One Thousand Two Hundred Ninety Nuclear Genes from a Genome-Wide Survey Support Lungfishes as the Sister Group of Tetrapods

Dan Liang,[1,†] Xing Xing Shen,[1,†] and Peng Zhang[1,*]

[1]Key Laboratory of Gene Engineering of the Ministry of Education, State Key Laboratory of Biocontrol, School of Life Sciences, Sun Yat-Sen University, Guangzhou, China

[†]These authors contributed equally to this work.

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

Associate editor: Nicolas Vidal

Abstract

The only currently unresolved portion of the backbone phylogeny of the vertebrates involves the relationships among coelacanths, lungfishes, and tetrapods. Despite active research on this question over the past three decades, it is still difficult to determine statistically whether lungfishes alone or both lungfishes and coelacanths together are closely related to tetrapods. To resolve this controversy, we assembled a data set comprising 1,290 nuclear genes encoding 690,838 amino acid residues by analyzing available genome and transcriptome data. Phylogenetic analyses of this data set provided overwhelming evidence that the lungfishes are the closest living relatives of the land vertebrates. This result is strongly supported by high bootstrap values from maximum likelihood and maximum parsimony analyses, Bayesian posterior probabilities of CAT model analysis, and topological tests. Additionally, a species tree analysis without data concatenation also strongly supported this result.

Key words: **transcriptome, phylogenomics, Sarcopterygii, coelacanth, species tree.**

The "living fossils," coelacanths and lungfishes, are the only two extant lobe-finned fish groups. Determining which group is more closely related to tetrapods is of importance if we are to understand how land vertebrates originated and colonized the land. However, despite extensive molecular and morphological research on this question during the last 3 decades, the phylogenetic relationships among coelacanths, lungfishes, and tetrapods remain unresolved and are still debated (Hedges 2009). The majority of molecular, morphological, and paleontological studies have favored the hypothesis that the lungfishes are the closest living relatives of the tetrapods (Tree 1; fig. 1*a*) (Panchen and Smithson 1991; Hedges et al. 1993; Yokobori et al. 1994; Zardoya et al. 1998; Venkatesh et al. 2001; Zhu et al. 2001; Brinkmann et al. 2004). This is currently the prevailing view in many general biology textbooks and educational websites and has been used as the framework for recent comparative evolutionary studies (Christensen-Dalsgaard et al. 2011; King et al. 2011). Nevertheless, there is also evidence supporting the view that the coelacanths are most closely related to the tetrapods (Tree 2; fig. 1*b*) (Fritzsch 1987; Gorr et al. 1991) or that the lungfishes and coelacanths form a clade and are equally related to tetrapods (Tree 3; fig. 1*c*) (Zardoya and Meyer 1996; Shan and Gras 2011). However, in most studies that support Tree 1, alternative hypotheses (especially Tree 3) cannot be excluded with statistical support. By far, the most powerful analysis using 44 nuclear genes still generated conflicting results, indicating that the coelacanth, lungfish, and tetrapod lineages diverged within a very short time interval and that their relationships may represent an irresolvable trichotomy (Tree 4; fig. 1*d*) (Takezaki et al. 2004).

Recently, the genome sequence data of the coelacanth (*Latimeria chalumnae*) were released, providing an opportunity to revisit the phylogenetic relationships among coelacanths, lungfishes, and tetrapods at a genome level. Unfortunately, there is no genome sequencing project for lungfishes because these animals have extremely large genomes (up to 40 times as large as the human genome) that are intractable for current DNA sequencing technology. However, modern RNA-Seq transcriptome technology may solve this problem because it can effortlessly generate genome-wide protein-coding sequences for phylogenetic studies. Taking advantage of the available RNA-Seq data for the African lungfish (*Protopterus annectens*), we assembled a phylogenomic data set of 1,290 nuclear genes (690,838 aligned amino acids, 8.0% missing data), including one lungfish, one coelacanth, and three tetrapods, with two ray-finned fishes and three cartilaginous fishes as the outgroup.

Phylogenetic analyses of the 1,290-gene data set using maximum parsimony (MP), maximum likelihood (ML), and CAT-model Bayesian inference all converged to the same topology (fig. 2). All noncontroversial groups (chondrichthyes, actinopterygians, sarcopterygians, tetrapods, and amniotes) were recovered with unambiguous bootstrap support (bootstrap percentage $BP_{MP-ML} = 100\%$; fig. 2) and Bayesian posterior probabilities ($PP_{CAT} = 1.0$; fig. 2). For the node of interest, MP strongly supported the grouping of lungfishes and tetrapods—Tree 1 (BP = 98%). The more accurate ML method

Letter

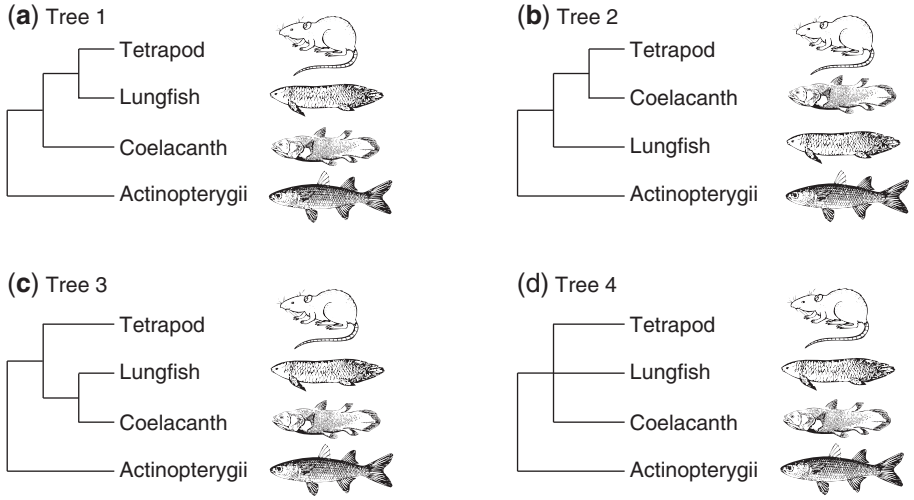

FIG. 1. The four possible phylogenetic relationships among coelacanths, lungfishes, and tetrapods.

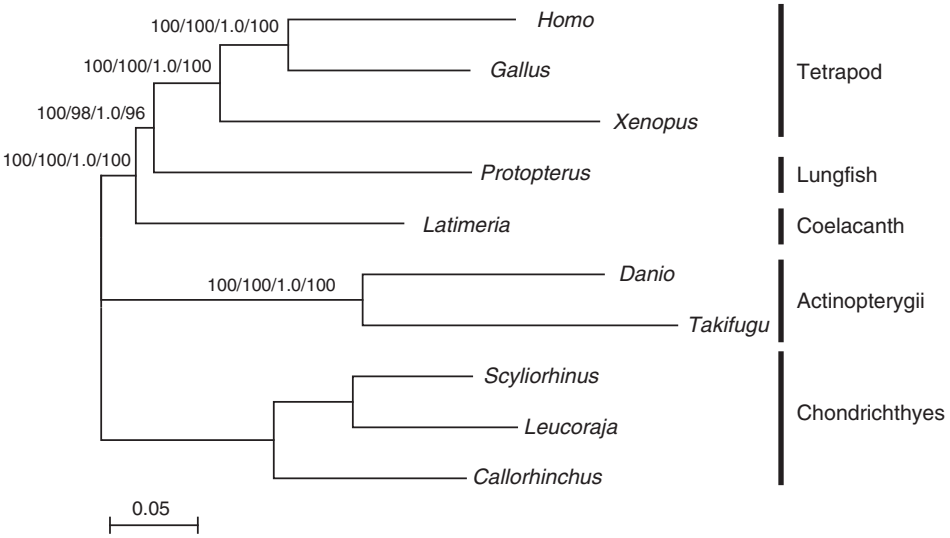

FIG. 2. The backbone phylogenetic tree of jawed vertebrates inferred from the phylogenomic data sets comprising 1,290 nuclear genes and 690,838 amino acid positions. The tree was inferred from concatenation analyses using MP, ML, and a Bayesian mixture model (CAT) and from a species-tree analysis using the pseudo-ML approach (MP-EST). Branch support values are indicated beside nodes in order of ML bootstrap, MP bootstrap, Bayesian CAT posterior probability, and MP-EST bootstrap, from left to right. Branch lengths are from the ML analysis.

provided unambiguous bootstrap support (100%) for this grouping (fig. 2). Our analysis of data subsets shows a progressive increase in bootstrap support value for the lungfish–tetrapod grouping if more and more genes are analyzed. To recover this relationship with $BP_{ML} > 90\%$, at least 100 genes are needed (fig. 3). To further test the stability of phylogenetic relationships among lungfishes, coelacanths, and tetrapods, we used a likelihood framework to evaluate the three possible topologies connecting the three groups. As a result, Tree 2 and Tree 3 were significantly rejected at the 5% confidence level by all statistical tests (table 1). When we introduced the lamprey as the outgroup, the data set was reduced to 811 genes, 418,696 sites, and 8.4% missing data. The reduced data set also supports the lungfish–tetrapod grouping ($BP_{MP} = 68\%$; $BP_{ML} = 100\%$; $PP_{CAT} = 1.0$; supplementary fig. S1, Supplementary Material online) and significantly rejects the alternative hypotheses (supplementary table S1, Supplementary Material online). However, the 811-gene data set does not recover a well-defined actinopterygian–sarcopterygian clade in the MP analysis, and the lamprey has a particularly long branch (supplementary fig. S1, Supplementary Material online). It is known that the inclusion of a distantly related and rapidly evolving outgroup can decrease the probability of recovering the correct tree topology. Therefore, this data set was not used for further analyses.

Note that the above analyses were all based on concatenation methods, which do not accommodate gene tree heterogeneity. Theoretically, concatenation methods may yield misleading results if a high level of gene tree heterogeneity occurs in phylogenomic data (Mossel and Vigoda 2005; Kubatko and Degnan 2007). We inferred phylogenetic trees for each gene using ML (RAxML) and found that the gene tree heterogeneity is evident (supplementary fig. S2, Supplementary Material online). Such a problem can be

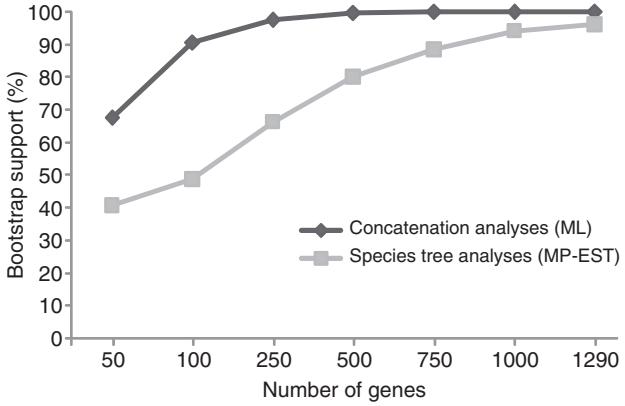

FIG. 3. The effect of increasing the number of nuclear loci on recovering the lungfish–tetrapod grouping. Each data point represents the mean of support values estimated from 10 random sampling subsets.

Table 1. Topological Test of the Three Phylogenetic Hypotheses among Tetrapods, Lungfishes, and Coelacanths Based on the 1,290-Gene Data Set.

Hypothesis	−ln *L*	Δln *L*	AU	SH	RELL BP
Tree 1	6397799.6	best	1.000	1.000	1.000
Tree 2	6398385.2	585.6	2E-34[a]	0[a]	0[a]
Tree 3	6398305.1	505.4	3E-102[a]	0[a]	0[a]

NOTE.—L, likelihood value; AU, approximately unbiased test; SH, Shimodaira–Hasegawa test; RELL BP, resampling of estimated log-likelihood bootstrap percentage.
[a]Statistically significant at the 5% level.

resolved through the use of phylogenomic data and coalescent methods that explicitly address gene tree heterogeneity (Song et al. 2012). Therefore, we used a recently developed coalescent method, the maximum pseudolikelihood estimation of the species tree (MP-EST) method (Liu et al. 2010), to reanalyze the 1,290-gene data. As a result, the tree obtained by the coalescent analysis was identical to those from concatenation analyses, favoring Tree 1, namely, that the lungfishes are the closest living relatives of the tetrapods (Bootstrap percentage = 96% by MP-EST; fig. 2).

Our results, therefore, indicate a strong phylogenetic affinity between the lungfishes and the tetrapods to the exclusion of the coelacanths. However, obtaining high statistical support for a given topology does not necessarily indicate that the phylogenetic inference is correct (Delsuc et al. 2006). It is common in phylogenomic analyses that misleading results are strengthened if systematic errors occur, for instance, long-branch attraction (LBA) artifacts and compositional biases. A saturation plot (supplementary fig. S3, Supplementary Material online) indicates that there are few saturated characters in our data set. Moreover, lungfishes evolve slightly faster than coelacanths (fig. 2); if LBA occurs, the rapidly evolving lungfishes would be attracted to the rapidly evolving outgroup and would not be retained in their current position. In addition, using different combinations of tetrapods does not affect the tree topology and branch support (supplementary table S2, Supplementary Material online).Therefore, the lungfish branch was not attracted to the *Xenopus* branch either, which is longer than the human and chicken branches. These three observations demonstrate that LBA is not responsible for the inferred grouping of lungfishes and tetrapods. To assess possible compositional biases, we calculated amino acid composition for the 10 studied taxa and found relatively stationary usage frequencies for 20 amino acid residues (supplementary table S3, Supplementary Material online). In conclusion, the strongly supported grouping between lungfishes and tetrapods cannot be explained by identifiable systematic biases (LBA and compositional bias). Therefore, it most likely represents the evolutionary history.

Our results highlight the power of mining genome and transcriptome data to construct large-scale data sets to resolve species relationships in difficult biological scenarios. The proposed vertebrate backbone phylogeny strengthens the view that the lungfishes, not the coelacanths, are the closest living relatives of the land vertebrates. This result improves the current understanding of the origin of the major evolutionary novelties for living on land, such as alveolated lungs and paired pectoral and pelvic appendages. Comparative studies of tetrapods with our closest "fish" relatives from both biological and genomics perspectives will be valuable for understanding the early evolution of land vertebrates.

Materials and Methods

Data Assembly

Proteome data of human (*Homo sapiens*), chicken (*Gallus gallus*), frog (*Xenopus tropicalis*), zebrafish (*Danio rerio*), fugu (*Takifugu rubripes*), coelacanth (*L. chalumnae*), and lamprey (*Petromyzon marinus*) were downloaded from the Ensembl (http://www.ensembl.org/). RNA-seq data of the African lungfish (*P. annectens*) and three cartilaginous fishes (*Callorhinchus milii*, *Leucoraja erinacea*, and *Scyliorhinus canicula*) were retrieved from the Sequence Read Archive of NCBI under accession SRR505721–SRR505726, SRR088619–SRR088625. Transcriptome assembly and putative ORF prediction were performed with the program suite Trinity (Grabherr et al. 2011). We used the mutual best hit (MBH) in Basic Local Alignment Search Tool to identify genes that are putatively orthologous between two species. Only hits with an *e* value lower than 10^{-20} were retained. To avoid the use of confounding paralogs, the MBH lists were filtered, such that an MBH-pair was retained only if the second best hit of either gene in the other genome had a score smaller than half the score of the best hit. Only gene sets including all species of interest were retained for further analysis. Each gene set was aligned using MUSCLE (Edgar 2004) and trimmed using GBlocks (Castresana 2000) without a gap penalty (-b5 = a). Refined alignments shorter than 200 aa were discarded. Because the proteome of the lamprey is rather incomplete (~10,000 genes), we prepared two data sets, one including the lamprey and comprising 811 genes and another excluding the lamprey and comprising 1,290 genes. All selected genes of the two data sets are described in supplementary table S4, Supplementary Material online.

Phylogenetic Analyses

Unweighted MP heuristic searches were conducted using MEGA5 (Tamura et al. 2011) with 20 random additions of species and the close neighbor interchange method. The ML analyses were implemented using RAxML version 7.2.6 (Stamatakis 2006) with a concatenated LG + F + Γ_4 model. Bootstrap support for MP and ML was evaluated with 500 replicates. Bayesian inference under a mixture model CAT + Γ4 was performed in PhyloBayes 3.3 (Lartillot et al. 2009) with two independent MCMC runs for 10,000 cycles. Stationarity was reached when the largest discrepancy (maxdiff) was less than 0.1 between two independent runs. Species tree estimation was conducted using the pseudo-ML approach in the program MP-EST (Liu et al. 2010) under the coalescent model. The robustness of the species tree was evaluated with nonparametric bootstrapping of 500 replicates. To explore the effect of the number of genes on tree reconstruction, we randomly sampled 50, 100, 250, 500, 750, and 1,000 genes from the original 1,290-gene data set 10 times and redid the ML and MP-EST analyses. The branch support of the lungfish–tetrapod clade from these subsets is summarized in figure 3.

Likelihood-based tests of alternative topologies were calculated using CONSEL (Shimodaira and Hasegawa 2001). Sitewise log-likelihood values were computed with RAxML with a concatenated LG + F + Γ_4 model. The *P* values of the different likelihood-based tests were finally calculated with CONSEL.

Supplementary Material

Supplementary figures S1–S3 and tables S1–S4 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 31172075 and 30900136) to P.Z. and the Cultivation Project of the Department of Education of Guangdong Province (2012330004201913) to D.L.

Note Added in Proof

We used the coelacanth genome data and the lungfish RNA-seq data from Amemiya et al. (2013), which was published the same week in which our manuscript was accepted and released online in Advanced Access in *Molecular Biology and Evolution*. Amemiya et al. obtained the same relationship among coelacanths, lungfishes, and tetrapods as we have reported in this work. We commend Amemiya et al. for freely providing new genome data for use by everyone soon after its completion, which enabled us to conduct analysis concurrently, but independently, of their team. We suggest that researchers refer to the Amemiya et al. publication as the first source when discussing the relationship reported among coelacanths, lungfishes, and tetrapods.

References

Amemiya CT, Alföldi J, Lee AP, et al. (91 co-authors). 2013. The African coelacanth genome provides insights into tetrapod evolution. *Nature* 496:311–316.

Brinkmann H, Venkatesh B, Brenner S, Meyer A. 2004. Nuclear protein-coding genes support lungfish and not the coelacanth as the closest living relatives of land vertebrates. *Proc Natl Acad Sci U S A.* 101: 4900–4905.

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

Christensen-Dalsgaard J, Brandt C, Wilson M, Wahlberg M, Teglberg PM. 2011. Hearing in the African lungfish (*Protopterus annectens*): pre-adaptation for pressure hearing in tetrapods? *Biol Lett.* 7:139–141.

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

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32:1792–1797.

Fritzsch B. 1987. The inner ear of the coelacanth fish Latimeria has tetrapod affinities. *Nature* 327:153–154.

Gorr T, Kleinschmidt T, Fricke H. 1991. Close tetrapod relationships of the coelacanth Latimeria indicated by haemoglobin sequences. *Nature* 351:394–397.

Grabherr MG, Haas BJ, Yassour M, et al. (21 co-authors). 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol.* 29:644–652.

Hedges SB. 2009. Vertebrates (Vertebrata). In: Hedges SB, Kumar S, editors. The timetree of life. New York: Oxford University Press. p. 309–314.

Hedges SB, Hass CA, Maxson LR. 1993. Relations of fish and tetrapods. *Nature* 363:501–502.

King HM, Shubin NH, Coates MI, Hale ME. 2011. Behavioral evidence for the evolution of walking and bounding before terrestriality in sarcopterygian fishes. *Proc Natl Acad Sci U S A.* 108: 21146–21151.

Kubatko LS, Degnan JH. 2007. Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Syst Biol.* 56:17–24.

Lartillot N, Lepage T, Blanquart S. 2009. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics* 25:2286–2288.

Liu L, Yu L, Edwards SV. 2010. A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. *BMC Evol Biol.* 10:302.

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

Panchen AL, Smithson TS. 1991. Character diagnosis, fossils and the origin of tetrapods. *Biol Rev.* 62:341–438.

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

Song S, Liu L, Edwards SV, Wu S. 2012. Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. *Proc Natl Acad Sci U S A.* 109:14942–14947.

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

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.

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol.* 28:2731–2739.

Venkatesh B, Erdmann MV, Brenner S. 2001. Molecular synapomorphies resolve evolutionary relationships of extant jawed vertebrates. *Proc Natl Acad Sci U S A.* 98:11382–11387.

Yokobori AI, Hasegawa M, Ueda T, Okada N, Nishikawa K, Watanabe K. 1994. Relationship among coelacanths, lungfishes, and tetrapods: a phylogenetic analysis based on mitochondrial cytochrome oxidase I gene sequences. *J Mol Evol.* 38: 602–609.

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, lungfish, and tetrapods based on the 28S ribosomal RNA gene. *Proc Natl Acad Sci U S A.* 93:5449–5454.

Zhu M, Yu XB, Ahlberg PE. 2001. A primitive sarcopterygian fish with an eyestalk. *Nature* 410:81–84.