# Three Genome-scale Approaches Support that Lungfish is the Closest Living Relative of Land Vertebrate, but not Coelacanth

Yunfeng Shan[1]

[1]Guangxi Science and Technology Normal University

July 12, 2022

## Abstract

The origin of tetrapod has been one of intense debating open questions for decades between coelacanth(Latimeria chalumnae) and lungfish (Protopterus annectens). For resolving this incongruence in phylogenies, a genome-wide data mining approach is used to retrieve 43 shared genes of seven taxa from GenBank and further 1001 orthologous genes of ten taxa from the Ensembl and NCBI. We used the maximum gene-support tree approach and the majority-rule branch approach to analyze 43 nuclear genes encoding amino acid residues and compared these results to those inferred with the concatenation approach. Our results successfully provide strong evidence in favor of the lungfish-tetrapod hypothesis, but rejecting the coelacanth-tetrapod hypothesis based on significantly fewer gene supports and lower taxon jackknife probabilities for the coelacanth-tetrapod clade than the lungfish-tetrapod one with the maximum gene-support tree approach and the jackknife method for taxon subsampling. When more and more genomic data become available in recent years, sequence data of 1001 shared genes was mined. We used the maximum gene-support approach with this larger dataset successfully to infer that lungfish is the closest relative of land vertebrates with a significant difference at p $<$ 0.01 (Chi-Square test) in gene support values between a maximum gene-support tree and the second most gene support tree with ML methods. The second most support to the maximum (SM ratio), a relative value, is a better support index than a single absolute value of support to show the insight of the phylogenetic support. Our results also show increasing the number of shared genes is much more effective than increasing the number of taxa.

Three Genome-scale Approaches Support that Lungfish is the Closest Living Relative of Land Vertebrate, but not Coelacanth

Yunfeng Shan[1.5*], Youjun Zhou[1], Guo Li[2], Xiu-Qing Li[3], and Robin Gras[4]

[1] College of Mathematics and Computer Sciences, Guangxi Science and Technology Normal University, Laibin 546199, China;

[2] College of Electronic Information and Automation, Guilin University of Aerospace Technology, Guilin 541004, China;

[3] Fredericton Research and Development Centre, Agriculture and Agri-Food Canada, Fredericton, New Brunswick, E3B 4Z7, Canada;

[4]School of Computer Science, University of Windsor, 401 Sunset Avenue, Windsor, Ontario N9B 3P4, Canada;

[5]Department of Natural History, Royal Ontario Museum, 100 Queen's Park, Toronto, Ontario M5S 2C6, Canada

Corresponding authors: Yunfeng Shan (shanyunfeng2018@163.com)

This work was supported in part by the Grant for Outstanding Scholars from Guangxi Science and Technology Normal University. This work was also supported in part by the National Natural Science Foundation of China under Grant 61861014; in part by the Guangxi Nature Science Fund under Grant 2015GXNSFAA139298 and Grant 2016GXNSFAA380226; in part by the Guangxi University High Level Innovation Team and Outstanding Scholar Program; in part by the Guangxi Thousand Bones Project.

**ABSTRACT** The origin of tetrapod has been one of intense debating open questions for decades between coelacanth(*Latimeria chalumnae* ) and lungfish (*Protopterus annectens* ). For resolving this incongruence in phylogenies, a genome-wide data mining approach is used to retrieve 43 shared genes of seven taxa from GenBank and further 1001 orthologous genes of ten taxa from the Ensembl and NCBI. We used the maximum gene-support tree approach and the majority-rule branch approach to analyze 43 nuclear genes encoding amino acid residues and compared these results to those inferred with the concatenation approach. Our results successfully provide strong evidence in favor of the lungfish-tetrapod hypothesis, but rejecting the coelacanth-tetrapod hypothesis based on significantly fewer gene supports and lower taxon jackknife probabilities for the coelacanth-tetrapod clade than the lungfish-tetrapod one with the maximum gene-support tree approach and the jackknife method for taxon subsampling. When more and more genomic data become available in recent years, sequence data of 1001 shared genes was mined. We used the maximum gene-support approach with this larger dataset successfully to infer that lungfish is the closest relative of land vertebrates with a significant difference at $p < 0.01$ (Chi-Square test) in gene support values between a maximum gene-support tree and the second most gene support tree with ML methods. The second most support to the maximum (SM ratio), a relative value, is a better support index than a single absolute value of support to show the insight of the phylogenetic support. Our results also show increasing the number of shared genes is much more effective than increasing the number of taxa.

**INDEX TERMS** Origin of tetrapod, lungfish, coelacanth, genome-scale approaches, data mining.

## I. INTRODUCTION

More and more genomic data become available publicly due to recently many genome projects such as the 10k animal genome project and the 10k fish genome project. It provides a good opportunity and resources to resolve some open-standing questions by analyzing data through data mining from big genomic data resources.

The origin of tetrapods (land vertebrates) has been debated for many decades. Since its discovery in 1938, the coelacanth (*Latimeria chalumnae* ), the "living fossil" [1,2], has generally been thought the closest living relative of the land vertebrates [3], the missing transition from aquatic to terrestrial vertebrates. Three hypotheses have been proposed for the phylogenetic relationship: lungfish-tetrapod (Hypothesis 1, Fig. 1a); coelacanth-tetrapod (Hypothesis 2, Fig.1b); and lungfish-coelacanth sister grouping (Hypothesis 3, Fig. 1c). The lungfish-coelacanth-tetrapod trichotomy (Fig. 1d) is not generally considered a hypothesis. The coelacanth-tetrapod sister hypothesis (Fig. 1b) was proposed by some comparative morphologists and paleontologists [1, 4-7] since the coelacanth's discovery, although the lungfishes were historically thought to be the closest living relative, which was also supported by recent researchers [8-12]. The hypothesis that coelacanth and lungfish form a monophyletic sister group that is closely related to tetrapod (tree III) was also favored [13-15].

Molecular data from single genes, whole mitochondrial genomes, especially recent whole transcriptomes, and whole nuclear genomes have widely been used over the last three decades for inferring phylogenetic relationships. Lungfish-tetrapod sister hypothesis was favored by molecular data [16-27], while the hypothesis of the coelacanth as the closest living relative of tetrapod was preferred [3, 28-31]. The coelacanth and lungfish sister hypothesis was suggested by a single gene or multiple genes [20-21, 32] and by the whole mitochondrial genome [24], whereas an unresolved coelacanth-lungfish-tetrapod trichotomy has resulted from the 12S rRNA gene [18] or 44 genes [33]. With sequence data of whole transcriptomes and genomes, three analyses (genes with Phylobayes and ML, 251 genes with Phylobayes) from four analyses showed that lungfish was the closest living relative to tetrapod [34]. Similar results were reported [35-38], but one analysis (251 genes with ML) showed that the latest common ancestor of coelacanth and lungfish was the closest living relative to the tetrapod [34]. Another phylogenomic analysis with three datasets showed that in all the data sets with ML and Bayesian methods, the sister relationship of lungfish and tetrapod was reconstructed with the use of cartilaginous fish as the outgroup with high support, but when ray-finned fish were used as the outgroup, the sister relationship of coelacanth and tetrapod was supported most strongly [39]. Therefore, no consistent result is shown up to now. The origin of the tetrapod continued to be contentious and still is one of the longest-standing open questions in land vertebrate evolution.

Significant progress in genome sequencing technology and recently completed genome projects produce huge amounts of sequence data from a lot of organisms, which results in an inference that in the near future, recovering the tree of life (TOL) will simply be a matter of enough sequence data collection with the concatenated multiple gene approach. However, concatenated multiple gene analyses of some key clades in life's history have not resolved phylogenetic trees due to homoplasy [40]. Therefore, an alternative methodology is still necessary.

It is assumed that all genes share the same evolutionary history and that increasing the number of nucleotides increases the signal for that evolutionary history for the concatenation approach. However, genes do not always share an identical evolutionary history, because of horizontal gene transfer, introgression, incomplete lineage sorting, or gene duplication/loss or do not share the same evolutionary divergence speed. In some cases, it may be more appropriate to analyze genes separately rather than concatenating all the genes [41-49]. A gene has its own function and exclusive role in biological processes, so it has its own evolutionary history with exclusive function constraints. Here, we test maximum gene-support tree approach [50] and the maximum gene-support branch approach to infer species tree from single gene trees that each single gene tree was reconstructed independently.

The number of alternative trees increases exponentially with the increasing number of taxa [51]. Reducing taxa can increase accuracy and decrease the probability of distorting the tree topology [52]. We used a jackknife approach to subsample six, five, and four taxa from the seven taxon dataset each time in order to reduce the number of taxa and, subsequently, to reduce the number of alternative trees. In this way, we expect to increase gene support values and study the feasibility of subsampling for the maximum gene-support approach.

It is clear that not all phylogenetic methods and genome-scale (or multiple-gene) approaches are equally powerful and reliable [27]. However, it is generally accepted that the convergence of several phylogenetic methods and approaches on the same topology can be taken as added evidence in support of a particular hypothesis [27]. To resolve the open question of the origin of tetrapod, we use three genome-scale approaches, e.g., the maximum gene-support tree approach [50], the majority-rule branch approach and the widely used concatenation approach, with three common phylogenetic methods, e.g., maximum likelihood, maximum parsimony, and neighbor-joining, to analyze all 43 nuclear genes that are available in GenBank at that time.

Adding characters can always increase the accuracy [52-54], and as many genes as possible should therefore be included. When 1001 shared genes become available recently, 1001 shared genes from proteome and transcriptomes were mined to infer phylogenetic relationships among lungfish, coelacanth, and tetrapod with the maximum gene-support tree approach.

## II. MATERIALS AND METHODS

### 2.1. Sequence Collection

The two datasets from previous studies were provided by the authors upon our request. The sequences of encoding amino acid residues of 43 genes were resampled and re-analyzed from 44 genes dataset [33] for our study. Having been compared with the supplementary materials [33], the lengths of some sequences were different (Supplementary Material Table S1). One gene (FSCN1) was not included, because some taxa lacked the FSCN1 sequence in GenBank [33]. To compare the results with the multiple-gene concatenated gene approach [33]), we used the same seven taxa: African lungfish (*P. annectens* ), coelacant (*L. chalumnae* ), zebrafish (*Danio rerio* ), frog (*Xenopus tropicalis* ), chicken (*Gallus gallus* ), human (*Homo sapiens* ) and catshark (*Scyliorhinus canicula* ), corresponding to lungfish (L), the coelacanth (C), ray-finned fish (R), amphibian (A), bird (B), mammal (M), and shark (S) with the 43 genes. Amino acid sequences were used for phylogenetic analysis.

The second dataset provided by Liang et al. (2013)[36] consisted of alignments of 1465 individual genes. We extracted alignments of 1001 genes for seven taxa and ten taxa sets, consisting of African lungfish (*P. annectens* ), coelacanth (*L. chalumnae* ), frog (*X. tropicalis* ), chicken (*G. gallus* ), human (*H. sapiens* ), zebrafish (*D. rerio* ) and elephant shark (*C. milii* ) from the 1290 gene dataset of 10 taxa set[36].

### 2.2. Phylogenetic Analysis

Sequences of an individual gene were aligned with the ClustalX tool using the default settings [57]. All alignments of single genes were edited to exclude insertions or deletions and uncertain positions from further analysis.

For 43 shared genes, the PAUP* phylogenetic analysis software program (version 4.0b10) [58] was used for tree inference with the maximum parsimony (MP) method. Each set of sequences of single genes or concatenated genes was analyzed under the optimality criteria for MP. The MP analyses were performed with unweighted parsimony. The sequences were also analyzed with the maximum likelihood (ML) and the neighbor-joining (NJ) with the PHYLIP phylogenetic analysis package using the default settings [51]. For the 1001 shared genes, all single gene trees were inferred with RAxML [59].

### *2.3. Concatenated Multiple-Gene Approach*

The first step was to concatenate small alignments of single genes into one large alignment, which was then used to reconstruct a tree [60-64]. The bootstrap consensus tree was searched using the branch-and-bound algorithm for MP, and the full heuristic search was used for NJ and ML-based on a 50% majority rule. 1,000 replicates were used except for ML, where 100 replicates were completed.

### *2.4. Maximum Gene-Support Tree Approach*

All single gene trees were recovered using all 43 individual genes as above with the MP, ML, and NJ methods, separately. The tree distances for all pairwise comparisons among trees were calculated, using the symmetric difference metric, with PAUP* [58] and PHYLIP [49]. This distance is the number of steps required to convert between two trees, that is, the number of branches that differ between a pair of trees [65]. Two trees with identical topology have a tree distance of zero. A maximum gene-support tree was defined as a unique tree that was recovered by the most genes of all the ones used [50]. A computer program in C language for calculating gene support is also available upon request (henry.*shan@gmail.com* ).

### *2.5. Maximum Gene-Support Branch Approach*

Based on all single gene trees recovered using all 43 individual genes as above, a majority rule consensus tree with a parameter setting of less than 50% was calculated with PAUP* [58]. Gene support was obtained for each branch using the corresponding support values. We also call it as the majority rule branch approach.

### *2.6. Taxon Jackknife Subsampling*

We used a jackknife approach to subsample six, five, and four taxa from the seven taxa each time in order to reduce alternative trees.

### 2.7. Chi-Square Test

The statistically significant difference between gene supports of tree I, tree II, or tree III was determined by the Chi-Square test, respectively. The statistically significant difference in the taxon jackknife support averages between the six-, five-, and four-taxon sets were also analyzed by means of the Chi-Square test.

## III. RESULTS

### 3.1. Seven-Taxon Set with 43 shared locus

Gene support is the number of genes that reconstruct a unique topology. As shown in Table 1, the gene supports were equal, namely two, for all four tree types with the MP method. Four tree types of seven taxa are shown in Fig. 2a to 2d. No unique maximum gene-support tree was identified. Therefore, this phylogeny was irresolvable using the maximum gene-support tree approach for these seven taxa with MP. The irresolvable results were also observed with the ML and NJ methods (Table 1).

Phylogenetic analysis with these three common phylogenetic methods and these three approaches did not converge, as shown in Table 1. The results clearly varied with the method and the approach. The hypothesis that lungfishes are the closest living relatives was inferred with ML with 100% bootstrap support, but the lungfish and coelacanth sister group was recovered with NJ with 87% support using the concatenated multiple-gene approach. The maximum gene-support tree approach clearly showed that 43 genes were not able to resolve the phylogenetic relationship for these seven taxa, regardless of the phylogenetic method.

### *3.2. Six-Taxon Sets*

Table 2 shows that gene support values of tree II were lower than those of tree I or tree III for all the methods. Significant differences in the gene support between tree II and tree I or between tree II and tree III inferred with MP were observed for MBACLR and MACLRS (Table 2) at $P < 0.10$ level by means of the Chi-Square test. With MP, there were significantly more gene supports for tree III than tree I for MACLRS at $P < 0.10$ level, but there were no significant differences in the gene supports for tree I and tree III for the other four six-taxon sets. Tree IV was supported by one gene for the MBCLRS taxon set only.

### *3.3. Five-Taxon Sets*

The Chi-Square test showed that, with MP, the gene support was significantly lower for tree II compared to tree III for MBCLR at$P < 0.05$ significance level. With ML, significantly lower gene supports were detected for tree II compared to tree I for BACLS at $P < 0.05$, MACLS at $P < 0.10$, and MBCLS at $P < 0.10$ (Table 2). There were no significant differences in gene supports between tree I and tree III of all nine five-taxon sets (Table 2).

### *3.4. Four-Taxon Sets*

Based on the Chi-Square test, significantly lower gene supports were observed for tree II compared to tree III or tree I for ACLR and BCLS at$P < 0.05$ significance level with NJ (Table 2) and for BCLS at $P < 0.05$ level with ML (Table 2). Significantly higher gene support was observed for tree III compared to tree I for ACLR at $P < 0.05$ level with NJ, but no significant differences were observed between tree I and tree III in the gene supports for the other five four-taxon sets (Table 2).

### 3.5. Taxon Jackknife Analysis

The taxon jackknife analysis (Table 3) showed that jackknife probability was 10.0% for tree II, 27.5% for tree I, and 62.5% for tree III with the maximum gene-support tree approach using MP. Zero probability for tree II, 40% jackknife probability for tree I, and 50% for tree III were observed with the concatenated multiple-gene approach using MP. Jackknife probability was 10% for tree II, 30% for tree I, and 60% for tree III with the maximum gene-support branch approach. Jackknife probability for tree IV was zero for the three approaches (Table 3) with MP. The Chi-Square test showed no significant differences between these taxon sampling sets. The results show that taxon sampling has no significant effect on phylogenetic inference for these taxon sets.

### 3.6. Seven-Taxon Set with the 1001 shared genes

$S$ equences data of 1001 shared genes become available in recent years, used the maximum gene approach successfully inferred the phylogeny of seven taxa with the ML method. Tree I (lungfish hypothesis) was reconstructed from 1001 ML single gene trees with the maximum gene-support tree approach, in which the most gene support value is 89. The second and the third most supported tree is tree III (lungfish-coelacanth sister hypothesis) with a gene support value of 56, and tree II (coelacanth hypothesis) with a gene support value of 50, while the difference in the two support values is no significant. The Chi-Square test shows a significant difference at $p < 0.01$ significant level (12.23**) and (17.09**) between tree I and tree III in gene support values and between tree I and tree II. The results show that lungfish is the closest living relative of vertebrates, which is supported by the most genes at 89 from the 1001 genes.

### 3.7. *Ten-Taxon Set with 1001 shared genes*

Similar results were observed for ten-taxon compared to those of the above seven-taxon set with 1001 shared genes. The maximum gene support value is 92. The second and the third most gene support tree is tree III (lungfish-coelacanth sister hypothesis) with a gene support value of 59, and tree II (coelacanth hypothesis) with a gene support value of 51, while the difference in the two gene support values was no significant. The Chi-Square test shows a significant difference at $p < 0.01$ significant level (11.83**) and (18.27**) between tree I and tree III in gene support values and between tree1 and tree II (Table 4).

## IV. DISCUSSION

When 43 shared genes are used to reconstruct the phylogeny of seven taxa, the maximum gene-support tree approach provides no resolution. However, the maximum gene-support branch approach infers tree III by MP and ML and tree II by NJ. Additionally, the concatenation approach recovers tree I with MP and ML, but tree III with NJ. These results show incongruence with the phylogenetic methods. The results of the maximum gene-support tree approach clearly show that 43 genes do not reach the threshold of the minimum number of genes required for the resolution of the phylogeny of these seven taxa. Because the number of alternative trees increases exponentially with the number of taxa (6945, 710,395, n(2n-3)!!) [51], the minimum number of required genes increases as the number of taxa increases. One way to meet the minimum gene requirement is to increase the number of genes, whereas another way is to decrease the number of taxa. The debate over taxon sampling has not ended. On the one hand, accuracy is enhanced with the addition of taxa [51-52, 66-67). On the other hand, adding taxa can reduce accuracy and increase the probability of distorting the tree topology [52]. We used a jackknife approach to subsample six, five, and four taxa from the seven taxa each time in order to reduce the number of taxa and, subsequently, the number of alternative trees. The results show no significant differences in taxon jackknife probabilities when the size of the taxon sets was reduced from six to five, then to four (Tables 3). However, the gene supports increased as the size of the taxon sets was reduced (Table 2). When the number of genes increases to 1001, consistent results are shown for seven taxa and ten taxa. Therefore, an increasing number of genes is a more effective way than increasing taxa. In recent years, the concatenated multiple-gene approach has been widely used to reconstruct phylogenetic relationships [63, 68-70]. Currently, the number of sampled genes seems to be arbitrary. The minimum number of genes required to resolve a phylogenetic tree for eight yeasts was 20 without statistical significance test [63] for the concatenated multiple-gene approach. In earlier cases, 15 to 50 genes could meet the minimum gene requirement to obtain congruent trees with the concatenated multiple-gene approach and the maximum gene-support approach without a statistical significance test [50, 71]. In this study, 43 shared genes of seven-taxon do not meet the minimum number of genes required with the maximum gene-support tree approach. The position of stramenopiles (a group of eukaryotes) and the relationships among Conosa (amoeba and slime mold), Opisthokonta (fungi and animal), and plants could not be settled by the use of more than 100 genes [72]. The minimum number of genes required varies with the method, taxon set, and type of base. When a reliable tree is not known, the determination of the minimum number of genes required is difficult. Bootstrap support of 100% does not mean that the branch is 100% correct. The level of bootstrap support of 100% may occur in an alternative branch [64]. High bootstrap support does not necessarily signify “the truth” [73]. Tree 1 (lungfish hypothesis) was confirmed with the 1001 shared gene with maximum gene support tree approach and hundreds up to five thousand of 1:1:1 orthologous genes from other whole-genome data sets with concatenation approach [35-38]. When a maximum gene-support value is not significantly different from the second most gene support, for example in the case of seven taxa with 43 shared genes, it can be recognized that the number of genes used does not meet the minimum gene requirement. It means information is not enough to get a solution for this clad. More genes are required. This is the outstanding advantage of the maximum gene-support tree approach with the Chi-Square test.

It always is recommended to perform cross-validation using several approaches and methods while dealing with recalcitrant nodes. In this study, we used the above three approaches and three methods and obtained more confident conclusions than a single approach. It is highly possible to apply the proposed pipeline to prove other biological hypotheses. In the tree of life, there are other contentious phylogenetic relationships to wait for resolution for decades, such as the root of placental mammals, the phylogenetic relationships among lineages at the origin of land plants, and among lamprey, hagfish, jawed vertebrates [34], and so on. Especially, the maximum gene support approach, an alternative approach, may be especially suitable for resolving incongruence on ancient and short internodes or short divergence time internodes of phylogenetic trees. When the second most support value is close to the maximum gene support value, even though the maximum gene support value is very great, a “true tree” or species tree cannot be identified. In contrast, when the second most support value is significantly different from the maximum gene support value with the Chi-square test, even though the maximum gene support value is not great, a “true tree” or species tree can be identified. Therefore, the second-to-maximum ratio (SM ratio) in gene support values is a very proper support index to identify a species tree from gene trees. The SM ratio is the ratio of the second most gene support value to the maximum gene support (the most gene support) value in a set of all gene support values in detail. It is somehow like the noise to signal ratio. The SM ratio, a relative value, is a better support index than a single absolute value of phylogenetic supports to show the insight of phylogenetic branch supports. The SM ratio should be further studied in the future as an alternative index of the widely used bootstrap support. The maximum gene support tree approach can avoid distortions caused by systematic errors, long branch attraction (LBA) artifacts, and compositional heterogeneity because each single gene tree is independently reconstructed and all distorted trees are not identical to the maximum gene supported tree. Therefore, those distorted gene trees are excluded as species tree with this approach. Large data of many genes and valid models are always desirable and basic because they increase the probabilities of showing significant differences in the SM ratio and Chi-square test, and further generating accurate phylogenies. The selection of a more balanced taxon set and proper outgroup will also be beneficial to congruence in the species tree inference from single gene trees by the maximum gene tree approach, which should be addressed in the future, particularly when dealing with very subtle nodes with conflictive clad inferences for reconstructing the tree of life. Considering that the co-evolving patterns from the sequence information of proteins (74-75) can be extracted, these findings should be integrated into the proposed pipeline as future work.

Tree IV received nearly zero gene support and taxon jackknife probability (Tables 2 and 3). Therefore, the results sufficiently show that the irresolvable trichotomy of lungfish, coelacanth, and tetrapod is not a hypothesis. A reinvestigation using 44 genes with a concatenation genome-scale approach showed an unresolved trichotomy [33]. The major paleontological studies published in the last decade proposed that lungfishes (Dipnoi) are the closest living relatives of tetrapods or, alternatively, that coelacanths and lungfishes form a monophyletic group that is equally closely related to tetrapods [76-77]. The cause of this puzzle is the fact that the divergence of coelacanths and lungfishes happened over a relatively short period within a small window in time (20–30 million years) around 400 million years ago [3, 12]. The result was due to little time and opportunity for lineage-specific molecular changes to happen, yet considerable time and opportunity for multiple and parallel changes and their accumulation since the origin of these two lineages [3].

In our study, no consistent results were observed between tree I and tree III in taxon jackknife probabilities with the concatenation approach across the three phylogenetic methods with 43 shared gene datasets (Table 3). However, the clear consensus is that tree II received significantly lower gene support than tree I or tree III and, evidently lower taxon jackknife probabilities with all the phylogenetic methods and multiple-gene approaches (Tables 2 and 3). These results favor lungfish, but not coelacanth, as the closest living relative of tetrapod, based on all these phylogenetic analyses from all 43 shared gene datasets. These results are consistent with previous molecular and paleontological phylogenetic analyses [3]. A similar suggestion was made based on mitochondrial DNA sequences [16].

For 43 shared gene datasets with the gene support tree approach, tree I is consistently reconstructed with the ML method for 4 taxon sets with $P < 0.10^+$ or $0.05^*$(Chi-Square test) compared to Tree II in gene support values (Table 2). Significant higher gene-support values of tree I were consistently observed for the

BACLR at $P < 0.05$, MACLS at $P < 0.10$, MBCLS at $P < 0.10$, and BCLS at $P < 0.05$ compared to those of tree II using the ML method with maximum gene-support approach (Table 2). However, no consistent tree is inferred with NJ and MP methods (Table 2). Generally, the accuracy of the ML method is higher than the MP or NJ method. Moreover, only tree I was observed with 100% bootstrap supports for the MBACLRS, MBACLR, ACLS with ML method, the BACLS, MACLS, MBCLS with MP method, BACLS with NJ method among all combinations of taxon samplings and methods using concatenation approach (Table 3), but no 100% bootstrap supports of tree II or tree III (Table 3). Therefore, these results also clearly support lungfish-tetrapod (tree I) hypothesis, but not the coelacanth-tetrapod (tree II) hypothesis.

Adding characters can always increase the accuracy [52-54], and as many genes as possible should therefore be included. When sequences data of 1001 shared genes become available, we use this large dataset with the maximum gene-support approach and successfully inferred the lungfish as the closest living relative of land vertebrates with a significant difference at $p < 0.01$ (Chi-Square test) with the ML methods. Tree I (lungfish hypothesis) is reconstructed by the 1001 ML single gene trees with the maximum gene-support tree approach, in which the most gene support value is 89. The second most gene support tree is tree III (lungfish-coelacanth sister hypothesis) with a gene support value of 56. The third most gene support tree is tree II (coelacanth hypothesis) with a gene support value of 50. A significant difference is observed at $p < 0.01$ significant level (12.23**) and (17.09**) between tree I and tree III in gene support values and between tree1 and tree II, respectively. The fourth and fifth most gene support value are 38 and 36, separately, which are significantly lower than those of tree I, tree II, or III. No evident difference was observed when taxa were increased from seven to ten in this study. Our results show that lungfish is the closest living relative of landing vertebrates, which is supported by the most genes of 89 of 1001 shared genes of seven –taxon dataset and 92 of ten-taxon (Table 4). Tree II and tree III also received significantly more gene support than tree IV, tree V, and others. These observations show tree II and tree III hypotheses are not arbitrary guesses, but somehow intrinsic support from the second most and third most genes. Additional insight into the arguments of the three hypotheses can be gained by examining the gene support values for each hypothesis.

Our results also show that increasing the number of shared genes is more effective than increasing the number of taxa. When the number of genes or sequence data is not enough, trees are very incongruent and vary widely with multiple-gene approaches, taxon samples, and phylogenetic inference methods. When sequence data of genes are not enough, whichever phylogenetic method is used, it is not able to produce a valid phylogenetic tree.

In conclusion, we have successfully provided evidence in favor of the lungfish-tetrapod hypothesis (tree I), but rejecting the coelacanth-tetrapod hypothesis (tree II) based on the above phylogenetic analysis results using 43 genes with all three common phylogenetic/phonetic methods and three genome-scale approaches, and further additional analysis using the 1001 shared genes with the maximum gene support. This result is consistent with those based on whole genome data with a concatenation approach [35-38], although there were incongruent cases of phylogenomic analysis [34, 39].

## ACKNOWLEDGMENT

We thank Richard Winterbottom for critically reading an earlier draft and providing helpful comments and valuable suggestions.

## REFERENCES

[1] B. Fritzsch, “Inner ear of the coelacanth fish *Latimeria*has tetrapod affinities,’ *Nature,* vol. 327, pp. 153-154, 1987.

[2] K. S. Thomson, *Living fossil* , W. W. Norton & Company, New York, 1991.

[3] A. Meyer, “Molecular evidence on the origin of tetrapods and the relationships of the coelacanth,’ *Trends Ecol Evol.,* vol. 10, pp. 111–116, 1995.

[4] A. S. Romer, *Vertebrate paleontology* , University of Chicago Press, Chicago, 1966.

[5] J. A. Long, [1989] "A new rhizodontiform fish from the Early Carboniferous of Victoria, Australia, with remarks on the phylogenetic position of the group," *J. Vert. Paleontol.,* vol. 9, pp. 1–17.

[6] G. C. Young, J. A. Long and A. Ritchie, "Crossopterygian fishes from the Devonian of Antarctica: systematics, relationships, and biogeographic significance," *Rec. Austral. Mus.,* vol. 14 [Suppl], pp. 1–77, 1992.

[7] M. Zhu and H. P. Schultze, "The oldest sarcopterygian fish,"*Lethaia,* vol. 30, pp. 293–304, 1997.

[8] D. E. Rosen, P. L. Forey, B. G. Gardiner and C. Patterson, "Lungfishes, tetrapods, paleontology, and plesimorphy," *Bull. Am. Mus. Nat. Hist.,* vol. 167, pp. 159–276, 1981.

[9] B. G. Gardiner, "The relationships of the palaeoniscid fishes, a review based on new specimens of Mimia and Moythomasia from the Upper Devonian of Wester Australia," *Bull. Brit. Mus. Nat. Hist. (Geol.),* vol. 37, pp. 173–428, 1984.

[10] J. G. Maisey, "Heads and tails: a chordate phylogeny,"*Cladistics,* vol. 2, pp. 201–256, 1986.

[11] A. L. Panchen and T. S. Smithson, "Character diagnosis, fossils and the origin of tetrapods," *Biol. Rev.,* vol. 62, pp. 341–438, 1987.

[12] P. E. Ahlberg, "Postcranial stem tetrapod remains from the Devonian Scat Craig, Morayshire," *Scotland Zool. J. Linn. Soc.,*vol. 103, pp. 241–287, 1991.

[13] R. G. Northcutt, *The biology and evolution of lungfishes* , eds W. E. Bemis, W. W. Burggren, N. E. Kemp and R. A. Liss, New York, pp 277–297, 1986.

[14] M. M. Chang, *Origins of the higher groups of tetrapods: controversy consensus* , eds H. P. Schultze, L. Trueb, Cornell University Press, Ithaca, New York, pp 3–28.

[15] P. L. Forey, B. G. Gardiner and C. Patterson, *Origins of the higher groups of tetrapods: controversy consensus* , eds. P. H. Schultze, L. Trueb (Cornell University Press, Ithaca, New York), pp. 145-172, 1991.

[16] A. Meyer and A. C. Wilson, "Origin of tetrapods inferred from their mitochondrial DNA affiliation to lungfish," *J. Mol. Evol.,*vol. 31, pp. 359–364, 1990.

[17] B. B. Normark, A. R. McCune and R. G. Harrison, "Phylogenetic relationships of neopterygian fishes, inferred from mitochondrial DNA sequences." *Mol. Biol. Evol.,* vol. 8, pp. 819–834, 1991.

[18] A. Meyer and S. I. Dolven, "Molecules, fossils, and the origin of tetrapods," *J. Mol. Evol.,* vol. 35, pp. 102–113, 1992.

[19] S. B. Hedges, C. A. Hass and L. R. Maxson, "Relations of fish and tetrapods," *Nature,* vol. 363, pp. 501–502, 1993.

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

[21] R. Zardoya and A. Meyer, "Evolutionary relationships of the coelacanth, lungfish, and tetrapods based on the 28S ribosomal RNA gene," *Proc Natl Acad Sci USA,* vol. 93, pp. 5449–5454, 1996.

[22] R. Zardoya and A. Meyer, "The complete DNA sequence of the mitochondrial genome of a 'living fossil,' the coelacanth (Latimeria chalumnae)," *Genetics* , vol. 146, pp. 995-1010, 1997.

[23] Y. Cao, P. J. Waddell, N. Okada and M. Hasegawa, "The complete mitochondrial DNA sequence of the shark Mustelus manazo: evaluating rooting contradictions to living bony vertebrates,". *Mol. Biol. Evol.,* vol. 15, pp. 1637–1646, 1998.

[24] R. Zardoya, Y. Cao, M. Hasegawa and A. Meyer, "Searching for the closest living relative[s] of tetrapods through evolutionary analyses of mitochondrial and nuclear data," *Mol. Biol. Evol.,*vol. 15, pp. 506–517, 1998.

[25] Y. Tohyama, T. Ichimiya, H. Kasama-Yoshida, Y. Cao, M. Hasegawa, H. Kojima, Y. Tamai and T. Kurihara, "Gene structure and amino acid sequence of *Latimeria chalumnae* [coelacanth] myelin DM20: phylogenetic relation of the fish," *Mol. Brain Res* ., vol. 80, pp. 256–259, 2000.

[26] B. Venkatesh, M. V. Erdmann and S. Brenner, "Molecular synapomorphies resolve evolutionary relationships of extant jawed vertebrates," *Proc Natl Acad Sci USA,* vol. 98, pp. 11382– 11387, 2001.

[27] H. Brinkmann, B. Venkatesh, S. Brenner and A. Meyer, "Nuclear protein-coding genes support lungfish and not the coelacanth as the closest living relatives of land vertebrates," *Proc Natl Acad Sci USA,* vol. 101, pp. 4900-4905, 2004.

[28] T. Gorr, T. Kleinschmidt and H. Fricke, "Close tetrapod relationships of the coelacanth Latimeria indicated by haemoglobin sequences," *Nature,* vol. 351, pp. 394–397, 1991.

[29] P. M. Sharp and A. T. Lloyd, "Coelacanth's relationships,"*Nature,* vol. 353, pp. 218–219, 1991.

[30] D. W. Stock and D. L. Swofford, "Coelacanth's relationships,"*Nature* , vol. 353, pp. 217–218, 1991.

[31] D. W. Stock, K. D. Moberg, L. R. Maxson and G. S. Whitt, "A phylogenetic analysis of the 18S ribosomal RNA sequence of the coelacanth Latimeria chalumnae," *Env Biol Fishes* 32:99–117, 1991.

[32] Y. Shan and R. Gras, "43 genes support the lungfish-coelacanth grouping related to the closest living relative of landing vertebrates with the Bayesian method under the coalescence model," *BMC Research Notes,* **vol. 4, pp.** 49, 2011.

[33] N. Takezaki, F. Figueroa, Z. Zaleska-Rutczynska, N. Takahata and J. Klein, "The Phylogenetic Relationship of Tetrapod, Coelacanth, and Lungfish Revealed by the Sequences of Forty-Four Nuclear Genes,"*Mol. Biol. Evol.,* Vol. 21, pp. 1512-152, 2004.

[34] I. Irisarri and A. Meyey, "The Identification of the Closest Living Relative(s) of Tetrapods: Phylogenomic Lessons for Resolving Short Ancient Internodes, " *Systematic Biology,* Vol. 65, No. 6, pp. 1057-1075, Nov. 2016.

[35] C. T. Amemiya, J. Alfoldi, A. P. Lee, S. Fan, H. Philippe, I. Maccallum, I. Braasch, T. Manousaki, I. Schneider, N. Rohner et al., "The African coelacanth genome provides insights into tetrapod evolution," *Nature* , vol. 496, pp. 311–316, 2013.

[36] D. Liang, X. X. Shen and P. Zhang, "One Thousand Two Hundred Ninety Nuclear Genes from a Genome-Wide Survey Support Lungfishes as the Sister Group of Tetrapods," *Journal of Molecular and Evolutionary Biology* , vol. 30, pp. 1803–1807, 2014.

[37] W. Kun, J. Wang, C. Zhu et al., (20 co-authors), "African lungfish genome sheds light on the vertebrate water-to-land transition," *Cell* , Vol. 184, pp. 1-15, 2021.

[38] A. Meyer, S. Schloissnig, P. Franchini and Kang Du *et al.,* "Giant lungfish genome elucidates the conquest of land by vertebrates," *Nature,* vol. 590, pp. 284–289, 2021 .

[39] N. Takezaki and H. Nishihara, "Resolving the Phylogenetic Position of Coelacanth: The Closest Relative Is Not Always the Most Appropriate Outgroup," Genome Biology and Evolution, vol. 8, no. 4, pp. Pages 1208–1221, Apr. 2016.

[40] A. Rokas and S. B. Carroll, "Bushes in the Tree of Life,"*PLoS Biol* ., vol. 4, no. 11, Nov. 2006.

[41] E. N. III, Adams, "Consensus techniques and the comparison of taxonomic trees," *Syst. Zool* ., vol. 21, pp. 390-397, 1972.

[42] M. F. Mickevich, "Taxonomic congruence," *Syst. Zool.* , vol. 27, pp. 143-158, 1978.

[43G. J. Nelson, 1979. Cladistic analysis and synthesis: Principles and definitions, with a historical note on Adanson's Families des Plantes (1763-1764). *Syst. Zoo.* , vol. 28, pp. 1-21, 1979.

[44]D. M. Hillis, "Molecular versus morphological approaches to systematics," *Annu. Rev. Ecol. Syst.* , vol. 18, pp. 23-42, 1987.

[45] D. I. Swofford, "When are phylogeny estimates from molecular and morphological data incongruent? "Phylogenetic analysis of DNA sequences, M. M. Miyamoto and J. Cracraft, Eds. Oxford Univ. Press, New York, pp. 295-333, 1991.

[46]A. De Queiroz, "For consensus (sometimes)," Syst. Bio., vol. 42, pp. 368-372, 1993.

[47]A. G. Rodrigo, M. Kelly-Borges, P. R. Bergquist and P. L. Bergquist, "A randomization test of the null hypothesis that two cladograms are sample estimates of a parametric phylogenetic tree," N. Z. J. Bot., vol. 31, pp. 257-258, 1993.

[48]C. Ane, J. G. Burleigh, M.M. McMahon and M. J. Sanderson, "Covarion structure in plastid genome evolution: A new statistical test," Molecular Biology and Evolution, vol. 22, pp. 914–924, 2005.

[49]T. A. Heath, S. M. Hedtke and D. M. Hills, "Taxon sampling and the accuracy of phylogenetic analyses," Journal of Systematics and Evolution, vol. 46, pp. 239–257, 2008a.

[50]Y. Shan and X. Li, "Maximum Gene-support tree. Evolutionary Bioinformatics.," vol. 4, pp. 181-191, 2008.

[51]I. Felsenstein, Inferring Phylogenies, Sinauer Associates, Inc., Sunderland, Massachusetts, pp. 19-36, 2004.

[52]A. Graybeal, "is it better to add taxa or characters to a difficult phylogenetic problem? "Syst. Biol., vol. 47, pp. 9–17, 1998.

[53]S. Poe and D. L. Swofford, "Taxon sampling revisited," Nature, vol. 398, pp. 299–300, 1999.

[54]M. S. Rosenberg and S. Kumar, "incomplete taxon sampling is not a problem for phylogenetic inference," Proc Natl Acad Sci USA, vol. 98, pp. 10751-10756, 2001.

[55]Y. Shan and R. Gras, "Genome-wide EST data mining approaches to resolving incongruence of molecular phylogenies," In: Arabnia HA, editors. Advances in Computational Biology. New York: Springer, pp. 237-243, 2010.

[56]M. G. Grabherr, B. J. Haas, M. Yassour, J. Z. Levin, D. A.Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, Z. Chen, E. Mauceli, N. Hacohen, A. Gnirke, N. Rhind, F. di Palma, B. W. Birren, C. Nusbaum, K. Lindblad-Toh, N. Friedman and A. Regev, "Full-length transcriptome assembly from RNA-Seq data without a reference genome," Nat. Biotechnol., vol. 29, No. 7, pp. 644-52, 2011, doi: 10.1038/nbt.1883.

[57]J. D. Thompson, T. J. Gibson, F. Plewniak, F. Jeanmougin and D. G. Higgins, "The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools," Nucleic Acids Res., vol. 24, pp. 4876-4882, 1997.

[58]D. L. Swofford, PAUP*: Phylogenetic Analysis Using Parsimony (*and Other Methods), Version 4. Sinauer Associates, Sunderland, Massachusetts, 2002.

[59]A. Stamatakis, "RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies," Bioinformatics, vol. 30, pp. 1 1312–1313, 2014.

[60]A., G. Kluge, "A concern for evidence, and a phylogenetic hypothesis of relationships among Epicrates [Boidae, Serpentes]," Syst. Zool., vol. 38, pp. 7-25, 1989.

[61] J. P. Huelsenbeck, J. J. Bull and C. W. Cunningham, “Combining data in phylogenetic analysis,” TREE, vol. 11, pp.152-158, 1996.

[62] Z. Yang, “Maximum-likelihood models for combined analysis of multiple sequence data,” J. Mol. Evol., vol. 42, pp. 587-596.

[63]A. Rokas, B. L. Williams, N. King and S. B. Carroll, “Genome-scale approaches to resolving incongruence in molecular phylogenies,” Nature, vol. 425, pp.798–804, 2003.

[64] D. E. Soltis et al., “Genome-scale data, angiosperm relationships, and ending incongruence: a cautionary tale in phylogenetics,” Trends Plant Sci., vol. 9, pp. 477-483, 2004

[65]D. R. Robinson, and L. R. Foulds, “Comparison of Phylogenetic Trees,” Math Biosciences, vol. 3, pp. 131-147, 1981

[66] S. M. Hedtke, T. M. Townsend and D. M. Hillis ”Resolution of phylogenetic conflict in large data sets by increased taxon sampling,” Systematic Biology, vol. 55, vol. 522–529, 2006.

[67] T. A. Heath, D. J. Zwickl, J. Kim and D. M. Hillis, “Taxon sampling affects inferences of macroevolutionary processes from phylogenetic trees,” Systematic Biology, vol. 57, pp.160–166, 2008b.

[68] J. W. Murphy, E. Eizirik, W. E. Johnson, Y. P. Zhang, O. A. Ryder and O. J. O’Brien, “Molecular phylogenetics and the origins of placental mammals,” Nature vol. 409, pp. 14–618, 2001.

[69] O. Madsen, M. Scally, C. J. Douady, D. J. Kao, R. W. DeBry, R. Adkins, H. M. Amrine, M. J. Stanhope, W. W. de Jong and M. S. Springer, “Parallel adaptive radiations in two major clades of placental mammals., “Nature, vol. 409, pp. 610–614, 2001.

[70] Y. I. Wolf, I, G. Rogozin and E. V. Koonin, “Coelomata and not Ecdysozoa: evidence from genome-wide phylogenetic analysis, “Genome Res., vol. 14, pp.9–36, 2004.

[71] E. A. Herniou, T. Luque, X. Chen, J. M. Vlak, D. Winstanley, J. S. Cory and D. R. O’Reilly, “Use of whole genome sequence data to infer baculovirus phylogeny,”’ J. Virol., vol. 5, pp. 8117-8126, 2001.

[72] E. Bapteste, H. Brinkmann, A. J. Lee et al [11 co-authors], “The analysis of 100 genes supports the grouping of three highly divergent amoebae: Dictyostelium, Entamoeba, and Mastigamoeba,” Proc Natl Acad Sci USA, vol. 99, pp. 1414–1419, 2002.

[73] M. J. Phillips, F. Delsuc and D. Penny, “Genome-scale phylogeny and the detection of systematic biases,” Mol. Biol. Evol., vol.21, pp. 1455-1458, 2004.

[74] L. Hu and K. C. Chan, “Extracting coevolutionary features from protein Sequences for predicting protein-protein interactions,” IEEE/ACM Trans Comput Biol Bioinform, vol. 14, No. 1, pp. 155-166, Jan. 2017.

[75] Hu, J. Zhang, X. Pan, H. Yan, and Z-H. You, “HiSCF: leveraging higher-order structures for clustering analysis in biological networks, ” Bioinformatics, vol. 37, No.4, pp. 542-550, May 2021.[76]M.Zhu and X. Yu, “A primitive fish close to the common ancestor of tetrapods and lungfish,” Nature, vol. 418, pp. 767-770, 2002.

[77] M. Zhu, X. B. Yu and P.E. Ahlberg, “A primitive sarcopterygian fish with an eyestalk,” Nature, vol. 410, pp. 81-84, 2001.

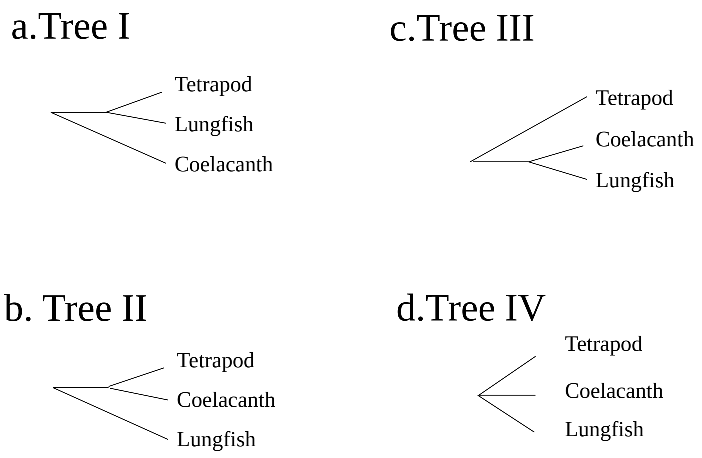
a.Tree I
Tetrapod
Lungfish
Coelacanth
c.Tree III
Tetrapod
Coelacanth
Lungfish
b. Tree II
Tetrapod
Coelacanth
Lungfish
d.Tree IV
Tetrapod
Coelacanth
Lungfish

Fig. 1.

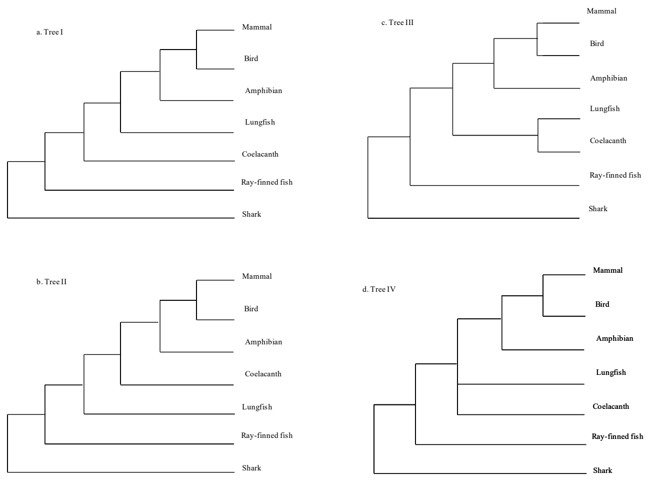
a. Tree I
Mammal
Bird
Amphibian
Lungfish
Coelacanth
Ray-finned fish
Shark
c. Tree III
Mammal
Bird
Amphibian
Lungfish
Coelacanth
Ray-finned fish
Shark
b. Tree II
Mammal
Bird
Amphibian
Coelacanth
Lungfish
Ray-finned fish
Shark
d. Tree IV
Mammal
Bird
Amphibian
Lungfish
Coelacanth
Ray-finned fish
Shark

**Fig. 2.**

**Table 1**

Tree types of seven taxa with three methods and three genome-scale approaches

| | Phylogenetic Methods | | |
|---|---|---|---|
| Genome-Scale Approaches | MP | ML | NJ |
| Concatenation | I (67%) | I (100%) | III (87%) |
| Maximum Gene Support Tree | I(2)/II(2)/III(2)/IV(2) | I(2)/II(1)/III(2)/IV(0) | I(2)/II(1)/III(1)/IV(0) |
| Maximum Gene Support Branch | III (9) | III (9) | II (10) |

Notes: The numbers in ( ) are gene supports for the maximum gene support tree and maximum gene support branch approaches , bootstrap supports (%) for the concatenation approach, respectively. MP = Maximum parsimony; ML = Maximum likelihood; NJ = Neighbor joining.

**Table 2**

Gene supports for four tree types and six-, five-, and four-taxon sets inferred with MP, ML, and NJ

| | Tree type | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | MP | | | | ML | | | | NJ | | | |
| Taxon set | Tree I | Tree II | Tree III | Tree IV | Tree I | Tree II | Tree III | Tree IV | Tree I | Tree II | Tree III | Tree IV |
| Six taxon sets: | | | | | | | | | | | | |
| BACLRS | 2 | 3 | 3 | 0 | 3 | 1 | 5 | 0 | 4 | 3 | 3 | 0 |
| MACLRS | 1 | $1^{+}$ | 6 | 0 | 3 | 3 | 2 | 0 | 1 | 1 | 2 | 0 |
| MBACLR | 6 | $1^{+}$ | 6 | 0 | 4 | 3 | 5 | 0 | 3 | 3 | 8 | 0 |
| MBACLS | 4 | 3 | 5 | 0 | 6 | 3 | 4 | 0 | 2 | 7 | 5 | 0 |
| MBCLRS | 5 | 3 | 7 | 1 | 5 | 4 | 5 | 0 | 7 | 3 | 3 | 0 |
| Five taxon sets: | | | | | | | | | | | | |
| ACLRS | 6 | 5 | 6 | 0 | 8 | 4 | 3 | 0 | 7 | 4 | 3 | 0 |
| BACLR | 6 | 6 | 8 | 0 | 7 | 4 | 10 | 0 | 8 | 4 | 10 | 0 |
| BACLS | 9 | 4 | 8 | 0 | 13 | $4^{*}$ | 6 | 0 | 10 | 9 | 6 | 0 |
| BCLRS | 8 | 4 | 6 | 0 | 8 | 5 | 5 | 0 | 8 | 7 | 3 | 0 |
| MACLR | 7 | 5 | 9 | 0 | 6 | 6 | 9 | 0 | 5 | 4 | 10 | 0 |
| MACLS | 4 | 5 | 10 | 0 | 9 | $3^{+}$ | 8 | 0 | 3 | 9 | 8 | 0 |
| MBCLR | 10 | $4^{*}$ | 14 | 0 | 12 | 8 | 15 | 0 | 12 | 8 | 15 | 0 |
| MBCLS | 10 | 11 | 8 | 0 | 14 | $6^{+}$ | 12 | 0 | 15 | 10 | 9 | 0 |
| MCLRS | 5 | 6 | 9 | 0 | 5 | 4 | 6 | 0 | 6 | 4 | 3 | 0 |
| Four taxon sets: | | | | | | | | | | | | |
| ACLR | 13 | 13 | 16 | 1 | 16 | 12 | 15 | 0 | 9 | $10^{*}$ | 24 | 0 |
| ACLS | 14 | 14 | 13 | 2 | 16 | 13 | 14 | 0 | 14 | 12 | 17 | 0 |
| BCLR | 19 | 11 | 13 | 0 | 15 | 11 | 17 | 0 | 16 | 16 | 11 | 0 |
| BCLS | 18 | 11 | 13 | 1 | 21 | $8^{*}$ | 14 | 0 | 22 | $8^{*}$ | 13 | 0 |
| MCLR | 12 | 13 | 17 | 1 | 12 | 15 | 16 | 0 | 10 | 14 | 19 | 0 |
| MCLS | 11 | 14 | 16 | 2 | 13 | 13 | 17 | 0 | 13 | 13 | 17 | 0 |

Notes: MP = Maximum parsimony; ML = Maximum likelihood; NJ = Neighbor joining. The taxa used were mammal (M), bird (B), amphibian (A), coelacanth (C), lungfish (L), ray-finned fish (R), and shark (S). $^{+}$ and $^{*}$ indicate chi- square test significance levels of $P < 0.10$ and 0.05 between the gene supports for tree II and tree I/III, respectively.

**Table 3**

Bootstrap supports, gene supports, and taxon jackknife probabilities for four tree types and six-, five-, and four-taxon sets using three approaches recovered with MP, ML, and NJ

| | Approaches and methods | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | MP | | | ML | | | NJ | | |
| Taxon set | CT | MGT | MGB | CT | MGT | MGB | CT | MGT | MGB |
| Six-taxon sets: | | | | | | | | | |
| BACLRS | I (73%) | III (3) /II (3) | III (9) | I (53%) | III (5) | III (11) | III (79%) | I (4) | I (9) |
| MACLRS | AT (n/a) | III (6) | III (11) | I (47%) | III (3)/II (3) | I (17) | III (95%) | III (2) | III (6) |
| MBACLR | III (86%) | III (6) /I (6) | III (12) | I (100%) | III (5) | III (16) | III (96%) | II (7) | III (21) |
| MBACLS | I (80%) | III (5) | III (11) | I (53%) | I (6) | I (20) | III (69%) | III (8) | III (9) |
| MBCLRS | I (61%) | III (7) | III (14) | I (50%) | III (5)/I (5) | III (9) | III (85%) | I (7) | I (7) |
| Five-taxon sets: | | | | | | | | | |
| ACLRS | III (52%) | III (6)/I (6) | I (11) | I (32%) | I (8) | III (11) | III (92%) | I (7) | I (9) |
| BACLR | III (85%) | III (8) | II (20) | III (48%) | III (10) | III (15) | III (95%) | III (10) | II (12) |
| BACLS | I (100%) | I (9) | I (21) | I (22%) | I (13) | I (21) | I (100%) | I (10) | I (12) |
| BCLRS | I (87%) | I (8) | I (18) | I (46%) | I (8) | I (14) | III (59%) | I (8) | I (9) |
| MACLR | III (94%) | III (9) | III (21) | III (52%) | III (9) | III (16) | III (99%) | III (10) | III (12) |
| MACLS | I (100%) | III (10) | III (19) | I (46%) | I (9) | III (12) | III (81%) | II (9) | III (12) |
| MBCLR | I (73%) | III (14) | III (32) | II (43%) | III (15) | III (16) | III (88%) | III (15) | III (13) |
| MBCLS | I (100%) | II (11) | II (29) | II (40%) | I (14) | I (32) | III (75%) | I (15) | II (13) |
| MCLRS | III (53%) | III (9) | III (13) | III (45%) | III (6) | II (15) | III (96%) | I (6) | III (9) |
| Four-taxon sets: | | | | | | | | | |
| ACLR | III (95%) | III (16) | III (17) | III (55%) | I (16) | III (16) | III (99%) | III (24) | III (20) |
| ACLS | AT (n/a) | I/II (14) | I (14) | I (100%) | I (16) | II (17) | IV (49%) | III (17) | I (13) |
| BCLR | III (51%) | I (19) | I (19) | I (51%) | III (17) | III (17) | III (75%) | I (16) or II (16) | I (19) |
| BCLS | III (80%) | I (18) | I (18) | I (46%) | I (21) | I (21) | I (64%) | I (22) | I (15) |
| MCLR | III (80%) | III (17) | III (17) | III (49%) | III (16) | III (16) | III (93%) | III (19) | III (17) |
| MCLS | III (61%) | III (16) | III (16) | III (50%) | III (17) | III (18) | III (92%) | III (17) | III (16) |
| JKF: | I (40%) | I (27.5%) | I (30%) | I (60%) | I (47.5%) | I (30%) | I (10%) | I (42.5%) | I (40%) |
| | II (0) | II (10%) | II (10%) | II (10%) | II (2.5%) | II (10%) | II (0) | II (12.5%) | II (10%) |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | III (50%) | III (62.5%) | III (60%) | III (30%) | III (50%) | III (60%) | III (85%) | III (45%) | III (50%) |
| | IV (0) | IV (0) | IV (0) | IV (0) | IV (0) | IV (0) | IV (5%) | IV (0) | IV (0) |
| | AT (10%) | AT (0) | AT (0) | AT (0) | AT (0) | AT (0) | AT (0) | AT (0) | AT (0) |

Notes: MP = Maximum parsimony; ML = Maximum likelihood; NJ = Neighbor joining; CT = Concatenation tree; MGT = Maximum gene-support tree; MGB = Maximum gene-support branch; AT = Alternative tree; n/a = Not available; JKF = Taxon jackknife probabilities (%). The numbers in parentheses are bootstrap supports (%) for CT and gene supports for MGT and MGB.

**Table 4**

Gene supports for four tree types and seven- and ten-taxon sets of 1001 share genes inferred with ML

| | Tree type | | | |
|---|---|---|---|---|
| Taxon set | Tree I | Tree II | Tree III | Tree IV |
| Seven taxon set:MBACLRS | 89** | 50 | 56 | 0 |
| Ten taxon set: MBFACLRTES | 92** | 59 | 51 | 0 |

Notes: ML = Maximum likelihood; The taxa used were mammal (M), bird (B), Fugu (F), amphibian (A), coelacanth (C), lungfish (L), ray-finned fish (R), Little shark ( T), Elephant shark ( E ) and cat shark (S). ** indicate s chi- square test significance levels of $P < 0.01$ between the gene supports for tree II and tree I/III, respectively.