F., Fukui M., Fujita M., Greve C.,

Grobe P., Gu S., Huang Y., Jermiin L.S., Kawahara A.Y., Krogmann L., Kubiak M., Lanfear R., Letsch H., Li Y., Li Z., Li J., Lu H., Machida R., Mashimo Y., Kapli P., McKenna D.D., Meng G., Nakagaki Y., Navarrete-Heredia J.L., Ott M., Ou Y., Pass G., Podsiadlowski L., Pohl H., von Reumont B.M., Schütte K., Sekiya K., Shimizu S., Slipinski A., Stamatakis A., Song W., Su X., Szucsich N.U., Tan M., Tan X., Tang M., Tang J., Timelthaler G., Tomizuka S., Trautwein M., Tong X., Uchifune T., Walzl M.G., Wiegmann B.M., Wilbrandt J., Wipfler B., Wong T.K.F., Wu Q., Wu G., Xie Y., Yang S., Yang Q., Yeates D.K., Yoshizawa K., Zhang Q., Zhang R., Zhang W., Zhang Y., Zhao J., Zhou C., Zhou L., Ziesmann T., Zou S., Li Y., Xu X., Zhang Y., Yang H., Wang J., Wang J., Kjer K.M., Zhou X. 2014. Phylogenomics resolves the timing and pattern of insect evolution. Science 346:763–767.

Mooers A.Ø., Holmes E.C. 2000. The evolution of base composition and phylogenetic inference. Trends Ecol. Evol. 15:365–369.

Morgan C.C., Foster P.G., Webb A.E., Pisani D., McInerney J.O., O'Connell M.J. 2013. Heterogeneous models place the root of the placental mammal phylogeny. Mol. Biol. Evol. 30:2145–2156.

Mossel E., Vigoda E. 2005. Phylogenetic MCMC algorithms are misleading on mixtures of trees. Science 309:2207–2209.

Müller J., Reisz R.R. 2005. Four well-constrained calibration points from the vertebrate fossil record for molecular clock estimates. Bioessays 27:1069–1075.

Near T.J. 2009. Conflict and resolution between phylogenies inferred from molecular and phenotypic data sets for hagfish, lampreys, and gnathostomes. J. Exp. Zool. B 312B:749–761.

Nelson J.S. 2006. Fishes of the world. 4th ed. Hoboken (NJ): John Wiley & Sons.

Nieselt-Struwe K., von Haeseler A. 2001. Quartet-mapping, a generalization of the likelihood-mapping procedure. Mol. Biol. Evol. 18:1204–1219.

Nikaido M., Noguchi H., Nishihara H., Toyoda A., Suzuki Y., Kajitani R., Suzuki H., Okuno M., Aibara M., Ngatunga B.P., Mzighani S.I., Kalombo H.W.J., Masengi K.W.A., Tuda J., Nogami S., Maeda R., Iwata M., Abe Y., Fujimura K., Okabe M., Amano T., Maeno A., Shiroishi T., Itoh T., Sugano S., Kohara Y., Fujiyama A., Okada N. 2013. Coelacanth genomes reveal signatures for evolutionary transition from water to land. Genome Res. 23:1740–1748.

Northcutt R.G. 1986. Lungfish neural characters and their bearing on sarcoptergian phylogeny. In: Beamis W.E., Burggren W.W., Kemp N.E., editors. The biology and evolution of lungfishes. New York: Alan R. Liss.

Pabijan M., Crottini A., Reckwell D., Irisarri I., Hauswaldt J.S., Vences M. 2012. A multigene species tree for Western Mediterranean painted frogs (*Discoglossus*). Mol. Phylogenet. Evol. 64:690–696.

Pagel M., Meade A. 2005. Mixture models in phylogenetic inference. In: Gascuel O., editor. Mathematics of evolution & phylogeny. New York: Oxford University Press. p. 121–142.

Panchen A.L., Smithson T.R. 1987. Character diagnosis, fossils and the origin of tetrapods. Biol. Rev. 62:341–436.

Philippe H., Brinkmann H., Lavrov D.V., Littlewood D.T.J., Manuel M., Wörheide G., Baurain D. 2011. Resolving difficult phylogenetic questions: why more sequences are not enough. PLoS Biol. 9:e1000602.

Philippe H., Roure B. 2011. Difficult phylogenetic questions: more data, maybe; better methods, certainly. BMC Biol. 9:91.

Phillips M.J., Delsuc F., Penny D. 2004. Genome-scale phylogeny and the detection of systematic biases. Mol. Biol. Evol. 21: 1455–1458.

Phillips M.J., Penny D. 2003. The root of the mammalian tree inferred from whole mitochondrial genomes. Mol. Phylogenet. Evol. 28: 171–185.

Pol D., Siddall M.E. 2001. Biases in maximum likelihood and parsimony: a simulation approach to a 10-taxon case. Cladistics 17:266–281.

Prum R.O., Berv J.S., Dornburg A., Field D.J., Townsend J.P., Lemmon E.M., Lemmon A.R. 2015. A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing. Nature 526:569–573.

R Development Core Team. 2009. R: a language and environment for statistical computing. Vienna (Austria): R Foundation for Statistical Computing.

Robinson-Rechavi M., Huchon D. 2000. RRTree: Relative-rate tests between groups of sequences on a phylogenetic tree. Bioinformatics 16:296–297.

Rodríguez-Ezpeleta N., Brinkmann H., Roure B., Lartillot N., Lang F.B., Philippe H. 2007. Detecting and overcoming systematic errors in genome-scale phylogenies. Syst. Biol. 56:389–399.

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

Romiguier J., Ranwez V., Delsuc F., Galtier N., Douzery E.J.P. 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.

Rota-Stabelli O., Telford M.J. 2008. A multi criterion approach for the selection of optimal outgroups in phylogeny: Recovering some support for Mandibulata over Myriochelata using mitogenomics. Mol. Phylogenet. Evol. 48:103–111.

Roure B., Baurain D., Hervé P. 2013. Impact of missing data on phylogenies inferred from empirical phylogenomic data sets. Mol. Biol. Evol. 30:197–214.

Sahney S., Benton M.J., Ferry P.A. 2010. Links between global taxonomic diversity, ecological diversity and the expansion of vertebrates on land. Biol. Lett. 6:544–547.

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

Schmieder R., Edwards R. 2011. Quality control and preprocessing of metagenomic datasets. Bioinformatics 27:863–864.

Schulz M.H., Zerbino D.R., Vingron M., Birney E. 2012. Oases: Robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics 28:1086–1092.

Seo T.-K. 2008. Calculating bootstrap probabilities of phylogeny using multilocus sequence data. Mol. Biol. Evol. 25:960–971.

Shan Y., Gras R. 2011. 43 genes support the lungfish-coelacanth grouping related to the closest living relative of tetrapods with the Bayesian method under the coalescence model. BMC Res. Notes 4:49.

Shimodaira H. 2002. An approximately unbiased test of phylogenetic tree selection. Syst. Biol. 51:592–508.

Shimodaira H., Hasegawa M. 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol. Biol. Evol. 16:1114–1116.

Shimodaira H., Hasegawa M. 2001. CONSEL: for assessing the confidence of phylogenetic tree selection. Bioinformatics 17: 1246–1247.

Song S., Liu L., Edwards S.V., 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 M.S., Gatesy J. 2016. The gene tree delusion. Mol. Phylogenet. Evol. 94(Part A):1–33.

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

St John J. 2013. SeqPrep. https://github.com/jstjohn/SeqPrep.

Streicher J.W., Schulte J.A., Wiens J.J. 2015. How should genes and taxa be sampled for phylogenomic analyses with missing data? An empirical study in iguanian lizards. Syst. Biol. 65: 128–145.

Strimmer K., von Haeseler A. 1997. Likelihood-mapping: A simple method to visualize phylogenetic content of a sequence?alignment. Proc. Natl Acad. Sci. USA 94:6815–6819.

Susko E. 2015. Bayesian long branch attraction bias and corrections. Syst. Biol. 64:243–255.

Takezaki N., Figueroa F., Zaleska-Rutczynska Z., Klein J. 2003. Molecular phylogeny of early vertebrates: Monophyly of the agnathans as revealed by sequences of 35 genes. Mol. Biol. Evol. 20:287–292.

Takezaki N., Figueroa F., Zaleska-Rutczynska Z., Takahata N., Klein J. 2004. The phylogenetic relationship of tetrapod, coelacanth, and lungfish revealed by the sequences of forty-four nuclear genes. Mol. Biol. Evol. 21:1512–1524.

Than C., Ruths D., Innan H., Nakhleh L. 2007. Confounding factors in HGT detection: statistical error, coalescent effects, and multiple solutions. J. Comput. Biol. 14:517–535.

Townsend J.P., López-Giráldez F. 2010. Optimal selection of gene and ingroup taxon sampling for resolving phylogenetic relationships. Syst. Biol. 59:446–457.
Townsend J.P., Su Z., Tekle Y.I. 2012. Phylogenetic signal and noise: predicting the power of a data set to resolve phylogeny. Syst. Biol. 61:835–849.
Townsend T.M., Mulcahy D.G., Noonan B.P., Sites J.W.J., Kuczynski C.A., Wiens J.J., Reeder T.W. 2011. Phylogeny of iguanian lizards inferred from 29 nuclear loci, and a comparison of concatenated and species-tree approaches for an ancient, rapid radiation. Mol. Phylogenet. Evol. 61:363–380.
Turmel M., Gagnon M.C., O'Kelly C.J., Otis C., Lemieux C. 2009. The chloroplast genomes of the green algae *Pyramimonas*, *Monomastix*, and *Pycnococcus* shed new light on the evolutionary history of prasinophytes and the origin of the secondary chloroplasts of euglenids. Mol. Biol. Evol. 26:631–648.
Venkatesh B., Erdmann M.V., Brenner S. 2001. Molecular synapomorphies resolve evolutionary relationships of extant jawed vertebrates. Proc. Natl Acad. Sci. USA 98:11382–11387.
Venkatesh B., Lee A.P., Ravi V., Maurya A.K., Lian M.M., Swann J.B., Ohta Y., Flajnik M.F., Sutoh Y., Kasahara M., Hoon S., Gangu V., Roy S.W., Irimia M., Korzh V., Kondrychyn I., Lim Z.W., Tay B.-H., Tohari S., Kong K.W., Ho S., Lorente-Galdos B., Quilez J., Marques-Bonet T., Raney B.J., Ingham P.W., Tay A., Hillier L.W., Minx P., Boehm T., Wilson R.K., Brenner S., Warren W.C. 2014. Elephant shark genome provides unique insights into gnathostome evolution. Nature 505:174–179.
Wägele J., Mayer C. 2007. Visualizing differences in phylogenetic information content of alignments and distinction of three classes of long-branch effects. BMC Evol. Biol. 7:147.
Wägele J.-W., Rödding F. 1998. A priori estimation of phylogenetic information conserved in aligned sequences. Mol. Phylogenet. Evol. 9:358–365.
Wang Z., Pascual-Anaya J., Zadissa A., Li W., Niimura Y., Huang Z., Li C., White S., Xiong Z., Fang D., Wang B., Ming Y., Chen Y., Zheng Y., Kuraku S., Pignatelli M., Herrero J., Beal K., Nozawa M., Li Q., Wang J., Zhang H., Yu L., Shigenobu S., Wang J., Liu J., Flicek P., Searle S., Wang J., Kuratani S., Yin Y., Aken B., Zhang G., Irie N. 2013. The draft genomes of soft-shell turtle and green sea turtle yield insights into the development and evolution of the turtle-specific body plan. Nat. Genet. 45:701–706.
Wiens J.J. 2003. Missing data, incomplete taxa, and phylogenetic accuracy. Syst. Biol. 52:528–538.
Wiens J.J, Morrill M.C. 2011. Missing data in phylogenetic analysis: reconciling results from simulations and empirical data. Syst. Biol. 60:719–731.
Xi Z., Liu L., Rest J.S., Davis C.C. 2014. Coalescent versus concatenation methods and the placement of Amborella as sister to water lilies. Syst. Biol. 63:919–932.
Xi Z., Ruhfel B.R., Schaefer H., Amorim A.M., Sugumaran M., Wurdack K.J., Endress P.K., Matthews M.L., Stevens P.F., Mathews S., Davis C.C. 2012. Phylogenomics and a posteriori data partitioning resolve the Cretaceous angiosperm radiation Malpighiales. Proc. Natl Acad. Sci. USA 109:17519–17524.
Yang Y., Smith S. 2013. Optimizing de novo assembly of short-read RNA-seq data for phylogenomics. BMC Genomics 14:328.
Yang Z. 1996. Among-site rate variation and its impact on phylogenetic analyses. Trends Ecol. Evol. 11:367–372.
Yang Z., Rannala B. 2012. Molecular phylogenetics: principles and practice. Nat. Rev. Genet. 13:303–314.
Zardoya R., Cao Y., Hasegawa M., Meyer A. 1998. Searching for the closest living relative(s) of tetrapods through evolutionary analyses of mitochondrial and nuclear data. Mol. Biol. Evol. 15:506–517.
Zardoya R., Meyer A. 1996. Evolutionary relationships of the coelacanth, lungfishes, and tetrapods based on the 28S ribosomal RNA gene. Proc. Natl Acad. Sci. USA 93:5449–5454.
Zhong B., Liu L., Yan Z., Penny D. 2013. Origin of land plants using the multispecies coalescent model. Trends Plant Sci. 18:492–495.
Zhong B., Xi Z., Goremykin V.V., Fong R., Mclenachan P.A., Novis P.M., Davis C.C., Penny D. 2014. Streptophyte algae and the origin of land plants revisited using heterogeneous models with three new algal chloroplast genomes. Mol. Biol. Evol. 31:177–183.
Zhu M., Schultze H.-P. 2001. Interrelationships of basal osteichthyans. In: Ahlberg P.E., editor. Major eventsin early vertebrate evolution: paleontology, phylogeny and development. London: Taylor & Francis. p. 289–314.