Biology Faculty Publications

Department of Biology

2008

# Optimal Data Partitioning and a Test Case for Ray-Finned Fishes (Actinopterygii) Based on Ten Nuclear Loci

Chenhong Li
*University of Nebraska-Lincoln*

Guoqing Lu
*University of Nebraska at Omaha*

Guillermo Orti
*University of Nebraska-Lincoln*

# Optimal Data Partitioning and a Test Case for Ray-Finned Fishes (Actinopterygii) Based on Ten Nuclear Loci

CHENHONG LI,[1] GUOQING LU,[2] AND GUILLERMO ORTÍ[1]

[1]*School of Biological Sciences, University of Nebraska, Lincoln, NE 68588, USA; E-mail: cli@unlserve.unl.edu (C.L.); gorti@unlserve.unl.edu (G.O.)*
[2]*Department of Biology, University of Nebraska, Omaha, NE 68182, USA; E-mail: glu3@mail.unomaha.edu*

*Abstract.*—Data partitioning, the combined phylogenetic analysis of homogeneous blocks of data, is a common strategy used to accommodate heterogeneities in complex multilocus data sets. Variation in evolutionary rates and substitution patterns among sites are typically addressed by partitioning data by gene, codon position, or both. Excessive partitioning of the data, however, could lead to overparameterization; therefore, it seems critical to define the minimum numbers of partitions necessary to improve the overall fit of the model. We propose a new method, based on cluster analysis, to find an optimal partitioning strategy for multilocus protein-coding data sets. A heuristic exploration of alternative partitioning schemes, based on Bayesian and maximum likelihood (ML) criteria, is shown here to produce an optimal number of partitions. We tested this method using sequence data of 10 nuclear genes collected from 52 ray-finned fish (Actinopterygii) and four tetrapods. The concatenated sequences included 7995 nucleotide sites maximally split into 30 partitions defined a priori based on gene and codon position. Our results show that a model based on only 10 partitions defined by cluster analysis performed better than partitioning by both gene and codon position. Alternative data partitioning schemes also are shown to affect the topologies resulting from phylogenetic analysis, especially when Bayesian methods are used, suggesting that overpartitioning may be of major concern. The phylogenetic relationships among the major clades of ray-finned fish were assessed using the best data-partitioning schemes under ML and Bayesian methods. Some significant results include the monophyly of "Holostei" (*Amia* and *Lepisosteus*), the sister-group relationships between (1) esociforms and salmoniforms and (2) osmeriforms and stomiiforms, the polyphyly of Perciformes, and a close relationship of cichlids and atherinomorphs. [Cluster analysis; data partitioning; Holostei; nuclear loci; phylogenetics; ray-finned fish; Actinopterygii.]

Phylogenomic approaches in systematics based on the analysis of multilocus sequence data are becoming increasingly common. Large numbers of characters and independent evidence from many genetic loci often result in well-resolved and highly supported phylogenetic hypotheses (e.g., Rokas et al., 2003a, 2003b, 2005; Philippe et al., 2005; McMahon and Sanderson, 2006; Baurain et al., 2007; Comas et al., 2007). In spite of this success and initial optimism about the phylogenomic approach (Gee, 2003; Rokas et al., 2003b), large and complex data sets also exacerbate many unresolved methodological challenges. As model-based phylogenetic methods gain acceptance in systematic biology, discussion of model selection strategies has shifted to a central place in the recent literature (reviewed by Sullivan and Joyce, 2005). Many long-standing challenges such as sparse taxon-sampling (Soltis et al., 2004), base compositional bias (Phillips et al., 2004; Collins et al., 2005), missing data (Wiens, 2003; Waddell, 2005), or incomplete lineage sorting (Kubatko and Degnan, 2007) also increase in relevance as multilocus data sets grow in size and complexity.

One important challenge concerns heterogeneity in evolutionary rates among genes and nucleotide sites (Bull et al., 1993; Buckley et al., 2001). An increasingly common approach to address this heterogeneity involves the definition of relatively homogeneous data blocks and subsequent optimization of independent (unlinked) models for each block. Several methods are available to choose the optimal model for any particular data set (or data block), based on testing criteria such as likelihood-ratio tests (Sullivan et al., 1997; Posada and Crandall, 1998), the AIC (Akaike information criterion; Posada and Crandall, 2001), BF (Bayes factors; Nylander et al., 2004), the BIC (Bayesian information criterion; Schwarz, 1978), and performance-based criteria (Minin et al., 2003). The

relative merits of these alternative approaches have been recently reviewed by Posada and Buckley (2004) and Sullivan and Joyce (2005). For heterogeneous data sets such as multilocus protein-coding sequences, phylogenetic analyses are increasingly based on partitioning the data by gene and/or codon position (Reed and Sperling, 1999; Nylander et al., 2004). Simulation and empirical studies have demonstrated the benefits of this approach by significantly improving overall likelihood scores and nodal support (Caterino et al., 2001; Pupko et al., 2002; Castoe et al., 2004; Brandley et al., 2005; Brown and Lemmon, 2007). In general, several methods are available for comparing and selecting for evolutionary models, an area of recent growth in phylogenetics. But unfortunately, the best approach to optimally define the number of homogenous data blocks in complex multilocus data sets has received substantially less attention (e.g., Brandley et al., 2005; Poux et al., 2005), and currently there is no explicit method available to heuristically search among all plausible partitioning schemes from a potentially vast array of alternatives—but see the mixture model, an alternative way to accommodate heterogeneity among sites (e.g., Pagel and Meade, 2004).

A common strategy for partitioning data is to use a priori knowledge and divide the concatenated sequences by gene, codon position, or both. This method is reasonable because it may capture most of the heterogeneity in the sequences. Many studies, indeed, reported that partitioning by both gene and codon position resulted in the best fit of the data (Caterino et al., 2001; Brandley et al., 2005). Under this approach, a multilocus data set with, for example, 10 protein-coding genes would be divided into 30 blocks, each with its own specified model. However, "overpartitioning"— dividing the data into too many blocks—will naturally

result in overparametrization (Sullivan and Joyce, 2005), because too many parameters associated with excess data blocks need to be estimated from the data. For finite data with relatively small number of characters in each data block, the degree of uncertainty in parameter estimation could seriously compromise the performance of the model (Rannala, 2002). It has been shown that both underpartitioning and overpartitioning led to erroneous estimates of bipartition posterior probabilities (BPPs) and increased the risk of phylogenetic error (Brown and Lemmon, 2007).

Reducing the number of parameters by merging all sites that exhibit similar substitution patterns into a single data block should improve the overall fit of a partitioned approach. For example, first codon positions of two genes with similar evolutionary constraints might be analyzed more efficiently with one model than with two separate models. To choose the best partitioning strategy, ideally, analyses of all possible combinations of predefined data blocks should be compared. The number of combinations, however, becomes prohibitively large and impractical to evaluate exhaustively when many genes are used. The number of ways a set of $n$ elements can be partitioned into nonempty subsets is called a Bell number, $B_n$ (Bell, 1934). Bell numbers can be generated using the recurrence relation

$$B_{n+1} = \sum_{k=0}^{n} \binom{n}{k} B_k \qquad (1)$$

The first six Bell numbers are $B_1 = 1$, $B_2 = 2$, $B_3 = 5$, $B_4 = 15$, $B_5 = 52$, and $B_6 = 203$, but for 10 and 30 partitions the number of possible combinations grows exponentially to $B_{10} = 115,975$ and $B_{30} = 8.47\text{E}+23$ (Bell numbers are implemented in Mathematica as BellB[n]; http://mathworld.wolfram.com/BellNumber.html). A data set with 20 protein-coding genes split into 60 a priori data blocks (3 codons $\times$ 20 genes) can give rise to $B_{30} = 9.77\text{E}+59$ possible partitioning schemes. Clearly, a heuristic approach is necessary to search for the optimal partitioning scheme for large multilocus data sets.

In some studies, "background information" or estimated model parameter values (e.g., evolutionary rate) for each potential data block have been used to guide the process of defining the optimal number of partitions for analysis (Mueller et al., 2004; Brandley et al., 2005; Poux et al., 2005; Castoe and Parkinson, 2006; McGuire et al., 2007; Nishihara et al., 2007). For example, using background information for protein-coding genes, the first plus second but not the third codon positions were grouped (Brandley et al., 2005). In another study (Poux et al., 2005), data blocks were originally defined by gene and codon position and each was optimized independently for its best-fitting model; subsequently, the number of partitions was reduced by grouping together original blocks when none of their model parameters differed by more than 100%. In yet another study using 2789 genes (Nishihara et al., 2007), the evolutionary rate of each gene was used as a criterion to define alternative

partitioning schemes. These few attempts present simple methods to avoid overpartitioning while effectively dealing with heterogeneity by first detecting (or inferring) similarities and then grouping similar data blocks. However, they fail to provide a systematic and objective way to explore the combinatorial space and/or to assess the potential improvement resulting from alternative partitioning schemes. In this study, we use cluster analysis (Hartigan, 1975) to guide the process of choosing the optimal partitioning scheme for multilocus protein-coding genes. Starting with all possible predefined data blocks (by both gene and codon position), our method provides a strategy to explore systematically alternative groupings of blocks and define a reasonable partitioning scheme based on explicit model selection criteria.

An alternative to data partitioning is mixture analysis (Pagel and Meade, 2004, 2005). Mixture methods can accommodate rate and pattern heterogeneity among sites by fitting a set of models with different weights to each site. It has the advantage that no a priori assumptions are made about the data. Currently, both mixture and partitioned methods are increasingly used in phylogenetic studies. The comparison between the two approaches and the evidence favoring one or the other await more thorough examination. The method we developed in this study is a proposal to solve one of the methodological obstacles in a partitioned analysis.

We apply this new approach to explore optimal partitioning schemes in a new data set of 10 protein-coding genes sampled from 52 actinopterygian fishes. Ray-finned fishes (class Actinopterygii) are the most species-rich group of vertebrates, with high diversity in morphology, ecology, behavior, and physiology (see Helfman et al., 1997). They comprise nearly 27,000 described species, currently classified into three subclasses, 44 orders, and 453 families (Nelson, 2006). Phylogenetic relationships among the major groups of ray-finned fishes are still controversial and unresolved, as are many of the proposed higher-level taxa (e.g., Greenwood et al., 1973; Lauder and Liem, 1983; Stiassny et al., 1996; Kocher and Stepien, 1997; Meyer and Zardoya, 2003; Miya et al., 2003; Cloutier and Arratia, 2004; Springer and Johnson, 2004). A small sample of current controversies may include such classic debates as the evidence for "Holostei" (Jessen, 1972; Olsen, 1984; Grande and Bemis, 1996), a group that contains gars (*Lepisosteus*) and the bowfin (*Amia*); the identity of the most early diverging lineages of living teleosts (Patterson and Rosen, 1977; Arratia, 2000; Inoue et al., 2001), or the interrelationship among ostariophysans (Fink and Fink, 1981; Dimmick and Larson, 1996; Ortí and Meyer, 1996; Saitoh et al., 2003). In addition to these classic debates, unexpected relationships among several groups of teleosts have been proposed by recent molecular studies, such as the sister relationship of osmeriforms with stomiiforms (Lopez et al., 2004) and gonorynchiforms as a basal clupeomorph (Saitoh et al., 2003), which need more rigorous testing. Most notoriously without resolution is the crown of the teleost tree. One of the major questions in ichthyology is the pattern of phylogenetic relationships among the higher

"perch-like" fishes, the order Perciformes and their relatives (e.g., Chen et al., 2003; Miya et al., 2003; Orrell and Carpenter, 2004; Dettai and Lecointre, 2005; Smith and Craig, 2007). Given the high taxonomic diversity and the lack of unambiguous morphological synapmorphies for many traditionally proposed groupings, it has been difficult to resolve higher-level phylogenies of ray-finned fish with morphological characters alone. More recently, molecular data have been used with varying degrees of success (Kocher and Stepien, 1997; Wiley et al., 2000; Chen et al., 2003; Miya et al., 2003, 2005; Lopez et al., 2004). Many of the early molecular studies used relatively short sequences from few genetic loci and/or limited taxonomic coverage. A successful strategy to collect large data sets has been to sequence complete mitochondrial genomes, assisted by well-established laboratory procedures and uncomplicated identification of orthologous genes (Curole and Kocher, 1999; Miya and Nishida, 2000). Pioneering phylogenetic results have been obtained with mitogenomic data spanning a huge taxonomic diversity and the resolution of many parts of the ray-finned fish phylogeny has been improved (Miya et al., 2001, 2003, 2005; Inoue et al., 2003; Ishiguro et al., 2003; Saitoh et al., 2003). However, the major caveat with mitogenomic data is that all mitochondrial genes represent but a single genetic locus, increasing the risk of systematic error (Curole and Kocher, 1999). In fact, many of the novel hypotheses proposed on the basis of mitogenomic evidence await scrutiny from morphological evidence and independent corroboration based on nuclear genes.

Currently, a truly comprehensive phylogenomic approach to fish phylogeny is not possible with only a handful of complete fish genomes available. Some recent studies have analyzed sequences from large chromosomal regions or a large number of genes representing relatively few actinopterygian (Amores et al., 2004; Chen et al., 2004; Chiu et al., 2004) and sarcopterygian (Noonan et al., 2004) species. These studies with a high gene-to-taxon ratio stand in contrast to the most commonly used "many taxa, few genes" approach attempting to span the diversity of actinopterygian fishes with few genetic markers (e.g., Chen et al., 2003). The strategy reported in this study stands in between. We analyze DNA sequences for 10 newly developed nuclear gene markers (Li et al., 2007) sequenced from 52 ray-finned fish taxa and four outgroups to assess their phylogenetic utility for higher-level systematics and assess some hypotheses of actinopterygian relationships. If successful, this approach could set the stage for future gene-taxon sampling schemes toward efficiently building the tree of life for ray-finned fishes.

## MATERIALS AND METHODS

### Taxon Sampling, Amplification, and Sequencing

We sampled 52 ray-finned fish taxa representing 41 of the 44 recognized orders of ray-finned fish (Nelson, 2006), except for Saccopharyngiformes, Ateleopodiformes, and Stephanoberyciformes due to the lack of viable tissue samples (see Appendix 1). Four tetrapods, *Xenopus tropicalis*, *Monodelphis domestica*, *Mus musculus*, and *Homo sapiens* were used as outgroup to root the phylogeny. Admittedly, our taxon sampling is not comprehensive enough to represent the diversity of ray-finned fishes, if nothing else because the delineation of many orders is still an open question and key taxa with unexpected affinities may be missing from the study. Nevertheless, this is the first report attempting to address phylogenetic relationships among ray-finned fishes using sequences of multiple nuclear genes at a relatively large taxonomic scale.

The nuclear gene makers chosen for this study were selected among putatively single-copy genes using a bioinformatics approach to scan available genomic data for fishes (Li et al., 2007). DNA fragments between 700 and 1000 bp were amplified and sequenced from the following genes: zic family member 1 (zic1), cardiac muscle myosin heavy chain 6 alpha (myh6), ryanodine receptor 3-like protein (RYR3), si:ch211-105n9.1-like protein (Ptr), T-box brain 1 (tbr1), ectodermal-neural cortex 1-like protein (ENC1), glycosyltransferase (Glyt), SH3 and PX domain-containing 3-like protein (SH3PX3), pleiomorphic adenoma protein-like 2 (plagl2), and brain superconserved receptor 2 (serb2) gene. Sequences of these 10 loci for the four tetrapod and two tetraodontiform species were retrieved from the ENSEMBL genome browser (www.ensembl.org; see Appendix 1). Sequences for the rest of the taxa were collected for this study (EU001863 to EU002148) or were previously reported by Li et al. (2007). Primers used for PCR and sequencing and the reaction conditions are as reported by Li et al. (2007).

### Alignment, Homology Assessment, and Quality Control

Because the sequenced 10 nuclear fragments correspond to exons of protein-coding genes, alignments were based on translated protein sequences using ClustalW (Thompson et al., 1994), implemented in MEGA3.1 (Kumar et al., 2004). After alignment, the aligned protein sequences were translated back into the original nucleotides for phylogenetic analysis.

The 10 nuclear genes were operationally defined as "single-copy" in the genomes of model organisms used for the bioinformatic analysis: zebrafish (*D. rerio*), torafugu (*T. rubripes*), stickleback (*G. aculeatus*), and medaka (*O. latipes*). This operational definition of a single-copy gene only requires that the fragment is not present as a second copy in the genome with similarity higher than 50%. Some single-copy genes may, in fact, have duplicates in the genome that are less than 50% similar (Li et al., 2007). Therefore, to test whether the sequences collected for each locus may have paralogous copies resulting from fish-specific genome duplication events (Taylor et al., 2003; Van de Peer et al., 2003), the most similar fragments, or putative "out-paralogs" (Remm et al., 2001), in the genome were download from the ENSEMBL database for zebrafish, stickleback, medaka, torafugu, and spotted-green pufferfish. These putative paralogs were aligned

with the sequences collected in the present study and neighbor-joining trees (NJ; Saitou and Nei, 1987) were constructed for each locus. "Confused paralogy" would be diagnosed by this procedure if any of the putative out-paralogs are nested among the sequences collected for ray-finned fishes in our study. If all sequences collected for this study are orthologous to each other, paralogous copies should be placed in a sister-group relationship to all sampled ray-finned fishes. This expectation assumes that the duplication events preceded the early diversification of actinopterygians or the split between tetrapods and the ray-finned fish lineages. A discrepancy between the gene trees and species tree could occur if a duplication event took place after the diversification of actinopterygians and the duplicated genes were lost in different lineages asymmetrically. In this case, recovering the species tree would require consistent and independent phylogenetic signal from many unlinked loci, such as the ones used in this study. During this preliminary analysis, sequences placed at unexpected positions in the NJ tree were identified and checked for accuracy. Quality control performed in this way also was aimed to minimize laboratory mistakes (e.g., contamination) in addition to problems that may arise due to confused paralogy.

### Data Partitioning, Parameter Estimation, and Cluster Analysis

As a first step, the concatenated data matrix of 10 nuclear gene sequences was partitioned by gene and by codon position, producing 30 blocks of data. A smaller number of data blocks, however, may be sufficient because some of them are likely to exhibit similar evolutionary properties. These properties were assessed by phylogenetic analysis (see below) using the ML method implemented in TreeFinder (Jobb, 2006) and a Bayesian approach implemented in MrBayes (Nylander et al., 2004). For these analyses, each of the 30 data blocks was optimized independently under a GTR+Γ model. Overall similarity among data blocks was evaluated on the basis of their estimated parameter values, counting five substitution rates, three base composition proportions, the gamma parameter (alpha), and the rate multiplier for each. We did not include the invariable parameter (theta) in the models, because the alpha and theta estimated under I+Γ model might be highly correlated (Sullivan et al., 1999). Hierarchical cluster analysis was used to analyze the level of similarity among data blocks, using the parameter values as input for PROC CLUSTER with the centroid method in SAS. We choose centroid as the amalgamation approach because this method is more robust to outliers than most other hierarchical methods (Milligan, 1980). Separate cluster analyses were run for parameters estimated by ML and Bayesian approaches. The resulting hierarchical clustering graphs were used to guide the grouping process to propose increasingly smaller numbers of data partitions. Starting with all 30 blocks at the tips of the cluster dendrogram, this method identifies and groups the two most similar, yielding 29 data partitions; it subsequently identifies and groups the

next most-similar data blocks to yield 28 partitions, and so on down the guide-tree to assemble progressively larger clusters of data blocks with high similarity, continuing down to the root until single group is defined. Although this approach is not an exhaustive exploration of all possible partitioning schemes, it prescribes a reasonable and explicit strategy to group data blocks on the basis of their overall similarity in evolutionary parameters. To test the performance of our clustering approach, we also carried out an exhaustive search of all possible combinations for a subset of our data (6 data blocks = 2 genes × 3 codon positions). The partitioning schemes determined by cluster analysis (6 alternative schemes, from 1 to 6 partitions) were compared to all possible combinations of 6 blocks into 1, 2, 3, 4, 5, or 6 partitions ($B_6 = 203$ different combinations). An alternative nonhierarchical clustering approach (k-means; Hartigan and Wong, 1979) was compared to the hierachical method implemented in this study. k-means clustering was implemented using CLUSTER (Hoon, 2002). Although it does not guarantee to return a global optimum, the k-means method is faster and could be useful for larger data sets, when the hierarchical approach would be computationally slow and tedious to implement.

The effects on phylogenetic estimation of all the resulting partitioning schemes, from 30 to a single block, were evaluated using several decision criteria. The procedures for phylogenetic analysis using each partitioning scheme (each scheme defines a model for the analysis) are explained in the next section. For the ML-based inference, the following two criteria were applied: (1) the Akaike information criterion (AIC) was calculated as $AIC_i = -2\ln L_i + 2k_i$, where $L_i$ is the maximum likelihood of the model and $k_i$ is the number of parameters in the model $i$. When the ratio of the number of nucleotides to the number of parameters $n/k_i \leq 40$, $AIC_C$ is used instead of AIC to correct for small sample size (Burnham and Anderson, 2002). The $AIC_C$ was computed by $AIC_{Ci} = -2\ln L_i + 2k_i + 2k_i (k_i + 1)/(n - k_i - 1)$. As a deciding rule, $\Delta_i = AIC_i - AIC_{min} \leq 10$ was used as an indication of nonsignificant difference between model $i$ and the best model (Burnham and Anderson, 2002), and model $i$ was preferred if it was based on a smaller number of data blocks (i.e., a smaller number of parameters). (2) The Bayesian information criterion (BIC) that penalizes free parameters more strongly than the AIC was calculated as: $BIC_i = -2\ln L_i + k_i \ln n$.

To evaluate partitioning schemes using estimates based on Bayesian analyses, the BIC and Bayes factor (BF) were used to compare models. The BF was calculated as the difference of harmonic means of likelihoods between models compared (Nylander et al., 2004). A recent study has shown that BF provides a statistically sound measure for evaluating alternative partitioning schemes (Brown and Lemmon, 2007).

In addition to evaluating the fit of the data to different partitioning models by all the above criteria, the effects of different partitioning schemes on the inferred topology also were examined. In addition to listing alternative topologies in the appendix, the meta-tree method

by Nye (2008) based on tree-to-tree distances was used to visualize differences among results, treating each phylogeny as a node, and using a tree to describe the relationship among them. The method is based on minimization of the total Robinson-Foulds distance (Robinson and Foulds, 1981). Another important issue related to phylogenetic analyses under alternative partitioning schemes concerns nodal-support values (Buckley et al., 2001). Based on results of Bayesian analysis, the effects of data partitioning on the bipartition posterior probabilities (BPPs) were examined by plotting BPP values for common nodes resulting from over- and underpartitioning relative to the preferred partitioning scheme.

### Phylogenetic Analysis

The basic summary information for each locus, such as the number of parsimony informative sites, average genetic p-distance among taxa, and consistency index were calculated using PAUP (Swofford, 2003). All data-partitioning schemes were tested using both ML and Bayesian methods. Bayesian phylogenetic analyses implemented in MrBayes v3.1.1 and ML analyses in TreeFinder (Jobb, 2006) were performed on the nucleotide sequences. For the first set of analyses, designed to assess the similarity among the initial 30 data blocks and to compare the alternative data partitioning schemes, unlinked GTR+$\Gamma$ models were used for all data blocks, and the model parameters were estimated independently for each. The GTR+$\Gamma$ model was used for all data blocks to allow direct comparison of results from different partitioning schemes and also because the GTR model is the most commonly used model reported in the literature (Kelchner and Thomas, 2007). The GTR model implemented in Treefinder or MrBayes has eight free parameters (see manual of Treefinder and MrBayes), so each data block adds 10 parameters to the overall model—8 parameters for the GTR model, 1 parameter for $\Gamma$, and 1 rate multiplier, used to accommodate the overall rate difference among partitions. The best partitioning scheme was chosen according to the decision criteria outlined above.

A second set of analyses was performed after the optimal partitioning scheme was chosen. In this case, ModelTest (Posada and Crandall, 1998) and PAUP* were applied on each data block separately to select the best model (under AIC), rather than arbitrarily fitting each block with a GTR+$\Gamma$ model. If this procedure finds simpler models than the GTR+$\Gamma$, additional savings in the number of parameters is possible. PAUP* was used to obtain the score for each model proposed by the ModelTest block on the tree topology estimated from the best partitioning scheme with the GTR+$\Gamma$ model for all data blocks (as described above). However, because both TreeFinder and MrBayes only implement a subset of models, the closer but more parameter-rich model to that suggested by ModelTest available in TreeFinder or MrBayes was used for the analyses. For all Bayesian analyses, MCMC was run for 3 million generations with four chains, with tree-sampling frequency of 1 in 100 (30,000 trees saved). The last 5000 trees sampled were used to compute the consensus tree and the posterior probabilities. Two independent runs were used to provide additional confirmation of convergence of posterior probability distribution. All Bayesian analyses were run for the same number of generations (3 million), to allow direct comparison of the convergence rate for different partitioning schemes, as indicated by the average standard deviation of split frequencies among two MCMC runs, printed by MrBayes at the end of each run. For the ML analysis, 200 bootstrap replicates were carried out for the best partitioning scheme. Alternative hypotheses were assessed by one-tailed Shimodaira and Hasegawa (SH) tests (Shimodaira and Hasegawa, 1999) with 1000 RELL bootstrap replicates, implemented in TreeFinder.

### RESULTS

#### Characteristics of the Ten Nuclear Loci Amplified in Ray-Finned Fishes

The aligned sequences concatenating all 10 loci produced 7995 nucleotides for each taxon. The complete alignment is available on the TreeBASE website (study accession number = S2044, Matrix accession number = M3827). Gaps resulting from the alignment were treated under the default setting in MrBayes and TreeFinder. Some gene sequences were excluded from further analysis due to poor sequencing quality (deficient amplification and/or incomplete sequence data), resulting in a data matrix with about 16% missing data (see Appendix 1). The summary information for each locus is listed in Table 1.

Preliminary NJ analyses for each individual gene fragment, plus additional sequences of varying degrees of similarity downloaded from the ENSEMBL database for zebrafish, stickleback, medaka, torafugu, and spotted green pufferfish, were performed to detect putative "out-paralogs." Resulting NJ trees showed that the putative paralogs detected in the databases were all positioned as either a sister group of the other actinopterygian sequences or that they joined at the root of the tree as a polytomy (results not shown), supporting the assumption that all the sequences directly collected for this study are orthologous for each locus.

#### Comparison among Alternative Partitioning Schemes and Models

The maximum number of data blocks defined a priori for the concatenated data set was 30 (3 codon positions × 10 genes). For each block, parameter values for the GTR+$\Gamma$ model estimated using both ML and Bayesian approaches are shown in Appendices 2 and 3. These values were used as input for cluster analysis to obtain a branching pattern reflecting overall similarity among data blocks (Fig. 1). Clustering diagrams obtained based on ML and Bayesian estimation of parameter values are similar, except for minor differences within the major clusters (Fig. 1). For both cases, the three major clusters correspond to codon position of the genes, suggesting a major effect of this a priori factor in overall similarity among data blocks. The importance of individual model parameters affecting the clustering

TABLE 1.   Characteristics of the 10 nuclear loci amplified in ray-finned fishes. PI: parsimony-informative sites; CI-MP: consistency index on the maximum parsimony tree.

| Gene | No. of bp | No. of variable sites | No. of PI sites | Average p-distance (Min–Max) | CI-MP | No. of taxa (out of 56) |
|------|-----------|----------------------|-----------------|------------------------------|-------|-------------------------|
| zic1 | 927 | 395 | 345 | 0.158 (0.010–0.267) | 0.232 | 54 |
| myh6 | 735 | 369 | 325 | 0.174 (0.034–0.300) | 0.232 | 48 |
| RYR3 | 834 | 497 | 425 | 0.215 (0.039–0.338) | 0.280 | 41 |
| Ptr | 705 | 426 | 375 | 0.206 (0.023–0.352) | 0.272 | 51 |
| tbr1 | 720 | 410 | 328 | 0.196 (0.021–0.337) | 0.367 | 42 |
| ENC1 | 810 | 405 | 359 | 0.180 (0.029–0.283) | 0.242 | 50 |
| Glyt | 888 | 589 | 509 | 0.215 (0.029–0.364) | 0.291 | 44 |
| SH3PX3 | 705 | 373 | 319 | 0.168 (0.051–0.285) | 0.270 | 45 |
| plagl2 | 684 | 410 | 344 | 0.179 (0.012–0.372) | 0.316 | 44 |
| sreb2 | 987 | 431 | 387 | 0.149 (0.015–0.273) | 0.254 | 51 |

structure can be gauged by their RSQ/(1 − RSQ) value (ratio of between-cluster variance to within-cluster variance). The parameters with greatest effect, ranked according to this criterion (and their respective values), are the gamma-function parameter alpha (195.2), the rate multiplier (48.5), and the C-G (15.2), C-T (7.9), and A-T (7.3) substitution rates.

The alternative groupings of data blocks prescribed by the clustering analysis, from a single data block (no partitions) to the maximum of 30 data blocks, that were used in subsequent analyses are shown in Table 2. All partitioning schemes, as well as the frequently used approach of partitioning only by gene only, were assessed for their effects on phylogenetic inference under different testing criteria to determine the best partitioning scheme. For parameter values estimated under ML, the likelihood and AIC scores improved consistently with increasing number of data blocks (Table 3), with the best values obtained by the model with 30 blocks (299 parameters). Similarly, BIC values also improved with increasing number of data blocks (Table 3 and Fig. 2), but interestingly the best BIC score corresponded to a partitioned model based on 10 data blocks (99 parameters). The 10-block partitioning scheme was chosen as the optimal model in this study according to its BIC value, but the result from the 30-block partitioning model selected by the AIC also is reported (Table 3 and Fig. 2).

ModelTest and PAUP* were used to determine the best model for ML analysis for each of the 10 data blocks in the optimal partitioning scheme. The models with the best AIC values and the models used in ML analysis are listed in Table 4. The models implemented in TreeFinder that were closer to the models suggested by ModelTest (under AIC) actually had eight fewer parameters, and the final likelihood and BIC scores were worse than simply applying a GTR+Γ for all data blocks (Table 4). Therefore, GTR+Γ was used for the final analysis to construct the phylogeny. The tree topologies obtained under the two models are, however, the same (topology C, see below).

The results of the Bayesian analysis to test the effects on phylogenetic inference of alternative partitioning schemes were similar to those presented above for ML (Table 3, Fig. 3). For the Bayesian approach, the best BIC value was obtained by grouping the data into 17 blocks, whereas the best partitioning scheme according to Bayes factor is 22 blocks (Table 3, Fig. 3). BIC has been used under Bayesian context for choosing optimal partitioning schemes (McGuire et al., 2007). The 17-block scheme was chosen as the preferred model but the result of the 22-block model is reported as well (Table 3, Fig. 3, Supplementary Materials, available online at www.systematicbiology.org). The models selected using ModelTest (under AIC) for each of these 17 data blocks and the closest models available in MrBayes are listed in Table 5. The 17-block partitioning model implemented in the Bayesian analysis had 14 parameters less (total 155) than the model using GTR+Γ for all blocks (total 169 parameters). The likelihood and BIC scores were best under the simpler models selected by ModelTest (Table 5), so these were used for the final Bayesian phylogenetic inference. Another interesting result of Bayesian analysis concerns the rate of convergence of posterior distributions obtained for the different partitioning schemes. Partitioning the data into higher number of blocks resulted in slower convergence of two MrBayes runs, as indicated by the average standard deviation of split frequencies (ASDSF) after 3 million MCMC generations (Table 3).

Finally, the frequently used partitioning scheme by gene only (10 blocks in this case) resulted in significantly worse likelihood and BIC scores for both ML and Bayesian analysis (Table 3, Figs. 2 and 3).

*Effects of Alternative Partitioning Schemes and Models on Tree Topology*

Overall, phylogenetic analyses based on 30 alternative partitioning schemes under ML and Bayesian approaches resulted in 23 different tree topologies (trees were labeled A to W; see Supplementary Materials). For each data-partitioning scheme, the tree topology obtained is shown in Figures 2 and 3 for ML and Bayesian analyses, respectively. ML analyses resulted in only six different topologies (A to F), the most frequently obtained topology was C (in 26 out of 31 cases), and this topology also resulted from analysis of the optimal partitioning scheme with substitution models selected by ModelTest using AIC. Interestingly, topology C was a stable outcome with partitioning schemes involving 18 or more data blocks; thus, overparametrization in this case had little effect on the resulting tree. Therefore, topology C is considered the best hypothesis under the ML criterion.
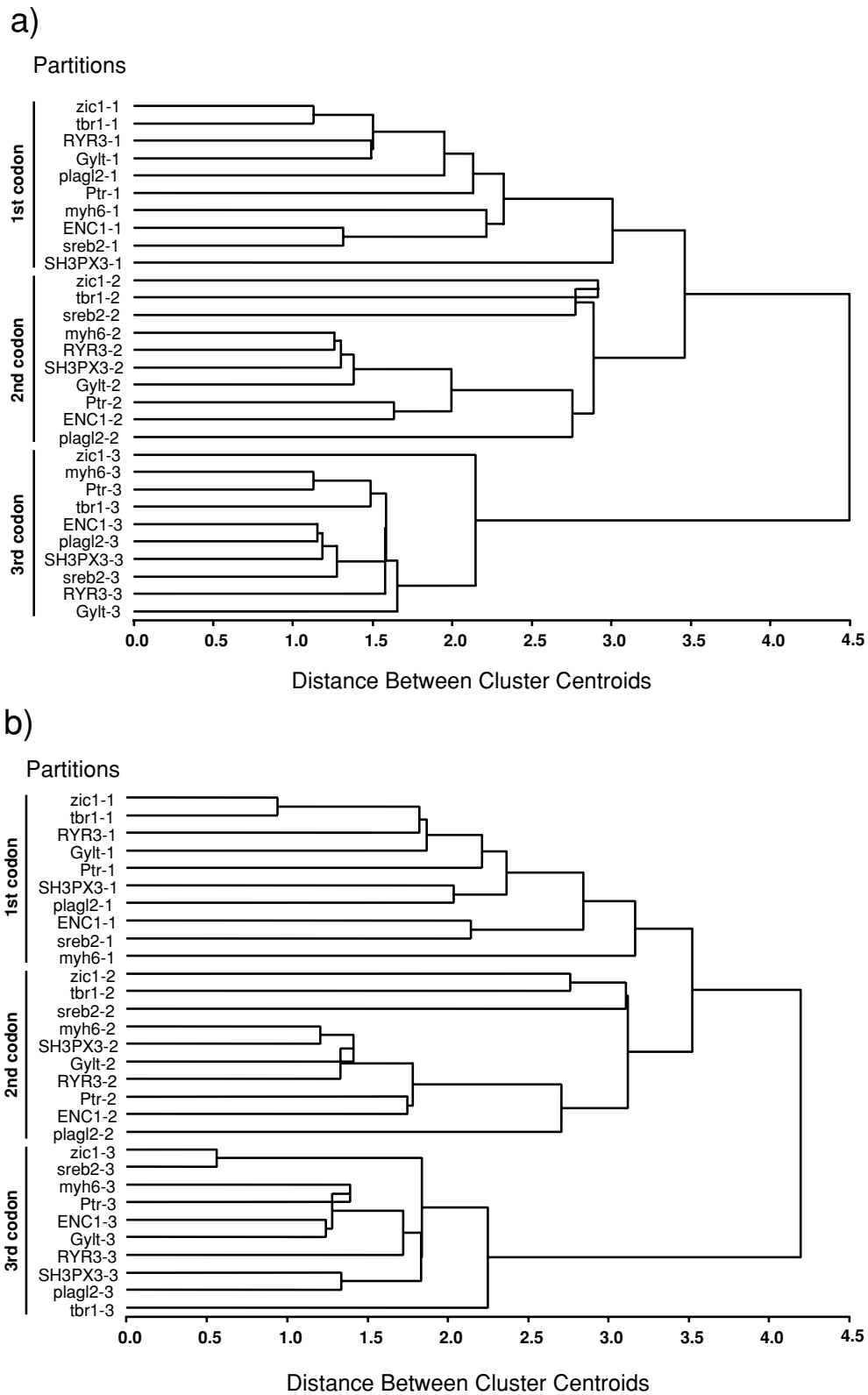
FIGURE 1. Clustering diagrams showing overall similarity among 30 data blocks of the full data set (10 genes × 3 codon positions). Each block is indicated at the tip of terminal branches by gene name (see Table 1 and Materials and Methods) and codon position. The shifting of some nodes (e.g., the node joining zic1-2 and tbr1-2) is the result of centroid method. (a) Cluster analysis of model parameters estimated using ML. (b) Cluster analysis of model parameters estimated using a Bayesian approach.

TABLE 2. Alternative partitioning schemes (from 30 to 1 partitions) suggested by cluster analysis on model parameters of the original 30 partitions. CL(number) indicates the composition of partitions grouped at each step, indicated by progressing from the tips to the root of the clustering diagrams shown in Figure 1 (e.g., CL 27 in the second column is composited of ENC1-3 plagl2-3).

| | Partitions | |
|---|---|---|
| Number of clusters | Joined based on ML-estimated parameters (Fig. 1a) | Joined based on Bayesian-estimated parameters (Fig. 1b) |
| 29 | myh6-3, Ptr-3 | zic1-3, sreb2-3 |
| 28 | zic1-1, tbr1-1 | zic1-1, tbr1-1 |
| 27 | ENC1-3, plagl2-3 | myh6-2, SH3PX3-2 |
| 26 | CL27, SH3PX3-3 | ENC1-3, Gylt-3 |
| 25 | myh6-2, RYR3-2 | SH3PX3-3, plagl2-3 |
| 24 | CL26, sreb2-3 | myh6-3, Ptr-3 |
| 23 | CL25, SH3PX3-2 | CL24, CL26 |
| 22 | ENC1-1, sreb2-1 | CL27, Gylt-2 |
| 21 | CL23, Gylt-2 | CL22, RYR3-2 |
| 20 | CL29, tbr1-3 | CL23, RYR3-3 |
| 19 | RYR3-1, Gylt-1 | Ptr-2, ENC1-2 |
| 18 | CL28, CL19 | CL21, CL19 |
| 17 | CL20, CL24 | CL28, RYR3-1 |
| 16 | CL17, RYR3-3 | CL20, CL25 |
| 15 | Ptr-2, ENC1-2 | CL29, CL16 |
| 14 | CL16, Gylt-3 | CL17, Gylt-1 |
| 13 | CL18, plagl2-1 | SH3PX3-1, plagl2-1 |
| 12 | CL21, CL15 | ENC1-1, sreb2-1 |
| 11 | CL13, Ptr-1 | CL14, Ptr-1 |
| 10 | zic1-3, CL14 | CL15, tbr1-3 |
| 9 | myh6-1, CL22 | CL11, CL13 |
| 8 | CL11, CL9 | CL18, plagl2-2 |
| 7 | CL12, plagl2-2 | zic1-2, tbr1-2 |
| 6 | zic1-2, tbr1-2 | CL9, CL12 |
| 5 | CL6, sreb2-2 | CL7, sreb2-2 |
| 4 | CL5, CL7 | CL5, CL8 |
| 3 | CL8, SH3PX3-1 | CL6, myh6-1 |
| 2 | CL3, CL4 | CL3, CL4 |
| 1 | CL2, CL10 | CL2, CL10 |

The effect of data partitioning on tree topology using Bayesian analysis was more pronounced, resulting in 17 alternative topologies (G to W). Topology I was the most frequent result (11 out of 31 cases). The preferred topology under the optimal-partitioning scheme and models selected by ModelTest (using AIC) was topology L (Fig. 3). In contrast to ML analysis, partitioning schemes with more than 16 blocks had highly heterogeneous outcomes (11 different topologies, L to V), suggesting a larger effect of overparametrization on the resulting posterior distribution of tree topology. Under the Bayesian approach, topology L is considered the best hypothesis. A strict consensus between topologies L and C is shown in Figure 4.

Tree-to-tree distances among alternative topologies (A to W) are presented using the meta-tree method (Fig. 5). The six topologies resulting from ML analysis (shown as circles A to F in Fig. 5) were split into two distinct groups. Topologies A and F (resulting from no partitioning and partitioning by genes alone, respectively) had the lowest BIC values (Fig. 2, Table 3) and are quite different from the rest (B to E), suggesting that partitioning by genes alone has a very similar effect on the topology and likelihood value has no partitioning at all. A similar pattern is

observed for the topologies obtained by Bayesian analyses (shown as squares G to W in Fig. 5). The topology obtained when no partitioning was assumed (G) also is close to that obtained by partitioning by gene alone (W). Interestingly, many topologies produced by our analyses located in the internal nodes of the meta-tree, suggesting high degree of congruence among them (all tree topologies are available at Supplementary Materials).

To check the effects of different partitioning schemes on nodal support, we plotted the BPPs resulting from over- and underpartitioning against the preferred partitioning schemes. Two extreme strategies (no partitioning and 30 partitions) were compared with the optimal 17 partitions selected by our approach (Fig. 6). Either under- or overpartitioning decreased the correlation among BPPs relative to values obtained with the preferred scheme (Fig. 6). The change of BPPs was found more severe in overpartitioned analysis ($r = 0.0083$) than underpartitioned analysis ($r = 0.0176$), which is similar to the results of a recent simulation study (Brown and Lemmon, 2007). In contrast to that simulation study, however, our results showed that there were 91 bipartitions supported by a PP = 1.0 among 112 bipartition patterns shared by both the underpartitioning and the preferred-partitioning schemes. Similarly, 91 bipartitions with PP = 1.0 were found among the 110 bipartition patterns shared by both the overpartitioning and the preferred scheme. Therefore, the different partitioning schemes had a minor effect on the BPPs, because most estimates with high support (BPP = 1) were unchanged in this study.

*Evaluation of Clustering Analysis as a Heuristic Approach for Data Partitioning*

For a subset of the data (two genes), the six partitioning schemes chosen by cluster analysis were compared with all 203 possible ways of partitioning the data (one to six partitions, $B_6 = 203$). Bayesian analyses using all different partitioning schemes were implemented and the $-\ln L$ values of the resulting trees are presented in Figure 7. In all cases, the partitioning schemes selected by our hierarchical clustering approach (shown as the squares in Fig. 7) have the best likelihood, except for the four-class partitioning where a random combination of data blocks outperformed the one selected by clustering analysis but the difference was not significant (Bayes factor < 5). The k-means methods (shown as the circles in Fig. 7) also resulted in best partitioning schemes in most of the cases, except that it was outranked by four other partitioning schemes in the four-class partitioning case (Fig. 7).

The thorough comparison using a subset of the data suggests that the heuristic data-partitioning approach based on cluster analysis should be useful when analyzing multilocus data. The k-means clustering method often is faster than the hierarchical methods, but it does not guarantee to return a global optimum (Hartigan and Wong, 1979), which may explain the suboptimal solution chosen by the k-means method in one of the cases reported.

TABLE 3. Comparison of log likelihoods, AIC, BIC, and Bayes factors among different partitioning schemes (from 1 to 30 partitions). For each type of analysis (ML or Bayesian), the following results are shown: total number of parameters; log likelihood calculated using TreeFinder ($L_{ML}$); uncorrected AIC values (when $n/k > 40$; n = the number of sites and $k$ = number of parameters) or corrected $AIC_C$ (when $n/k \leq 40$, only necessary when the number of partitions $\geq 20$); the difference in AIC values among model i and the best model ($\Delta_i = AIC_i - AIC_{min}$); the harmonic mean of $-\log$ likelihood calculated using MrBayes ($L_{BA}$); the Bayes factor calculated by comparing model $i$ to the model with maximum likelihood, BF $= (-\ln L_i) - (-\ln L_{best})$; and average standard deviation of split frequencies of two independent runs of MrBayes. Boxed text indicates the best partitioning schemes chosen by different model selection criteria.

| Number of partitions | No. of parameters | Maximum likelihood | | | | Bayesian analysis | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | $L_{ML}$ | AIC or $AIC_C$ | $\Delta_i$ | $BIC_{ML}$ | $L_{BA}$ | Bayes factor | $BIC_{BA}$ | Split deviation |
| 1 | 9 | 130,936 | 261,890 | 10,167 | 261,953 | 131,050 | 5193 | 262,180 | 0.005943 |
| 2 | 19 | 127,075 | 254,188 | 2465 | 254,321 | 127,095 | 1238 | 254,361 | 0.004624 |
| 3 | 29 | 126,686 | 253,431 | 1708 | 253,633 | 126,720 | 863 | 253,701 | 0.007499 |
| 4 | 39 | 126,654 | 253,387 | 1664 | 253,659 | 126,694 | 837 | 253,739 | 0.005629 |
| 5 | 49 | 126,484 | 253,066 | 1343 | 253,408 | 126,542 | 685 | 253,525 | 0.006435 |
| 6 | 59 | 126,421 | 252,961 | 1238 | 253,373 | 126,474 | 617 | 253,478 | 0.008284 |
| 7 | 69 | 126,373 | 252,885 | 1162 | 253,367 | 126,364 | 507 | 253,349 | 0.008371 |
| 8 | 79 | 126,324 | 252,806 | 1083 | 253,358 | 126,327 | 470 | 253,364 | 0.008377 |
| 9 | 89 | 126,237 | 252,652 | 929 | 253,273 | 126,282 | 425 | 253,363 | 0.009426 |
| 10 | 99 | 126,190 | 252,579 | 856 | 253,270 | 126,261 | 404 | 253,412 | 0.010901 |
| 11 | 109 | 126,160 | 252,538 | 815 | 253,299 | 126,178 | 321 | 253,335 | 0.008561 |
| 12 | 119 | 126,119 | 252,475 | 752 | 253,307 | 126,136 | 279 | 253,342 | 0.014122 |
| 13 | 129 | 126,068 | 252,393 | 670 | 253,294 | 126,126 | 269 | 253,412 | 0.008394 |
| 14 | 139 | 126,038 | 252,353 | 630 | 253,325 | 126,114 | 257 | 253,477 | 0.015416 |
| 15 | 149 | 125,988 | 252,275 | 552 | 253,316 | 126,086 | 229 | 253,511 | 0.016578 |
| 16 | 159 | 125,966 | 252,249 | 526 | 253,360 | 125,947 | 90 | 253,324 | 0.015155 |
| 17 | 169 | 125,913 | 252,165 | 442 | 253,345 | 125,857 | 0 | 253,232 | 0.031614 |
| 18 | 179 | 125,861 | 252,079 | 356 | 253,330 | 125,907 | 50 | 253,423 | 0.020992 |
| 19 | 189 | 125,829 | 252,036 | 313 | 253,356 | 125,881 | 24 | 253,461 | 0.028444 |
| 20 | 199 | 125,816 | 252,041 | 318 | 253,421 | 125,865 | 8 | 253,517 | 0.039061 |
| 21 | 209 | 125,718 | 251,866 | 143 | 253,315 | 125,921 | 64 | 253,720 | 0.025118 |
| 22 | 219 | 125,703 | 251,856 | 133 | 253,374 | 125,840 | −17 | 253,649 | 0.035717 |
| 23 | 229 | 125,691 | 251,854 | 131 | 253,441 | 125,893 | 36 | 253,843 | 0.023924 |
| 24 | 239 | 125,678 | 251,849 | 126 | 253,504 | 125,885 | 28 | 253,918 | 0.048132 |
| 25 | 249 | 125,650 | 251,814 | 91 | 253,537 | 125,935 | 78 | 254,107 | 0.034249 |
| 26 | 259 | 125,630 | 251,795 | 72 | 253,587 | 125,903 | 46 | 254,133 | 0.035437 |
| 27 | 269 | 125,607 | 251,771 | 48 | 253,632 | 125,897 | 40 | 254,212 | 0.096736 |
| 28 | 279 | 125,600 | 251,779 | 56 | 253,708 | 125,897 | 40 | 254,302 | 0.064801 |
| 29 | 289 | 125,569 | 251,738 | 15 | 253,736 | 126,032 | 175 | 254,662 | 0.051778 |
| 30 | 299 | 125,551 | 251,723 | 0 | 253,788 | 125,937 | 80 | 254,560 | 0.132187 |
| 10 (by gene) | 99 | 130,509 | 261,216 | 9493 | 261,908 | 130,570 | 4713 | 262,030 | 0.021610 |

## Phylogenetic Relationships among Ray-Finned Fishes

The topology shown in Figure 4 summarizes the preferred hypothesis of relationships based on the 10-gene data set analyzed in this study. Many traditionally recognized high-order taxa (Holostei, Teleostei, Clupeocephala, Euteleostei, Neoteleostei, and Acanthopterygii) were supported by the data and received high bootstrap and posterior probability values. Other groups, such as Paracanthopterygii, Protacanthopterygii, and the order Perciformes were not supported as traditionally recognized.

A few a priori alternative hypotheses about the branching pattern near the base of the actinopterygian tree were tested by ML (Table 6). The SH test rejected hypotheses placing either *Amia* (Patterson, 1973; Grande and Bemis, 1996) or *Lepisosteus* (Olsen, 1984) as the sister taxon to Teleostei. The previously proposed grouping of *Amia* and *Lepisosteus* (Nelson, 1969; Jessen, 1972) and of *Polypterus* and *Polyodon* (Schaeffer, 1973; Nelson, 1994) could not be rejected. The position of *Polypterus* as the sister group to all other ray-finned fishes, however, could not be rejected by the SH test. A group of "ancient fishes" (Acipenseriformes plus Holostei) suggested on the basis of mitoge-

nomic data (Inoue et al., 2003) and indel distribution in RAG genes (Venkatesh et al., 2001) also could not be rejected by the SH test. Other significant results concerning fish phylogenies are addressed in Discussion.

## DISCUSSION

### Optimal Data Partitioning

Much effort has been devoted to select the best-fitting substitution model for maximum-likelihood or Bayesian analysis (reviewed by Posada and Buckley, 2004; Sullivan and Joyce, 2005) because it has long been known that model selection can have a significant effect on phylogenetic inference (Sullivan et al., 1997; Cunningham et al., 1998; Kelsey et al., 1999). As large and complex data sets with heterogeneous sequence data became more common, partitioned analyses were developed to accommodate this heterogeneity (Yang, 1996; Posada and Crandall, 2001; Nylander et al., 2004). Individually fitted parameter-rich models can be applied to predefined types of data blocks and combined into a "supermodel" that can be analyzed with Bayesian (Nylander et al., 2004) or ML (Jobb, 2006) approaches. The issue of
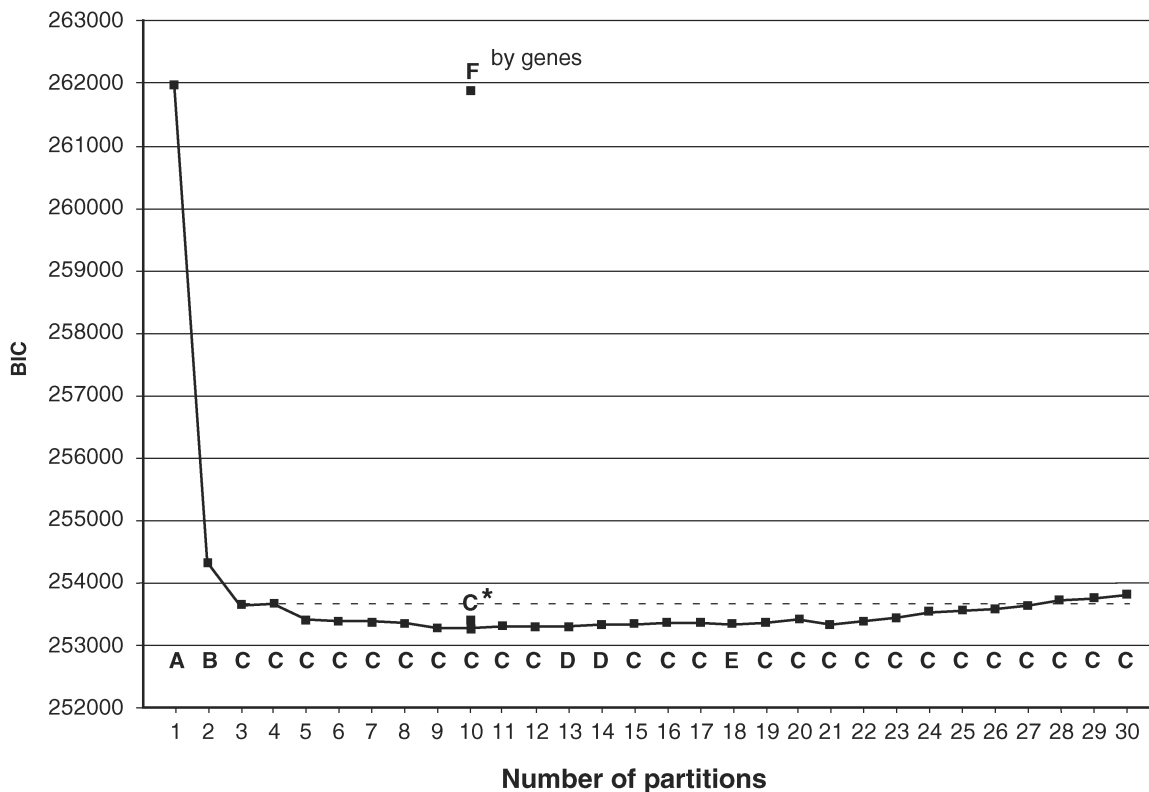
FIGURE 2.   BIC (Bayesian information criterion) values for maximum likelihood (ML) analyses under different data partitioning schemes. Letters A to F indicate alternative best tree topologies supported by different partitioning schemes; these topologies are shown in Supplementary Materials.

defining the optimal partitioning strategy, however, has not been addressed in a systematic way (but see McGuire et al., 2007). A recent study has shown that both over- and underpartitioning schemes may increase the risk of phy-logenetic error (Brown and Lemmon, 2007), but they did not provide an efficient solution to explore all possible partitioning schemes. Our study is the first to propose a heuristic approach to systematically define and evaluate

TABLE 4.   Models selected by ModelTest under the AIC criterion for the optimal 10-partition scheme for ML analysis. These models were implemented in TreeFinder and compared to using GTR+Γ for all partitions.

| Partition | Data included | Model chosen by ModelTest | Model implemented in TreeFinder | No. of parameters | No. of parameters for GTR+Γ |
|---|---|---|---|---|---|
| 1 | zic1-1, RYR3-1, ptr-1, tbr1-1, Glyt-1, plagl2-1 | GTR+I+G | GTR+I+G | 10 | 9 |
| 2 | zic1-2 | TVM+G | GTR+G | 9 | 9 |
| 3 | zic1-3, myh6-3, RYR3-3, ptr-3, tbr1-3, ENC1-3, Glyt-3, SH3PX3-3, plagl2-3, sreb2-3 | TVM+I+G | GTR+I+G | 10 | 9 |
| 4 | myh6-1 | TIM+I+G | GTR+I+G | 10 | 9 |
| 5 | myh6-2, RYR3-2, ptr-2, ENC1-2, Glyt-2, SH3PX3-2 | GTR+I+G | GTR+I+G | 10 | 9 |
| 6 | tbr1-2 | F81+I+G | HKY+I+G | 6 | 9 |
| 7 | ENC1-1, sreb2-1 | GTR+I+G | GTR+I+G | 10 | 9 |
| 8 | SH3PX3-1 | F81+I+G | HKY+I+G | 6 | 9 |
| 9 | plagl2-2 | F81+G | HKY+G | 5 | 9 |
| 10 | sreb2-2 | F81+I+G | HKY+I+G | 6 | 9 |
| No. of multipliers for relative rates | | | | 9 | 9 |
| Total no. of parameters | | | | 91 | 99 |
| ln Likelihood | | | | −126,287 | −126,190 |
| BIC | | | | 253,391 | 253,270 |

TABLE 5. Models selected by ModelTest under the AIC criterion for the optimal 17-partition scheme for Bayesian analysis. These models were implemented in MrBayes and compared to using GTR+Γ for all partitions.

| Partition | Data included | Model chosen by ModelTest | Model implemented in MrBayes | No. of parameters | No. of parameters for GTR+Γ |
|---|---|---|---|---|---|
| 1 | zic1-1, tbr1-1, RYR3-1 | TVM+I+G | GTR+I+G | 10 | 9 |
| 2 | zic1-2 | TVM+G | GTR+G | 9 | 9 |
| 3 | zic1-3, sreb2-3 | TVM+I+G | GTR+I+G | 10 | 9 |
| 4 | myh6-1 | TIM+I+G | GTR+I+G | 10 | 9 |
| 5 | myh6-2, SH3PX3-2, Glyt-2, RYR3-2, ptr-2, ENC1-2 | GTR+I+G | GTR+I+G | 10 | 9 |
| 6 | myh6-3, ptr-3, ENC1-3, Glyt-3, RYR3-3 | TVM+I+G | GTR+I+G | 10 | 9 |
| 7 | ptr-1 | GTR+G | GTR+G | 9 | 9 |
| 8 | tbr1-2 | F81+I+G | F81+I+G | 5 | 9 |
| 9 | tbr1-3 | TrN+G | GTR+G | 9 | 9 |
| 10 | ENC1-1 | SYM+I+G | GTR+I+G | 10 | 9 |
| 11 | Glyt-1 | HKY+G | HKY+G | 5 | 9 |
| 12 | SH3PX3-1 | F81+I+G | F81+I+G | 5 | 9 |
| 13 | SH3PX3-3, plagl2-3 | TVM+G | GTR+G | 9 | 9 |
| 14 | plagl2-1 | TVM+G | GTR+G | 9 | 9 |
| 15 | plagl2-2 | F81+G | F81+G | 4 | 9 |
| 16 | sreb2-1 | GTR+I+G | GTR+I+G | 10 | 9 |
| 17 | sreb2-2 | F81+I+G | F81+I+G | 5 | 9 |
| | No. of multipliers for relative rates | | | 16 | 16 |
| | Total no. of parameters | | | 155 | 169 |
| | ln Likelihood | | | −125,791 | −125,857 |
| | BIC | | | 252,975 | 253,232 |
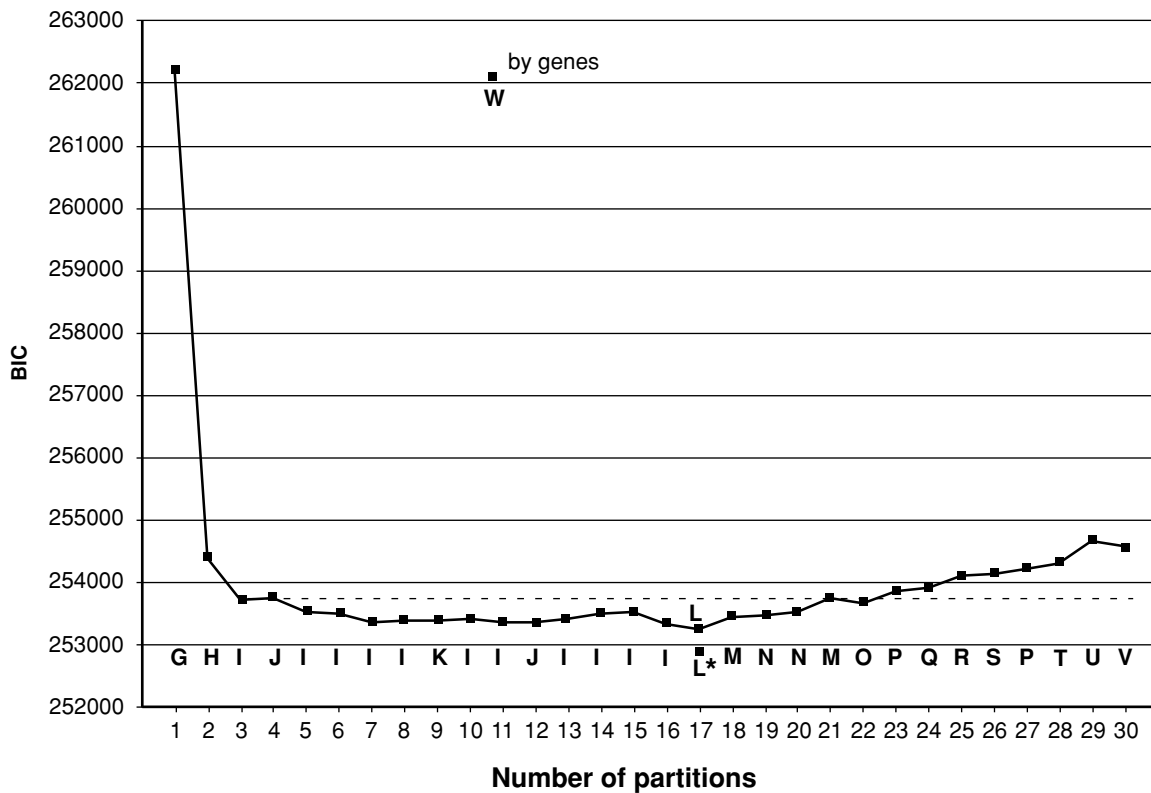


FIGURE 3. BIC values for Bayesian analyses under different data partitioning schemes. Letters G to W indicate alternative best tree topologies supported by different partitioning schemes; these topologies are shown in Supplementary Materials.
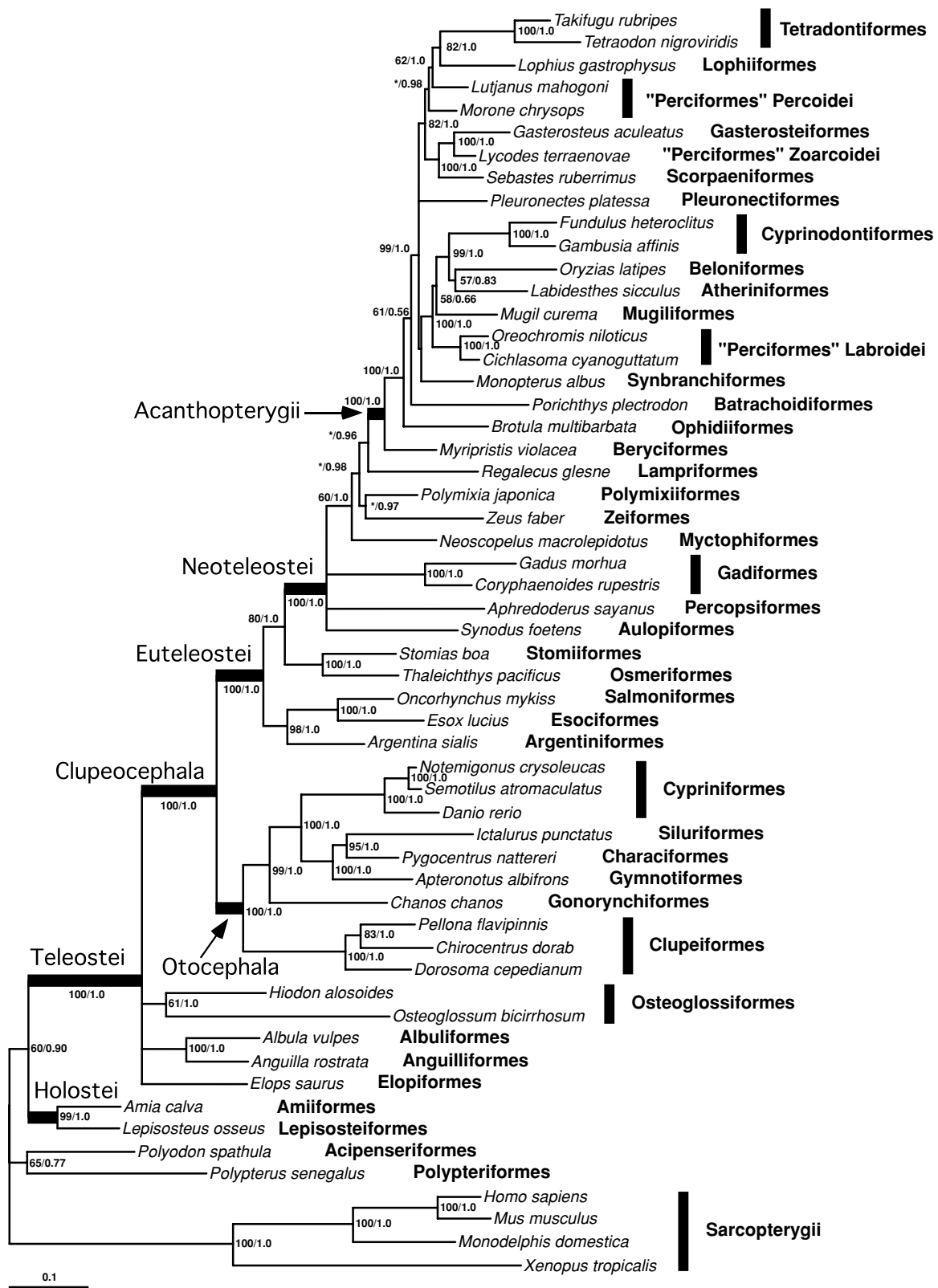
FIGURE 4.  Strict consensus tree of ray-finned fishes based on two different trees (topologies C and L) obtained by partitioned analyses of 10 nuclear genes (7995 bp), under ML and Bayesian criteria, respectively. Data were partitioned into 10 data blocks for the ML analysis and 17 blocks for the Bayesian analysis. The numbers on branches are ML bootstrap values and Bayesian posterior probabilities. Asterisks indicate a bootstrap value of <50%. The names of species, orders, and supraordinal taxa sampled for this study are indicated.
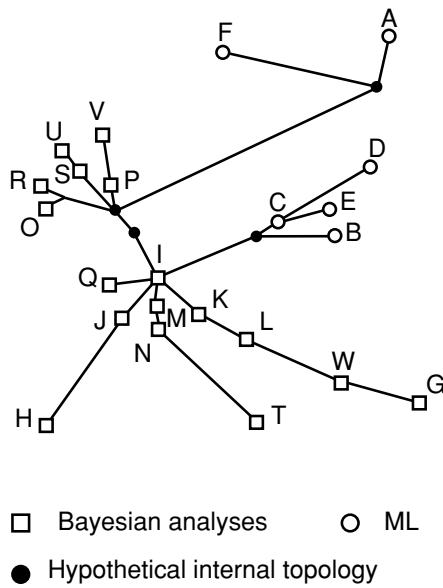
FIGURE 5. The meta-tree (or tree of trees) for all alternative topologies (A to W, see Supplementary Materials, www.systematicbiology. org) resulting from different partitioning strategies. The meta-tree was built by minimizing the total Robinson-Foulds distance (Nye, 2008, www.mas.ncl.ac.uk/~ntmwn/phylo_comparison/multiple.html).

alternative partitioning schemes for phylogenetic analysis of complex multilocus data sets.

In this article, cluster analysis is proposed as a method to explore alternative partitioning schemes based on overall similarity among predefined data blocks. Our results show that a relatively simple model based solely on partitioning by codon position (total data set split into two or three partitions) resulted in the largest improvement in AIC and BIC values (Figs. 2 and 3), indicating that most heterogeneity in this example is explained by
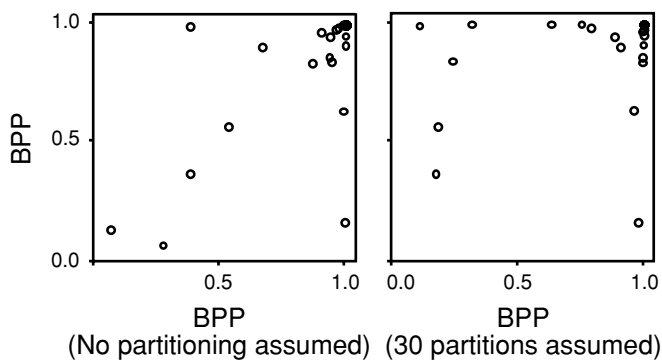


FIGURE 6. The effect of over- and underpartitioning on nodal support (bipartition posterior probabilities, BPPs) for Bayesian phylogenies inferred with alternative partitioning schemes. Models with two extreme partitioning strategies shown here include no partitioning (left) or 30 partitions (right). The BPPs for each of these cases were plotted against values obtained with the 17-partition model chosen as the optimal partitioning strategy. Diagonal lines imply equal values of nodal support for the compared models. On the left graph, 91 observations (among 112 shared bipartitions) for both models have BPP = 1, and in the right graph 91 observations (among 110 shared bipartitions) for both models also have BPP = 1.
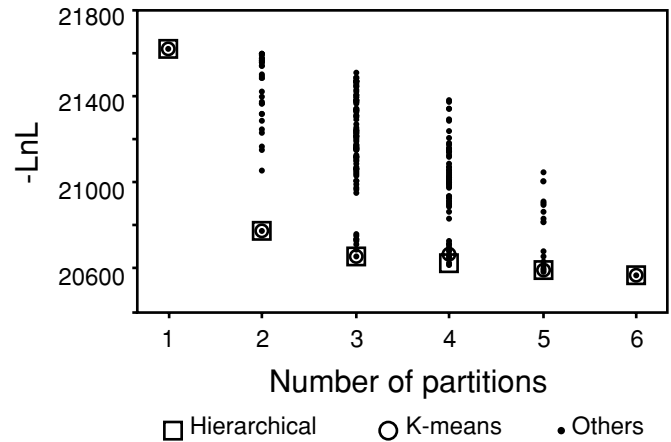


FIGURE 7. Likelihood values (vertical axis) for phylogenetic trees obtained under different partitioning schemes (from 1 to 6 partitions, horizontal axis). A subset of the data (2 genes) was used to exhaustively explore all 203 possible partitioning schemes for 6 data blocks (2 genes × 3 codon positions, $B_6 = 203$). The results of two different clustering methods, hierarchical (open squares) and k-means (open circles) are presented.

evolutionary differences among codon positions, especially due to third positions. Further partitioning of the data resulted in minor improvements. Partitioning the data by gene only (e.g., Nylander et al., 2004; Nishihara et al., 2007), as the present analysis shows, can be as ineffective as using a single data block. This pattern may be expected when genes do not differ significantly in their overall evolutionary rates ("fast" vs "slow" genes). The 10 genes chosen for our study exhibit relatively homogeneous rates (Appendices 2 and 3), most likely as a consequence of the way that they were chosen (Li et al., 2007). Therefore, this example may not be representative of complex data sets that combine fast and slow genes (e.g., mitochondrial and nuclear gene sequences). Our results show that the most important parameters defining the clustering schemes are the rate multiplier and the gamma function parameter *alpha*. Differences in these parameters among the a priori partitions (1st-, 2nd-, and 3rd-codon positions for each gene) determine low within-cluster variance and high between-cluster variance. This is most notably reflected in the clustering diagrams shown in Figure 1, where for three clusters the partitions are grouped according to codon position, and these are known to evolve at very different rates.

TABLE 6. Tests of alternative hypotheses of interrelationships among the early-branching actinopterygians. Based on the Shimodaira and Hasegawa test.

| Hypotheses tested | References | SH *P*-value |
|---|---|---|
| Polypteriformes basal | Nelson, 1994; Schaeffer, 1973 | 0.823 |
| "Ancient fish" | Inoue et al., 2003; Venkatesh et al., 2001 | 0.225 |
| *Amia* and teleosts sister-group | Grande and Bemis, 1996; Patterson, 1973 | 0.028 |
| *Lepisosteidae* and teleost sister-group | Olsen, 1984 | 0.023 |

Overpartitioning the data can have negative effects on phylogenetic inference, as has been discussed for cases using parameter-rich models that may lead to overparameterization in general (Nylander et al., 2004; Sullivan and Joyce, 2005). In this example, partitioning the data with more blocks always resulted in higher likelihood scores, but higher number of parameters may result in increasing sampling error. This problem can be most severe when some of the data blocks are relatively small, having a few characters, potentially hindering the rate of convergence of MCMC chains to a stationary posterior distribution. A slower rate of convergence was observed for Bayesian analyses when data were grouped into more than 10 to 15 blocks, as indicated by the standard deviation of split frequencies (Table 3, last column). The progression of averaged standard deviation of split frequencies (ASDSF) recorded for the MCMC runs with the worse final ASDSF values (models with 27, 28, 29, and 30 partitions) show that the best ASDSF values were achieved before reaching 2 million MCMC generations; therefore, the additional one million generations had little effect in achieving stationarity.

The optimal partitioning scheme identified with the clustering approach partitioned the data with an intermediate number of blocks, 10 for ML and 17 for Bayesian inference. Application of this approach to more complex data sets should yield reasonable partitioning schemes for analysis as a compromise solution to avoid systematic errors (underpartitioning) or overparametrization. Simulation studies similar to the one reported by Brown and Lemmon (2007) may be used to test the generality of our approach. These authors tested the utility of Bayes factors as a criterion for choosing among alternative partitioning strategies. They found that BF provided a robust method to determine the simulated partitioned model. Although we based our choice of optimal partitioning scheme on BIC or AIC, Bayes factors gave a similar result, selecting a 22-group partitioning scheme (see Table 3 and Supplementary Materials) that should also be considered as a candidate for "optimal partitioning." Although the BIC seems to prefer models with fewer parameters, the performance of the BIC as applied to phylogenetic models is not well understood and deserves further exploration.

A critical shortcoming of this method is that it depends on the definition of combinable data blocks a priori (e.g., partitions by gene, codon position, or other structural considerations). A more robust approach free of this constraint should be based on methods that do not require a priori definitions of data blocks, perhaps by exploration of all data simultaneously on a per site basis, such as the mixture model (Pagel and Meade, 2004). Combination of $k$ individual sites into an optimal number of homogeneous partitions by exploration of all possible ($B_k$) combinations would be computationally intractable (NP-hard), but it could be achieved heuristically by cluster analysis, as proposed here for a more limited number of predefined data blocks.

The effect of partitioning scheme on tree topology is shown in Figures 2 and 3 and the supplementary figures (available online at www.systematicbiology.org).

Earlier work has shown that contrary to the large change in the likelihood scores among alternative partitioning schemes, the topology of the resulting phylogenetic trees has been relatively stable (Buckley et al., 2001). In this study, however, changes in topology were obtained when data were analyzed under different number of blocks for both ML and Bayesian methods (Figs. 2 and 3). These effects were more conspicuous in Bayesian analysis than for ML methods (Fig. 3), which is consistent with the properties of Bayesian approaches accounting for model parameter value uncertainty. The failure to effectively reach convergence among MCMC runs with increasing number of partitions, indicated by the standard deviation of split frequencies, also is consistent with the observed changes in the supported topologies (Fig. 3).

### "Lower" Actinopterygians

The classic concept of "Chondrostei" that groups *Polypterus* and living sturgeons and paddlefishes and their fossil relatives (Schaeffer, 1973; Nelson, 1994) received some support in this study, albeit with a low bootstrap value of 65% and a posterior probability of 0.74 (Fig. 4). However, recent evidence from both morphological (Grande and Bemis, 1996; Gardiner et al., 2005) and molecular (Venkatesh et al., 2001; Inoue et al., 2003; Kikugawa et al., 2004) data suggests that "Chondrostei" is actually a paraphyletic group. The current consensus view places polypteriforms as the sister taxon to all other actinoterygians, while considering sturgeons and paddlefish as the sister group to neopterygians (*Lepisosteus*, *Amia*, and teleosts; Nelson, 2006).

Most morphological (Regan, 1923; Patterson, 1973) and molecular (Lê et al., 1993; Kikugawa et al., 2004; Crow and Wagner, 2006; Hurley et al., 2007) evidence supports the monophyly of Neopterygii, a group represented by extant lepisosteiforms, amiiforms, and teleosts. However, the relationships among these three lineages are hotly debated. Historically, *Lepisosteus* and *Amia* were grouped into a monophyletic clade as "Holostei," placed as the sister-group to teleosts (Nelson, 1969; Jessen, 1972). More recent morphological hypotheses suggest that either Amiiformes (Patterson, 1973; Grande and Bemis, 1996) or Lepisosteiformes (Olsen, 1984) is the sister-group of teleosts. However, mitogenome data and indel patterns in the nuclear gene RAG2 support a very different view, with Acipenseriformes, *Lepisosteidae,* and *Amia* forming a monophyletic "ancient fish" group. This group is placed as the sister-group to teleost (Venkatesh et al., 2001; Inoue et al., 2003). Our study supports the "Holostei" hypothesis with high probability. The "Holostei" hypothesis also was recovered in a study using multiple nuclear genes (Kikugawa et al., 2004) and in a reanalysis of morphological characters using both extant and fossil species (Hurley et al., 2007).

### Interrelationships among Major Teleostean Lineages

The monophyly of Teleostei is supported by many morphological characters (de Pinna, 1996; Arratia, 2000). Four major teleostean lineages, Elopomorpha,

Osteoglossomorpha, Ostarioclupeomorpha (or Otocephala = Clupeiformes plus Ostariophysi), and Euteleostei are currently recognized (Nelson, 2006). All these, except Elopomorpha, received strong support in this study. Ostarioclupeomorphs are generally placed as the sister-group to euteleosts (Lê et al., 1993; Arratia, 1997; Inoue et al., 2001), a grouping named Clupeocephala, which excludes elopomorphs and osteoglossomorphs. However, interrelationships among elopomorphs, osteoglossomorphs, and Clupeocephala are still controversial. Both morphological (Patterson and Rosen, 1977) and molecular (Inoue et al., 2001) studies support the position of osteoglossomorphs at the base of the teleosts, but this view was challenged by the alternative hypothesis suggesting that elopomorphs are the living sister-group of all other extant teleosts (Arratia, 1991, 1997, 2000; Shen, 1996). A third alternative was suggested by Lê et al. (1993) based on relatively weak evidence from 28S ribosomal gene sequences, with osteoglossomorphs and elopmorphs more closely related to each other than to the rest of the teleosts. The consensus phylogeny obtained in this study (Fig. 4) does not resolve these relationships.

Results of this study show strong support for the Ostarioclupeomorpha hypothesis (Otocephala), with Clupeiformes as a sister group to Ostariophysi (Fig. 4), in contrast to a recent result using mitogenomic data (Saitoh et al., 2003), suggesting that gonorynchiforms are more closely related to Clupeiformes. Relationships within Ostariophysi are consistent with the current view placing Cypriniforms as a sister to the rest (Fink and Fink, 1981), but relationships among Characiformes, Siluriformes, and Gymnotiformes cannot be resolved with confidence with such limited taxonomic sampling.

### Protacanthopterygians

The composition of Protacanthopterygii has changed drastically since Greenwood et al. (1966) defined the group as primitive teleosts of their division III (Fink, 1984; Williams, 1987; Arratia, 1997; Lopez et al., 2004). The current hypothesis of Protacanthopterygii (Nelson, 2006) includes Argentiniformes, Osmeriformes, Salmoniformes, and Esociformes, but esociforms were sometimes regarded as the sister-group to neoteleosts (Johnson and Patterson, 1996). Many recent studies (including this one) support a sister-taxon relationship between Esociformes and Salmoniformes (Williams, 1987; Arratia, 1997; Ishiguro et al., 2003; Lopez et al., 2004). Interestingly, Lopez et al. (2004) also suggested a novel sister-group relationship between osmeriforms and stomiiforms (Neoteleostei) based on RAG-1 and mtDNA sequences. Results from this study corroborate both findings of Lopez et al. (Fig. 4), suggesting that Stomiiformes should be excluded from Neoteleostei.

### Neoteleostei

Neoteleostei is a monophyletic group defined by few morphological characters (Johnson, 1992; Nelson, 1994), but this study provides strong support for Neoteleostei

(excluding stomiiforms) with 100% bootstrap value and 1.0 Bayesian posterior probability. Paracanthopterygii is a classical grouping of neoteleosts that has been extensively debated in the literature (Greenwood et al., 1966; Patterson and Rosen, 1989; Miya et al., 2003, 2005). None of the taxa traditionally assigned to the Paracanthopterygii (Gadiformes, Percopsiformes, Lophiiformes, Ophidiformes, Batrachoidiformes) formed a monophyletic group in this study (Fig. 4) but were instead scattered among other Acanthopterygian lineages. The hypothesis of paracanthopterygians proposed by mitogenomic analyses (Miya et al., 2003, 2005) also was not supported in this study since *Polymixia* and *Zeus* did not form a monophyletic group with gadiforms and percopsiforms.

### Acanthopterygii

If ophidiiforms, batrachoidiforms, and lophiiforms are included, Acanthopterygii also is strongly supported in this study as a monophyletic group. Beryciformes, Ophidiiformes, and Batrachoidifromes branch near the base of acanthopterygians, and the rest of the taxa included in this study formed a monophyletic group of crown acanthopterygians with a 99% bootstrap support. Taxa traditionally assigned to the order Perciformes (*Lutjanus, Morone, Lycodes, Oreochromis*, and *Cichlasoma*) do not form a monophyletic group, in agreement with several results supporting the polyphyletic nature of this group (Lauder and Liem, 1983; Johnson and Patterson, 1993; Miya et al., 2003; Nelson, 2006). One interesting group among the crown acanthopterygian taxa is a well-supported clade of atherinomorphs (Atheriniforms, Beloniformes, and Cyprinodontiformes), Mugiliformes, and cichlids. A close relationship among rice fish (*Oryzias*) and tilapia (*Oreochromis*) was first suggested by a phylogenomic study with model organisms (Chen et al., 2004).

Many relationships among major fish lineages still need to be resolved to obtain a solid phylogenetic framework for the ray-finned fishes. The current study presents an important contribution, both methodological and practical, towards developing appropriate strategies to achieve this goal. Future studies based on the set of 10 nuclear genes analyzed here but using a dense taxonomic sampling should provide promising results to resolve the persistent ichthyological dilemma appropriately labeled "the bush at the top" (Rosen, 1982).

REFERENCES

Amores, A., T. Suzuki, Y. L. Yan, J. Pomeroy, A. Singer, C. Amemiya, and J. H. Postlethwait. 2004. Developmental roles of pufferfish Hox clusters and genome evolution in ray-fin fish. Genome Res. 14:1–10.

Arratia, G. 1991. The caudal skeleton of Jurassic teleosts; a phylogenetic analysis. Pages 249–340 in Early vertebrates and related problems in evolutionary biology (M.-M. Chang, H. Liu, and G.-R. Zhang, eds.). Science Press, Beijing.

Arratia, G. 1997. Basal teleosts and teleostean phylogeny. Palaeo. Ichthyologica 7:5–168.

Arratia, G. 2000. Phylogenetic relationships of teleostei: Past and present. Estud. Oceanol. 19:19–51.

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

Bell, E. T. 1934. Exponential numbers. Am. Math. Monthly 41:411–419.

Brandley, M. C., A. Schmitz, and T. W. Reeder. 2005. Partitioned Bayesian analyses, partition choice, and the phylogenetic relationships of scincid lizards. Syst. Biol. 54:373–390.

Brown, J. M., and A. R. Lemmon. 2007. The importance of data partitioning and the utility of Bayes factors in Bayesian phylogenetics. Syst. Biol. 56:643–655.

Buckley, T. R., C. Simon, and G. K. Chambers. 2001. Exploring among-site rate variation models in a maximum likelihood framework using empirical data: Effects of model assumptions on estimates of topology, branch lengths, and bootstrap support. Syst. Biol. 50:67–86.

Bull, J. J., J. P. Huelsenbeck, C. W. Cunningham, D. L. Swofford, and P. J. Waddell. 1993. Partitioning and combining data in phylogenetic analysis. Syst. Biol. 42:384–397.

Burnham, K., and D. Anderson. 2002. Model selection and multimodel inference: A practical information-theoretic approach, 2nd edition. Springer-Verlag, New York.

Castoe, T. A., T. M. Doan, and C. L. Parkinson. 2004. Data partitions and complex models in Bayesian analysis: The phylogeny of Gymnophthalmid lizards. Syst. Biol. 53:448–469.

Castoe, T. A., and C. L. Parkinson. 2006. Bayesian mixed models and the phylogeny of pitvipers (Viperidae: Serpentes). Mol. Phylogenet. Evol. 39:91–110.

Caterino, M. S., R. D. Reed, M. M. Kuo, and F. A. Sperling. 2001. A partitioned likelihood analysis of swallowtail butterfly phylogeny (Lepidoptera:Papilionidae). Syst. Biol. 50:106–127.

Chen, W. J., C. Bonillo, and G. Lecointre. 2003. Repeatability of clades as a criterion of reliability: A case study for molecular phylogeny of Acanthomorpha (Teleostei) with larger number of taxa. Mol. Phylogenet. Evol. 26:262–288.

Chen, W. J., G. Ortí, and A. Meyer. 2004. Novel evolutionary relationship among four fish model systems. Trends Genet. 20:424–431.

Chiu, C. H., K. Dewar, G. P. Wagner, K. Takahashi, F. Ruddle, C. Ledje, P. Bartsch, J. L. Scemama, E. Stellwag, C. Fried, S. J. Prohaska, P. F. Stadler, and C. T. Amemiya. 2004. Bichir HoxA cluster sequence reveals surprising trends in ray-finned fish genomic evolution. Genome Res. 14:11–7.

Cloutier, R., and G. Arratia. 2004. Early diversification of actinopterygians. Pages 217–270 in Recent advances in the origin and early radiation of vertebrates (G. Arratia, M. V. H. Wilson, and R. Cloutier, eds.). Verlag Dr Friedrich Pfeil, Munich.

Collins, T. M., O. Fedrigo, and G. J. Naylor. 2005. Choosing the best genes for the job: The case for stationary genes in genome-scale phylogenetics. Syst. Biol. 54:493–500.

Comas, I., A. Moya, and F. Gonzalez-Candelas. 2007. From phylogenetics to phylogenomics: The evolutionary relationships of insect endosymbiotic gamma—Proteobacteria as a test case. Syst. Biol. 56:1–16.

Crow, K. D., and G. P. Wagner. 2006. Proceedings of the SMBE Tri-National Young Investigators' Workshop 2005. What is the role of genome duplication in the evolution of complexity and diversity? Mol. Biol. Evol. 23:887–892.

Cunningham, C. W., H. Zhu, and D. M. Hillis. 1998. Best-fit maximum-likelihood models for phylogenetic inference: Empirical tests with known phylogenies. Evolution 52:978–987.

Curole, J. P., and T. D. Kocher. 1999. Mitogenomics: Digging deeper with complete mitochondrial genomes. Trends Ecol. Evol. 14:394–398.

de Pinna, M. C. C. 1996. Teleostean monophyly. Pages 193–207 in Interrelationships of fishes (M. L. J. Stiassny, L. R. Parenti, and G. D. Johnson, eds.). Academic Press, San Diego.

Dettai, A., and G. Lecointre. 2005. Further support for the clades obtained by multiple molecular phylogenies in the acanthomorph bush. C. R. Biol. 328:674–689.

Dimmick, W. W., and A. Larson. 1996. A molecular and morphological perspective on the phylogenetic relationships of the otophysan fishes. Mol. Phylogenet. Evol. 6:120–133.

Fink, S. V., and W. L. Fink. 1981. Interrelationships of the ostariophysan fishes (Teleostei). Zool. J. Linnean Soc. 72:297–353.

Fink, W. L. 1984. Salmoniforms: Introduction. Pages 1–139 in Ontogeny and systematics of fishes. American Society of Ichthyologists and Herpetologists, Special Publication No. 1 (H. G. Moser, W. J. Richards, D. M. Cohen, M. P. Fahay, A. W. Kendall, and S. L. Richardson, eds.). Allen Press, Lawrence, Kansas.

Gardiner, B. G., B. Schaeffert, and J. A. Masserie. 2005. A review of the lower actinopterygian phylogeny. Zool. J. Linnean Soc. 144:511–525.

Gee, H. 2003. Evolution: Ending incongruence. Nature 425:782.

Grande, L., and W. E. Bemis. 1996. Interrelationships of Acipenseriformes, with comments on "Chondrostei." Pages 85–115 in Interrelationships of fishes (M. L. J. Stiassny, L. R. Parenti, and G. D. Johnson, eds.). Academic Press, San Diego.

Greenwood, P. H., R. S. Miles, C. Patterson, and Linnean Society of London. 1973. Interrelationships of fishes. Academic Press, London.

Greenwood, P. H., D. E. Rosen, S. H. Weitzman, and G. S. Meyers. 1966. Phyletic studies of teleostean fishes, with a provisional classification of living forms. Bull. Am. Mus. Nat. Hist. 131:339–456.

Hartigan, J. A. 1975. Clustering algorithms. Wiley, New York.

Hartigan, J. A., and M. A. Wong. 1979. A k-means clustering algorithm. Appl. Stat. 28:100–108.

Helfman, G. S., B. B. Collette, and D. E. Facey. 1997. The diversity of fishes. Blackwell Science, Malden, Massachusetts.

Hoon, M. 2002. Cluster 3.0 for Windows, Mac OS X, Linux, Unix. Distributed by the author at http://bonsai.ims.u-tokyo.ac.jp/~mdehoon/software/cluster/.

Hurley, I. A., R. L. Mueller, K. A. Dunn, E. J. Schmidt, M. Friedman, R. K. Ho, V. E. Prince, Z. Yang, M. G. Thomas, and M. I. Coates. 2007. A new time-scale for ray-finned fish evolution. Proc. R. Soc. B 274:489–498.

Inoue, J. G., M. Miya, K. Tsukamoto, and M. Nishida. 2001. A mitogenomic perspective on the basal teleostean phylogeny: Resolving higher-level relationships with longer DNA sequences. Mol. Phylogenet. Evol. 20:275–285.

Inoue, J. G., M. Miya, K. Tsukamoto, and M. Nishida. 2003. Basal actinopterygian relationships: A mitogenomic perspective on the phylogeny of the "ancient fish." Mol. Phylogenet. Evol. 26:110–120.

Ishiguro, N. B., M. Miya, and M. Nishida. 2003. Basal euteleostean relationships: A mitogenomic perspective on the phylogenetic reality of the "Protacanthopterygii." Mol. Phylogenet. Evol. 27:476–488.

Jessen, H. 1972. Schultergürtel und Pectoralflosse bei Actinopteriern. Fossils Strata 1:1–101.

Jobb, G. 2006. TREEFINDER, version May 2006. Distributed by the author at www.treefinder.de.

Johnson, G. D. 1992. Monophyly of the euteleostean clades: Neoteleostei, Eurypterygii, and Ctenosquamata. Copeia 1992:8–25.

Johnson, G. D., and C. Patterson. 1993. Percomorph phylogeny: A survey of acanthomorphs and a new proposal. Bull. Mar. Sci. 52:554–626.

Johnson, G. D., and C. Patterson. 1996. Relationships of lower euteleostean fishes. Pages 251–332 in Interrelationships of fishes (M. L. J. Stiassny, L. R. Parenti, and G. D. Johnson, eds.). Academic Press, San Diego.

Kelchner, S. A., and M. A. Thomas. 2007. Model use in phylogenetics: Nine key questions. Trends Ecol. Evol. 22:87–94.

Kelsey, C. R., K. A. Crandall, and A. F. Voevodin. 1999. Different models, different trees: The geographic origin of PTLV-I. Mol. Phylogenet. Evol. 13:336–347.

Kikugawa, K., K. Katoh, S. Kuraku, H. Sakurai, O. Ishida, N. Iwabe, and T. Miyata. 2004. Basal jawed vertebrate phylogeny inferred from multiple nuclear DNA-coded genes. BMC Biol. 2:3.

Kocher, T. D., and C. A. Stepien. 1997. Molecular systematics of fishes. Academic Press, San Diego.

Kubatko, L. S., and J. H. Degnan. 2007. Inconsistency of phylogenetic estimates from concatenated data under coalescence. Syst. Biol. 56:17–24.

Kumar, S., K. Tamura, and M. Nei. 2004. MEGA3: Integrated software for molecular evolutionary genetics analysis and sequence alignment. Brief Bioinformatics 5:150–163.

Lauder, G. V., and K. F. Liem. 1983. The evolution and interrelationships of the Actinopterygian fishes. Bull. Mus. Comp. Zool. 150:95–197.

Lê, H. L., G. Lecointre, and R. Perasso. 1993. A 28S rRNA-based phylogeny of the gnathostomes: First steps in the analysis of conflict and congruence with morphologically based cladograms. Mol. Phylogenet. Evol. 2:31–51.

Li, C., G. Ortí, G. Zhang, and G. Lu. 2007. A practical approach to phylogenomics: The phylogeny of ray-finned fish (Actinopterygii) as a case study. BMC Evol. Biol. 7:44.

Lopez, A. J., W. J. Chen, and G. Ortí. 2004. Esociform phylogeny. Copeia 2004:449–464.

McGuire, J. A., C. C. Witt, D. L. Altshuler, and J. V. Remsen, Jr. 2007. Phylogenetic systematics and biogeography of hummingbirds: Bayesian and maximum likelihood analyses of partitioned data and selection of an appropriate partitioning strategy. Syst. Biol. 56:837–856.

McMahon, M. M., and M. J. Sanderson. 2006. Phylogenetic supermatrix analysis of GenBank sequences from 2228 papilionoid legumes. Syst. Biol. 55:818–836.

Meyer, A., and R. Zardoya. 2003. Recent advances in the (molecular) phylogeny of vertebrates. Annu. Rev. Ecol. Evol. Syst. 34:311–318.

Milligan, G. W. 1980. An examination of the effect of six types of error perturbation on fifteen clustering algorithms. Psychometrika 45:325–342.

Minin, V., Z. Abdo, P. Joyce, and J. Sullivan. 2003. Performance-based selection of likelihood models for phylogeny estimation. Syst. Biol. 52:674–683.

Miya, M., A. Kawaguchi, and M. Nishida. 2001. Mitogenomic exploration of higher teleostean phylogenies: A case study for moderate-scale evolutionary genomics with 38 newly determined complete mitochondrial DNA sequences. Mol. Biol. Evol. 18:1993–2009.

Miya, M., and M. Nishida. 2000. Use of mitogenomic information in teleostean molecular phylogenetics: A tree-based exploration under the maximum-parsimony optimality criterion. Mol. Phylogenet. Evol. 17:437–455.

Miya, M., T. P. Satoh, and M. Nishida. 2005. The phylogenetic position of toadfishes (order Batrachoidiformes) in the higher ray-finned fish as inferred from partitioned Bayesian analysis of 102 whole mitochondrial genome sequences. Biol. J. Linn. Soc. Lond. 85:289–306.

Miya, M., H. Takeshima, H. Endo, N. B. Ishiguro, J. G. Inoue, T. Mukai, T. P. Satoh, M. Yamaguchi, A. Kawaguchi, K. Mabuchi, S. M. Shirai, and M. Nishida. 2003. Major patterns of higher teleostean phylogenies: A new perspective based on 100 complete mitochondrial DNA sequences. Mol. Phylogenet. Evol. 26:121–138.

Mueller, R. L., J. R. Macey, M. Jaekel, D. B. Wake, and J. L. Boore. 2004. Morphological homoplasy, life history evolution, and historical biogeography of plethodontid salamanders inferred from complete mitochondrial genomes. Proc. Natl. Acad. Sci. USA 101:13820–13825.

Nelson, G. J. 1969. Gill arches and the phylogeny of fishes, with notes on the classification of vertebrates. Bull. Am. Mus. Nat. Hist. 141:475–552.

Nelson, J. S. 1994. Fishes of the world, 3rd edition. John Wiley & Sons, New York.

Nelson, J. S. 2006. Fishes of the world, 4th edition. John Wiley & Sons, New York.

Nishihara, H., N. Okada, and M. Hasegawa. 2007. Rooting the eutherian tree: The power and pitfalls of phylogenomics. Genome Biol. 8:R199.

Noonan, J. P., J. Grimwood, J. Danke, J. Schmutz, M. Dickson, C. T. Amemiya, and R. M. Myers. 2004. Coelacanth genome sequence reveals the evolutionary history of vertebrate genes. Genome Res. 14:2397–2405.

Nye, T. 2008. Trees of trees: An approach to comparing multiple alternative phylogenies. Syst. Biol. In press.

Nylander, J. A., F. Ronquist, J. P. Huelsenbeck, and J. L. Nieves-Aldrey. 2004. Bayesian phylogenetic analysis of combined data. Syst. Biol. 53:47–67.

Olsen, P. E. 1984. The skull and pectoral girdle of the parasemionotid fish Watsonulus eugnathoides from the Early Triassic Sakamena Group of Madagascar, with comments on the relationships of the holostean fishes. J. Vert. Paleontol. 4:481–499.

Orrell, T. M., and K. E. Carpenter. 2004. A phylogeny of the fish family Sparidae (porgies) inferred from mitochondrial sequence data. Mol. Phylogenet. Evol. 32:425–434.

Ortí, G., and A. Meyer. 1996. Molecular evolution of Ependymin and the phylogenetic resolution of early divergences among Euteleost fishes. Mol. Biol. Evol. 13:556–573.

Pagel, M., and A. Meade. 2004. A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. Syst. Biol. 53:571–581.

Pagel, M., and A. Meade. 2005. Mixture models in phylogenetic inference. Pages 63–90 in Mathematics of evolution and phylogeny (O. Gascuel, ed.). Oxford University Press, New York.

Patterson, C. 1973. Interrelationships of holosteans. Pages 207–226 in Interrelationships of fishes (P. H. Greenwood, R. S. Miles, and C. Patterson, eds.). Academic Press, London.

Patterson, C., and D. E. Rosen. 1977. Review of ichthyodectiform and other Mesozoic teleost fishes and the theory and practice of classifying fossils. Bull. Am. Mus. Nat. Hist. 158:81–172.

Patterson, C., and D. E. Rosen. 1989. The Paracanthopterygii revisited: Order and disorder. Pages 5–36 in Papers on the systematics of gadiform fishes (D. M. Cohen, ed.). Natural History Museum of Los Angeles County, Los Angeles, California.

Philippe, H., N. Lartillot, and H. Brinkmann. 2005. Multigene analyses of bilaterian animals corroborate the monophyly of Ecdysozoa, Lophotrochozoa, and Protostomia. Mol. Biol. Evol. 22:1246–1253.

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

Posada, D., and T. R. Buckley. 2004. Model selection and model averaging in phylogenetics: Advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests. Syst. Biol. 53:793–808.

Posada, D., and K. A. Crandall. 1998. ModelTest: Testing the model of DNA substitution. Bioinformatics 14:817–818.

Posada, D., and K. A. Crandall. 2001. Selecting the best-fit model of nucleotide substitution. Syst. Biol. 50:580–601.

Poux, C., O. Madsen, E. Marquard, D. R. Vieites, W. W. de Jong, and M. Vences. 2005. Asynchronous colonization of Madagascar by the four endemic clades of primates, tenrecs, carnivores, and rodents as inferred from nuclear genes. Syst. Biol. 54:719–730.

Pupko, T., D. Huchon, Y. Cao, N. Okada, and M. Hasegawa. 2002. Combining multiple data sets in a likelihood analysis: Which models are the best? Mol. Biol. Evol. 19:2294–2307.

Rannala, B. 2002. Identifiability of parameters in MCMC Bayesian inference of phylogeny. Syst. Biol. 51:754–760.

Reed, R. D., and F. A. Sperling. 1999. Interaction of process partitions in phylogenetic analysis: An example from the swallowtail butterfly genus Papilio. Mol. Biol. Evol. 16:286–297.

Regan, C. T. 1923. The skeleton of Lepidosteus, with remarks on the origin and evolution of the lower neopterygian fishes. Proc. Zool. Soc. 1923:445–461.

Remm, M., C. E. Storm, and E. L. Sonnhammer. 2001. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. J. Mol. Biol. 314:1041–1052.

Robinson, D., and L. Foulds. 1981. Comparison of phylogenetic trees. Math. Biosci. 53:131–147.

Rokas, A., N. King, J. Finnerty, and S. B. Carroll. 2003a. Conflicting phylogenetic signals at the base of the metazoan tree. Evol. Dev. 5:346–359.

Rokas, A., D. Kruger, and S. B. Carroll. 2005. Animal evolution and the molecular signature of radiations compressed in time. Science 310:1933–1938.

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

Rosen, D. E. 1982. Teleostean interrelationships, morphological function and evolutionary inference. Am. Zool. 22:261–273.

Saitoh, K., M. Miya, J. G. Inoue, N. B. Ishiguro, and M. Nishida. 2003. Mitochondrial genomics of ostariophysan fishes: Perspectives on phylogeny and biogeography. J. Mol. Evol. 56:464–472.

Saitou, N., and M. Nei. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4:406–425.

Schaeffer, B. 1973. Interrelationships of chondrosteans. Pages 207–226 *in* Interrelationships of fishes (P. H. Greenwood, R. S. Miles, and C. Patterson, eds.). Academic Press, London.

Schwarz, G. 1978. Estimating the dimensions of a model. Ann. Stat. 6:461–464.

Shen, M. 1996. Fossil "osteoglossomorphs" in East Asia and their implications in teleostean phylogeny. Pages 261–272 *in* Mesozoic fishes: Systematics and plaleoecology (G. Arratia, and G. Viohl, eds.). Verlag Dr. F. Pfeil, München, Germany.

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

Smith, W. L., and M. T. Craig. 2007. Casting the percomorph net widely: The importance of broad taxonomic sampling in the search for the placement of Serranid and Percid fishes. Copeia 2007:35–55.

Soltis, D. E., V. A. Albert, V. Savolainen, K. Hilu, Y. L. Qiu, M. W. Chase, J. S. Farris, S. Stefanovic, D. W. Rice, J. D. Palmer, and P. S. Soltis. 2004. Genome-scale data, angiosperm relationships, and "ending incongruence": A cautionary tale in phylogenetics. Trends Plant Sci. 9:477–483.

Springer, V. G., and G. D. Johnson. 2004. Study of the dorsal gill-arch musculature of Teleostome fishes, with special reference to the Actinopterygii. Bull. Biol. Soc. Wash. 11:1–260.

Stiassny, M. L. J., L. R. Parenti, and G. D. Johnson. 1996. Interrelationships of fishes. Academic Press, San Diego.

Sullivan, J., and P. Joyce. 2005. Model selection in phylogenetics. Annu. Rev. Ecol. Evol. Syst. 36:445–466.

Sullivan, J., J. A. Markert, and C. W. Kilpatrick. 1997. Phylogeography and molecular systematics of the *Peromyscus aztecus* species group (Rodentia: Muridae) inferred using parsimony and likelihood. Syst. Biol. 46:426–440.

Sullivan, J., D. L. Swofford, and G. J. P. Naylor. 1999. The effect of taxon sampling on estimating rate-heterogeneity parameters of maximum-likelihood models. Mol. Biol. Evol. 16:1347–1356.

Swofford, D. L. 2003. PAUP*: Phylogenetic analysis using parsimony (*and other methods). Version 4. Sinauer Associates, Sunderland, Massachusetts.

Taylor, J. S., I. Braasch, T. Frickey, A. Meyer, and Y. Van de Peer. 2003. Genome duplication, a trait shared by 22000 species of ray-finned fish. Genome Res. 13:382–390.

Thompson, J. D., D. G. Higgins, and T. J. Gibson. 1994. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22:4673–4680.

Van de Peer, Y., J. S. Taylor, and A. Meyer. 2003. Are all fishes ancient polyploids? J. Struct. Funct. Genomics 3:65–73.

Venkatesh, B., M. V. Erdmann, and S. Brenner. 2001. Molecular synapomorphies resolve evolutionary relationships of extant jawed vertebrates. Proc. Natl. Acad. Sci. USA 98:11382–11387.

Waddell, P. J. 2005. Measuring the fit of sequence data to phylogenetic model: Allowing for missing data. Mol. Biol. Evol. 22:395–401.

Wiens, J. J. 2003. Missing data, incomplete taxa, and phylogenetic accuracy. Syst. Biol. 52:528–538.

Wiley, E. O., G. David Johnson, and W. Wheaton Dimmick. 2000. The interrelationships of Acanthomorph fishes: A total evidence approach using molecular and morphological data. Biochem. Syst. Ecol. 28:319–350.

Williams, R. R. G. 1987. The phylogenetic relationships of the salmoniform fishes based on the suspensorim and its muscles. PhD Thesis *in* Department of Zoology University of Alberta, Edmonton.

Yang, Z. 1996. Maximum-likelihood models for combined analyses of multiple sequence data. J. Mol. Evol. 42:587–596.

APPENDIX 1. Taxon sampling and GenBank accession numbers.

| Orders | Families | Genus | Species | zic1 | myh6 | RYR3 | ptr | tbr1 | ENC1 | Glyt | SH3PX3 | plagl2 | sreb2 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Outgroup | | *Xenopus* | *tropicalis* | Ensembl | | Ensembl | Ensembl | Ensembl | Ensembl | Ensembl | Ensembl | Ensembl | Ensembl |
| Outgroup | | *Monodelphis* | *domestica* | Ensembl | | Ensembl | Ensembl | Ensembl | Ensembl | Ensembl | Ensembl | Ensembl | Ensembl |
| Outgroup | | *Mus* | *musculus* | Ensembl | | Ensembl | Ensembl | Ensembl | Ensembl | Ensembl | Ensembl | Ensembl | Ensembl |
| Outgroup | | *Homo* | *sapiens* | Ensembl | | Ensembl | Ensembl | Ensembl | Ensembl | Ensembl | Ensembl | Ensembl | Ensembl |
| Acipenseriformes | Polyodontidae | *Polyodon* | *spathula* | EU001865* | EU001922* | EU001933* | EU001954* | | EU002011* | | | EU002094 | EU002119* |
| Albuliformes | Albulidae | *Albula* | *vulpes* | EU001888* | | EF032935 | EF032948 | EF032961 | EU002032* | EF032987 | EF033000 | EF033013 | EF033026 |
| Amiiformes | Amiidae | *Amia* | *calva* | EF032909 | EF032922 | | EU001977* | | EU002033* | | EU002083* | | |
| Anguilliformes | Anguillidae | *Anguilla* | *rostrata* | EU001889* | EU001923* | EU001947* | EU001979* | | EU002035* | | EU002085* | | |
| Argentiniformes | Argentinidae | *Argentina* | *sialis* | EU001891* | EU001924* | EU001944* | EU001974* | EU002004* | EU002029* | EU002059* | EU002081* | | EU002144* |
| Atheriniformes | Atherinopsidae | *Labidesthes* | *sicculus* | EU001883* | EU001919* | EU001943* | EU001973* | EU001999* | EU002028* | EU002057* | EU002080* | | EU002138* |
| Aulopiformes | Synodontidae | *Synodus* | *foetens* | EU001882* | EU001918* | | EU001965* | EU001998* | EU002021* | EU002050* | | | EU002137* |
| Batrachoidiformes | Batrachoididae | *Porichthys* | *plectrodon* | EU001876* | EU001911* | | | EU001992* | | EU002055* | | EU002104* | EU002131* |
| Beloniformes | Adrianichthyidae | *Oryzias* | *latipes* | EF032914 | EF032927 | EF032940 | EF032953 | EF032966 | EF032979 | EU002042* | EF033005 | EF033018 | EF033031 |
| Beryciformes | Holocentridae | *Myripristis* | *violacea* | EU001880* | EU001916* | EU001942* | EU001971* | EU001996* | EU002026* | EU002043* | EU002078* | EU002107* | EU002135* |
| Characiformes | Characidae | *Pygocentrus* | *nattereri* | | EU001902* | EU001936* | EU001956* | | EU002013* | EU002041* | EU002067* | EU002096* | EU002122* |
| Clupeiformes | Chirocentridae | *Chirocentrus* | *dorab* | EU001864* | EU001899* | EU001932* | EU001953* | | EU002010* | | | | EU002118* |
| Clupeiformes | Clupeidae | *Dorosoma* | *cepedianum* | EU001867* | EU001901* | EU001935* | | | EU002009* | | | EU002093* | EU002121* |
| Clupeiformes | Pristigasteridae | *Pellona* | *flavipinnis* | EU001863* | EU001898* | EU001931* | | | | | | EU002092* | EU002117* |
| Cypriniformes | Cyprinidae | *Danio* | *rerio* | EF032910 | EF032923 | EF032936 | EF032949 | EF032962 | EF032975 | EF032988 | EF033001 | EF033014 | EF033027 |
| Cypriniformes | Cyprinidae | *Notemigonus* | *crysoleucas* | EU001877* | EU001912* | EU001940* | EU001966* | EU001993* | EU002022* | EU002051* | EU002072* | EU002105* | EU002132* |
| Cypriniformes | Cyprinidae | *Semotilus* | *atromaculatus* | EF032921 | EF032934 | EF032947 | EF032960 | EF032973 | EF032986 | EF032999 | EF033012 | EF033025 | EF033038 |
| Cyprinodontiformes | Fundulidae | *Fundulus* | *heteroclitus* | EF032913 | EF032926 | EF032939 | EF032952 | EF032965 | EF032978 | EF032991 | EF033004 | EF033017 | EF033030 |
| Cyprinodontiformes | Poeciliidae | *Gambusia* | *affinis* | EU001872* | EU001907* | | EU001961* | EU001989* | EU002018* | EU002046* | EU002068* | EU002101* | EU002127* |
| Elopiformes | Elopidae | *Elops* | *saurus* | EU001868* | EU001903* | EU001937* | EU001957* | EU001987* | EU002014* | EU002044* | EU002070* | EU002097* | EU002123* |
| Esociformes | Esocidae | *Esox* | *lucius* | EU001870* | EU001905* | EU001938* | EU001959* | | EU002016* | EU002045* | EU002071* | EU002099* | EU002125* |
| Gadiformes | Gadidae | *Gadus* | *morhua* | EU001871* | EU001906* | | EU001960* | | EU002017* | EU002054* | EU002077* | EU002100* | EU002126* |
| Gadiformes | Macrouridae | *Coryphaenoides* | *rupestris* | | EU001915* | | EU001969* | | EU002025* | | | | |
| Gasterosteiformes | Gasterosteidae | *Gasterosteus* | *aculeatus* | EF032912 | EF032925 | EF032938 | EF032951 | EF032964 | EF032977 | EF032990 | EF033003 | EF033016 | EF033029 |
| Gonorynchiformes | Chanidae | *Chanos* | *chanos* | EU001869* | EU001904* | | EU001958* | EU001988* | EU002015* | | EU002069* | EU002098* | EU002124* |
| Gymnotiformes | Apteronotidae | *Apteronotus* | *albifrons* | EU001890* | | | EU001978* | EU002003* | EU002034* | | EU002084* | EU002112* | EU002143* |
| Lampriformes | Regalecidae | *Regalecus* | *glesne* | EU001874* | EU001909* | | EU001963* | | | | | EU002102* | EU002129* |
| Lepisosteiformes | Lepisosteidae | *Lepisosteus* | *osseus* | EU001886* | EU001921* | EU001946* | | | | EU002048* | EU002074* | EU002110* | EU002141* |

APPENDIX 1.  Taxon sampling and GenBank accesion numbers. (*Continued*)

| Orders | Families | Genus | Species | zic1 | myh6 | RYR3 | ptr | tbr1 | ENC1 | Glyt | SH3PX3 | plagl2 | sreb2 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Lophiiformes | Lophiidae | *Lophius* | *gastrophysus* | EU001884* | EU001920* | | EU001975* | EU002000* | EU002030* | | EU002082* | EU002108* | EU002139* |
| Mugiliformes | Mugilidae | *Mugil* | *curema* | EU001878* | EU001913* | | EU001967* | EU001994* | EU002023* | EU002052* | EU002075* | EU002106* | EU002133* |
| Myctophiformes | Neoscopelidae | *Neoscopelus* | *macrolepidotus* | EU001881* | EU001917* | EU001941* | EU001972* | EU001997* | EU002027* | EU002056* | EU002079* | | EU002136* |
| Ophidiiformes | Ophidiidae | *Brotula* | *multibarbata* | EF032920 | EF032933 | EF032946 | EF032959 | EF032972 | EF032985 | EF032998 | EF033011 | EF033024 | EF033037 |
| Osmeriformes | Osmeridae | *Thaleichthys* | *pacificus* | EU001892* | EU001925* | | EU001980* | EU002005* | EU002036* | EU002060* | EU002086* | EU002095* | EU002145* |
| Osteoglossiformes | Hiodontidae | *Hiodon* | *alosoides* | EU001866* | EU001900* | EU001934* | EU001955* | EU001986* | EU002012* | | EU002066* | EU002111* | EU002120* |
| Osteoglossiformes | Osteoglossidae | *Osteoglossum* | *bicirrhosum* | EU001887* | | | | EU002002* | EU002031* | | | | EU002142* |
| Perciformes | Cichlidae | *Cichlasoma* | *cyanoguttatum* | EU001875* | EU001910* | | EU001964* | EU001991* | EU002020* | EU002049* | | EU002103* | EU002130* |
| Perciformes | Cichlidae | *Oreochromis* | *niloticus* | EF032915 | EF032928 | EF032941 | EF032954 | EF032967 | EF032980 | EF032993 | EF033006 | EF033019 | EF033032 |
| Perciformes | Lutjanidae | *Lutjanus* | *mahogoni* | EF032919 | EF032932 | EF032945 | EF032958 | EF032971 | EF032984 | EF032997 | EF033010 | EF033023 | EF033036 |
| Perciformes | Moronidae | *Morone* | *chrysops* | EF032917 | EF032930 | EF032943 | EF032956 | EF032969 | EF032982 | EF032995 | EF033008 | EF033021 | EF033034 |
| Perciformes | Zoarcidae | *Lycodes* | *terraenovae* | EF032918 | EF032931 | EF032944 | EF032957 | EF032970 | EF032983 | EF032996 | EF033009 | EF033022 | EF033035 |
| Percopsiformes | Aphredoderidae | *Aphredoderus* | *sayanus* | EU001873* | EU001908* | EU001939* | EU001962* | EU001990* | EU002019* | EU002047* | EU002073* | | EU002128* |
| Pleuronectiformes | Pleuronectidae | *Pleuronectes* | *platessa* | EU001897* | EU001930* | EU001952* | EU001985* | EU002008* | | EU002065* | EU002091* | EU002116* | EU002148* |
| Polymixiiformes | Polymixiidae | *Polymixia* | *japonica* | EU001893* | EU001926* | EU001948* | EU001981* | | EU002037* | EU002061* | EU002087* | | |
| Polypteriformes | Polypteridae | *Polypterus* | *senegalus* | EU001885* | | EU001945* | EU001976* | EU002001* | | EU002058* | | EU002109* | EU002140* |
| Salmoniformes | Salmonidae | *Oncorhynchus* | *mykiss* | EF032911 | EF032924 | EF032937 | EF032950 | EF032963 | EF032976 | EF032989 | EF033002 | EF033015 | EF033028 |
| Scorpaeniformes | Sebastidae | *Sebastes* | *ruberrimus* | EU001896* | EU001929* | EU001951* | EU001984* | EU002007* | EU002040* | EU002064* | EU002090* | EU002115* | |
| Siluriformes | Ictaluridae | *Ictalurus* | *punctatus* | EF032916 | EF032929 | EF032942 | EF032955 | EF032968 | EF032981 | EF032994 | EF033007 | EF033020 | EF033033 |
| Stomiiformes | Stomiidae | *Stomias* | *boa* | EU001879* | EU001914* | | EU001968* | EU001995* | EU002024* | EU002053* | EU002076* | | EU002134* |
| Symbranchiformes | Synbranchidae | *Monopterus* | *albus* | EU001895* | EU001928* | EU001950* | EU001983* | EU002006* | EU002039* | EU002063* | EU002089* | EU002114* | EU002147* |
| Tetradontiformes | Tetradontidae | *Takifugu* | *rubripes* | Ensembl | Ensembl | Ensembl | Ensembl | Ensembl | Ensembl | Ensembl | Ensembl | Ensembl | Ensembl |
| Tetradontiformes | Tetradontidae | *Tetraodon* | *nigroviridis* | Ensembl | Ensembl | Ensembl | Ensembl | Ensembl | Ensembl | Ensembl | Ensembl | Ensembl | Ensembl |
| Zeiformes | Zeidae | *Zeus* | *faber* | EU001894* | EU001927* | EU001949* | EU001982* | | EU002038* | EU002062* | EU002088* | EU002113* | EU002146* |

*Sequences collected in this study.

APPENDIX 2. Parameters of 30 partitions estimated using ML method and GTR+Γ model.

| Partitions | Substitution rates | | | | | Base frequencies | | | Alpha | Multiplier |
|---|---|---|---|---|---|---|---|---|---|---|
| | AC | AG | AT | CG | CT | A | C | G | | |
| zic1-1 | 0.1906 | 0.2598 | 0.1003 | 0.0238 | 0.3345 | 0.2342 | 0.2518 | 0.3283 | 0.1221 | 0.1577 |
| zic1-2 | 0.1157 | 0.0786 | 0.0265 | 0.7191 | 0.0599 | 0.3194 | 0.2734 | 0.2001 | 0.1387 | 0.0304 |
| zic1-3 | 0.0641 | 0.3168 | 0.0635 | 0.0532 | 0.4324 | 0.1633 | 0.3873 | 0.2838 | 1.2711 | 1.9424 |
| myh6-1 | 0.1050 | 0.0930 | 0.0746 | 0.0366 | 0.5199 | 0.3153 | 0.1767 | 0.3428 | 0.2044 | 0.3096 |
| myh6-2 | 0.0506 | 0.2660 | 0.0552 | 0.4116 | 0.1442 | 0.3747 | 0.2219 | 0.1375 | 0.1301 | 0.1619 |
| myh6-3 | 0.0839 | 0.3966 | 0.1404 | 0.0172 | 0.3217 | 0.1678 | 0.3382 | 0.2758 | 2.9230 | 2.4993 |
| RYR3-1 | 0.1997 | 0.1627 | 0.0988 | 0.0717 | 0.3722 | 0.2961 | 0.1766 | 0.3384 | 0.3025 | 0.5128 |
| RYR3-2 | 0.0499 | 0.3284 | 0.0157 | 0.3639 | 0.1812 | 0.4051 | 0.1663 | 0.1238 | 0.2454 | 0.3727 |
| RYR3-3 | 0.0909 | 0.3623 | 0.1071 | 0.0111 | 0.3903 | 0.1355 | 0.3379 | 0.3619 | 1.7506 | 3.3205 |
| Ptr-1 | 0.1406 | 0.2433 | 0.0583 | 0.0168 | 0.4585 | 0.3355 | 0.1875 | 0.2310 | 0.4141 | 0.2293 |
| Ptr-2 | 0.1345 | 0.4321 | 0.0567 | 0.2608 | 0.0992 | 0.2762 | 0.1933 | 0.1408 | 0.2994 | 0.1456 |
| Ptr-3 | 0.0846 | 0.4030 | 0.1381 | 0.0269 | 0.3089 | 0.1149 | 0.4174 | 0.2850 | 2.6673 | 2.4090 |
| tbr1-1 | 0.1798 | 0.3229 | 0.1150 | 0.0672 | 0.2510 | 0.2316 | 0.2093 | 0.3433 | 0.3303 | 0.3971 |
| tbr1-2 | 0.1238 | 0.1900 | 0.0523 | 0.3634 | 0.1106 | 0.2291 | 0.4051 | 0.2063 | 0.2569 | 0.1662 |
| tbr1-3 | 0.0895 | 0.3815 | 0.1298 | 0.0726 | 0.2399 | 0.1549 | 0.3971 | 0.2831 | 1.7164 | 1.8345 |
| ENC1-1 | 0.1558 | 0.1168 | 0.0361 | 0.0315 | 0.5831 | 0.2524 | 0.2576 | 0.3182 | 0.2583 | 0.2106 |
| ENC1-2 | 0.1579 | 0.3591 | 0.0374 | 0.1761 | 0.2454 | 0.3115 | 0.2129 | 0.1806 | 0.1001 | 0.0580 |
| ENC1-3 | 0.0640 | 0.3225 | 0.1612 | 0.0252 | 0.3654 | 0.1241 | 0.3825 | 0.3347 | 2.2789 | 2.4439 |
| Gylt-1 | 0.1426 | 0.1985 | 0.1295 | 0.0643 | 0.3766 | 0.3009 | 0.2361 | 0.3133 | 0.4711 | 0.5441 |
| Gylt-2 | 0.0635 | 0.3615 | 0.0690 | 0.2342 | 0.2264 | 0.3254 | 0.2184 | 0.1588 | 0.2407 | 0.3698 |
| Gylt-3 | 0.0651 | 0.3058 | 0.1719 | 0.0267 | 0.3826 | 0.1726 | 0.3138 | 0.3459 | 2.7511 | 3.0404 |
| SH3PX3-1 | 0.1958 | 0.2066 | 0.2194 | 0.0287 | 0.1992 | 0.2783 | 0.2818 | 0.2636 | 0.2019 | 0.3483 |
| SH3PX3-2 | 0.1026 | 0.3464 | 0.0519 | 0.3587 | 0.1276 | 0.3698 | 0.1940 | 0.1650 | 0.1529 | 0.1636 |
| SH3PX3-3 | 0.0747 | 0.3388 | 0.1743 | 0.0101 | 0.3672 | 0.0938 | 0.4093 | 0.3696 | 1.3703 | 3.2724 |
| plagl2-1 | 0.1366 | 0.3133 | 0.1635 | 0.0055 | 0.3149 | 0.2790 | 0.2915 | 0.2572 | 0.3190 | 0.3504 |
| plagl2-2 | 0.2023 | 0.2048 | 0.0161 | 0.3840 | 0.1442 | 0.3455 | 0.2245 | 0.2081 | 0.3660 | 0.1182 |
| plagl2-3 | 0.0717 | 0.3906 | 0.1879 | 0.0154 | 0.2968 | 0.0949 | 0.3947 | 0.3561 | 1.8646 | 2.4753 |
| sreb2-1 | 0.2133 | 0.1440 | 0.0245 | 0.0114 | 0.5885 | 0.2491 | 0.2431 | 0.2747 | 0.1804 | 0.1706 |
| sreb2-2 | 0.0567 | 0.0716 | 0.0830 | 0.4465 | 0.3052 | 0.1791 | 0.2370 | 0.2024 | 0.1001 | 0.0206 |
| sreb2-3 | 0.0724 | 0.3239 | 0.1390 | 0.0264 | 0.3746 | 0.0929 | 0.4475 | 0.3234 | 1.3789 | 2.0777 |

APPENDIX 3. Parameters of 30 partitions estimated using Bayesian method and GTR+Γ model.

| Partitions | Substitution rates | | | | | Base frequencies | | | Alpha | Multiplier |
|---|---|---|---|---|---|---|---|---|---|---|
| | AC | AG | AT | CG | CT | A | C | G | | |
| zic1-1 | 0.1858 | 0.2723 | 0.1158 | 0.0205 | 0.3196 | 0.2194 | 0.2609 | 0.3532 | 0.1801 | 0.4900 |
| zic1-2 | 0.0988 | 0.0820 | 0.0322 | 0.7063 | 0.0578 | 0.2971 | 0.3129 | 0.2049 | 0.1902 | 0.0565 |
| zic1-3 | 0.0690 | 0.3513 | 0.0816 | 0.0447 | 0.3863 | 0.1255 | 0.4225 | 0.2969 | 1.4395 | 1.4379 |
| myh6-1 | 0.0742 | 0.0686 | 0.0719 | 0.0214 | 0.5713 | 0.3135 | 0.2050 | 0.3334 | 0.2476 | 0.6342 |
| myh6-2 | 0.0534 | 0.2933 | 0.0573 | 0.3956 | 0.1361 | 0.3542 | 0.2231 | 0.1513 | 0.1605 | 0.3246 |
| myh6-3 | 0.0940 | 0.3731 | 0.1213 | 0.0227 | 0.3487 | 0.1867 | 0.2921 | 0.2856 | 2.9869 | 1.9368 |
| RYR3-1 | 0.2086 | 0.1857 | 0.0941 | 0.0793 | 0.3383 | 0.2905 | 0.1866 | 0.3096 | 0.3396 | 0.4495 |
| RYR3-2 | 0.0489 | 0.3350 | 0.0197 | 0.3231 | 0.1995 | 0.4130 | 0.1910 | 0.1325 | 0.2677 | 0.4003 |
| RYR3-3 | 0.0908 | 0.3672 | 0.0664 | 0.0239 | 0.4063 | 0.1781 | 0.3086 | 0.3100 | 1.7262 | 2.6833 |
| Ptr-1 | 0.1343 | 0.2491 | 0.0652 | 0.0175 | 0.4441 | 0.3247 | 0.2123 | 0.2379 | 0.4299 | 0.2085 |
| Ptr-2 | 0.1312 | 0.4234 | 0.0549 | 0.2648 | 0.1025 | 0.2858 | 0.1865 | 0.1257 | 0.3442 | 0.0902 |
| Ptr-3 | 0.0752 | 0.3902 | 0.1115 | 0.0410 | 0.3414 | 0.1521 | 0.3776 | 0.2596 | 2.6146 | 2.3452 |
| tbr1-1 | 0.1753 | 0.3161 | 0.1261 | 0.0644 | 0.2509 | 0.2252 | 0.2221 | 0.3504 | 0.3733 | 0.3418 |
| tbr1-2 | 0.1225 | 0.1895 | 0.0581 | 0.3594 | 0.1123 | 0.2250 | 0.4063 | 0.2058 | 0.2535 | 0.0934 |
| tbr1-3 | 0.0878 | 0.3721 | 0.1777 | 0.0545 | 0.2221 | 0.1196 | 0.4492 | 0.2983 | 2.0430 | 1.2163 |
| ENC1-1 | 0.1333 | 0.1348 | 0.0462 | 0.0300 | 0.5647 | 0.2381 | 0.3201 | 0.2687 | 0.2890 | 0.2863 |
| ENC1-2 | 0.1516 | 0.3426 | 0.0385 | 0.1829 | 0.2509 | 0.3262 | 0.1999 | 0.1700 | 0.1075 | 0.0628 |
| ENC1-3 | 0.0586 | 0.3124 | 0.1325 | 0.0315 | 0.3995 | 0.1497 | 0.3607 | 0.3175 | 2.1514 | 2.5027 |
| Gylt-1 | 0.1278 | 0.2057 | 0.1402 | 0.0604 | 0.3702 | 0.3109 | 0.2576 | 0.2937 | 0.5161 | 0.5746 |
| Gylt-2 | 0.0572 | 0.3950 | 0.0683 | 0.2161 | 0.2114 | 0.3442 | 0.2121 | 0.1648 | 0.2644 | 0.5850 |
| Gylt-3 | 0.0700 | 0.3356 | 0.1561 | 0.0302 | 0.3620 | 0.1752 | 0.3128 | 0.3185 | 2.8676 | 2.8350 |
| SH3PX3-1 | 0.1757 | 0.1905 | 0.2310 | 0.0280 | 0.2074 | 0.2790 | 0.3008 | 0.2626 | 0.2337 | 0.3452 |
| SH3PX3-2 | 0.1038 | 0.3412 | 0.0594 | 0.3412 | 0.1314 | 0.3421 | 0.2024 | 0.1709 | 0.1561 | 0.1919 |
| SH3PX3-3 | 0.0751 | 0.3658 | 0.1455 | 0.0117 | 0.3682 | 0.1014 | 0.4130 | 0.3388 | 1.3172 | 3.5048 |
| plagl2-1 | 0.1313 | 0.3316 | 0.1793 | 0.0060 | 0.2876 | 0.2534 | 0.3252 | 0.2559 | 0.3706 | 0.3553 |
| plagl2-2 | 0.1949 | 0.2159 | 0.0210 | 0.3719 | 0.1382 | 0.3427 | 0.2515 | 0.1975 | 0.4314 | 0.1553 |
| plagl2-3 | 0.0625 | 0.3963 | 0.1560 | 0.0195 | 0.3205 | 0.1202 | 0.3892 | 0.3261 | 1.6625 | 2.4051 |
| sreb2-1 | 0.2351 | 0.1780 | 0.0319 | 0.0134 | 0.5203 | 0.2078 | 0.2865 | 0.2704 | 0.1539 | 0.1123 |
| sreb2-2 | 0.0654 | 0.0840 | 0.0910 | 0.4099 | 0.2982 | 0.1766 | 0.2449 | 0.1910 | 0.0828 | 1.4711 |
| sreb2-3 | 0.0681 | 0.3294 | 0.0877 | 0.0470 | 0.3948 | 0.1376 | 0.4170 | 0.2758 | 1.3135 | 1.7771 |