August 2007

A Genome-scale Approach to Phylogeny of Ray-finned Fish (Actinopterygii) and Molecular Systematics of Clupeiformes

Chenhong Li
Univesity of Nebraska, cli@unlserve.unl.edu

A GENOME-SCALE APPROACH TO PHYLOGENY OF RAY-FINNED FISH (ACTINOPTERYGII) AND MOLECULAR SYSTEMATICS OF CLUPEIFORMES

CHENHONG LI, Ph. D.

2007

DISSERTATION TITLE

A Genome-scale Approach to Phylogeny of Ray-finned Fish (Actinopterygii) and

Molecular Systematics of Clupeiformes

BY

Chenhong Li

SUPERVISORY COMMITTEE:

Approved / Date

27 APR 07

Signature

Guillermo Ortí

Typed Name

27 APR 07

Signature

Scott L. Gardner

Typed Name

27 APR 07

Signature

Brett C. Ratcliffe

Typed Name

27 APR 07

Signature

Etsuko Moriyama

Typed Name

Thomas O Powers

27 APR 07

Signature

Thomas O. Powers

Typed Name

Signature

Typed Name

UNIVERSITY OF Nebraska Lincoln

A Genome-scale Approach to Phylogeny of Ray-finned Fish (Actinopterygii) and Molecular Systematics of Clupeiformes

by

Chenhong Li

A DISSERTATION

Presented to the Faculty of

The Graduate College at the University of Nebraska

In Partial Fulfillment of Requirements

For the Degree of Doctor of Philosophy

Major: Biological Sciences

Under the Supervision of Professor Guillermo Ortí

Lincoln, Nebraska

August, 2007

A Genome-scale Approach to Phylogeny of Ray-finned Fish (Actinopterygii) and Molecular Systematics of Clupeiformes

Chenhong Li, Ph. D.
University of Nebraska, 2007

Adviser: Guillermo Ortí

The current trends in molecular phylogenetics are towards assembling large data matrices from many independent loci and employing realistic probabilistic models. Large genome-scale data sets shall reduce the sampling error, whereas complex models accommodating heterogeneity among sites and along the phylogenetic tree can decrease systematic errors. The theme of this dissertation project is using both bioinformatic and experimental approaches to develop genome-scale nuclear gene markers and applying them in studies of phylogeny of ray-finned fish (Actinopterygii) and systematics of clupeiforms. Bioinformatic tools and computer programs were developed to search for conserved single-copy nuclear genes with long exons. By comparing within and between genomes of zebrafish and pufferfish, I have found 138 candidate markers. Ten of fifteen candidates tested were found as good phylogenetic markers, showing similar performance as the popular nuclear marker, recombination activating gene 1 (RAG1). Using the ten newly developed nuclear markers, I conducted a phylogenetic analysis on 52 taxa representing 41 of 44 ray-finned fish orders along with four tetrapods as outgroups. The effects of different data partitioning methods were also tested. Some classic hypotheses about phylogenetic interrelationships of ray-finned fish based on morphological characters were rediscovered in this study, such as the “Holostei” group. In the last two chapters, I present the results of phylogenetic analyses of clupeiforms based on mitochondrial 12S and 16S ribosomal RNA genes, RAG1, RAG2 and six new nuclear loci. Clupeiforms include herrings, anchovies, etc. They have worldwide distribution and important commercial values. The most significant result of the study on clupeiforms is that Clupeidae is not monophyletic. Finally, the last chapter showed that

adding sequences from the six new loci significantly improved the resolution and suggested a different relationship at the basal clupeiods.

Acknowledgments

I would like to thank my advisor, Dr Guillermo Ortí. His help started even before I moved to Lincoln, Nebraska. It was very nice for him to provide my first year research assistantship, so I could have some time to tune up as an international student. Molecular phylogenetics is an academic area requiring strong theoretical background. It was his wisdom from which I benefited most when I started form knowing nothing about this area to being able to think what should be done next. I also learned a lot from him in writing papers and proposals. Thank you, Guillermo. What can be better than having a smart boss?

I am also grateful to the rest of my committee members Dr Scott Gardner, Dr Brett Ratcliffe, Dr Etsuko Moriyama and Dr Thomas Powers. Thanks for all your advices and help that guiding me through the process of my graduate study.

The most memorable part of my time in the lab was from all the people I worked with, although I probably did not spend much time in socializing because I was so busy in finishing my numerous projects. I would like to thank all the following people for their friendship, assistance and inspiration: Agustin Jimenez, Federico Ocampo, Annie Paradis, Corinna Ross, Mike Bessert, Wei-jen Chen, Obdulia Segura Leon, Michelle Steirauer, Chad Brock, Jeremy Brozek, Federico Hoffman, Julie Sommer, Adela Roa Varon, Jason Macrander, Stuart Willis, and TJ Bliss. One guy gave me the most help and became my best friend is Mike. He spent a lot of time to help me with writing and proofreading my grant proposals. He shared with me all his brilliant ideas and enthusiasms in studies of fishes and biology in general. He introduced me to his family and showed me his values. I came from China, but I feel much closer between the two apparently very different cultures because of him. The best memories I have include the walks on the sand bar in western Nebraska, the night drives on the meander road in Missouri, and certainly those wipers we caught at Enders Lake. Thank you, Mike. No matter where I will end up, I would like to keep our friendship and continue to collaborate like what we had for those fruitful side projects.

I also like to thank my parents: my mother, Li, Hua and my father, Li, Jinhao. I grew up at Langxi, a small town 300 km from Shanghai, China. I went to Shanghai

Fisheries University at age sixteen. Since then, I spend most of my time in schools and little time at home. My mom taught me to be always confident in myself. She told me "you can do it if someone else can". She suffered from Rheumatoid arthritis disease for almost 30 years and passed away in 1999. I would like to dedicate this dissertation to her. She would be glad to know her son made this achievement. My dad brought me to fish with him when I was three years old. I am grateful that he kindled my interests in fish and nature.

Finally, I would like to thank my beautiful wife, Zhang, Zhenyu. It is so wonderful that love can tie two person's life together. She quitted her decent job, came with me to United State far away from her family. I cannot get anything done without her supports. It is needless to say who changed the life for the other, but I know she is more important than anything else in my life.

Preface

In the dawn of genomic era, molecular systematics studies are under a transition from typically using a single gene or a few gene markers to seeking genome-scale multiple loci data. The arrangement of this thesis followed the thread of developing new phylogenetic markers and applying them onto the phylogeny of ray-finned fish (Actinopterygii), with an emphasis on interrelationships of Clupeiformes, herrings, anchovies and *etc*.

In the first Chapter, I reviewed the current problems and trends in molecular evolution and systematics. Also, the rational of developing genome-scale nuclear makers was illustrated in this Chapter. In Chapter two, I proposed three criteria for a good phylogenetic marker. The strategy and a computerized tool to develop single-copy nuclear gene markers were the major contributions of this Chapter. Also, results of testing the newly developed markers in fourteen ray-finned fish taxa were reported. Parts of material in the Chapter have been published:

Li, C., Ortí, G., Zhang, G., and Lu, G., A practical approach to phylogenomics: the phylogeny of ray-finned fish (Actinopterygii) as a case study. BMC: Evol. Biol. 7(1), 44.

As Chapter two focused on the development of new markers, in Chapter three, I presented the phylogenetic study of ray-finned fish using ten newly developed nuclear gene markers and 52 taxa representing 41 of 44 orders of ray-finned fishes. Several interesting phylogenetic relationships were found and discussed. In Chapter four, the phylogenetic relationships of Clupeiformes were assessed using both mitochondrial rDNA (12S and 16S) sequences and nuclear recombination activating gene (RAG1 and RAG2) sequences. Some relationships supported by old morphological studies were rediscovered, while deep nodes among some lineages were still unresolved. The results shown in this Chapter have been published in a recent paper:

Li, C., Ortí, G., Molecular phylogeny of Clupeiformes (Actinopterygii) inferred from nuclear and mitochondrial DNA sequences, Mol. Phylogenet. Evol. 44, 386-398

As a follow-up study of Chapter four, more taxa and six more newly developed nuclear gene markers were used to address the interrelationships in Clupeiformes that were not able to be answered by using mitochondrial and RAG genes. The results were summarized in Chapter five.

Besides high-lever (deep) phylogeny in ray-finned fish, my other research interests lie in population genetics and phylogeography of fishes. I have worked on two projects: "Phylogeography of *Prochilodus* (Charaicformes) in South America" and "Conservation genetics of the plains topminnow, *Fundulus sciadicus*". However, I did not write them in this dissertation because of the large volume already included. Out of these two projects, one primer note is in press and two more papers are in preparation:

Li, C., Bessert, M. L., Macrander, J. and Ortí, G., Microsatellite loci for the plains topminnow (*Fundulus sciadicus*, Fundulidae). Molecular Ecology Notes (2007), in press.

Li, C., Bessert, M. L., Macrander, J. and Ortí, G., Conservation genetics of the plains topminnow, (*Fundulus sciadicus*, Fundulidae). in prep.

Ortí, G., Li, C., Farias, I., Vasconcelos, W. R., Lima D. N. E., Saturnino, A., Phylogeny and Population Genetics of *Prochilodus* (Characiformes) based on mtDNA and nuclear intron DNA sequences. in prep.

Table of contents

List of Tables

List of Figures

Chapter 1 - Introduction

1.1. Abstract

In this Chapter, I introduce the major issues in phylogenetic studies: morphological *vs.* molecular data, parsimony *vs.* probabilistic methods, assumptions in likelihood models, analytical and biological systematic errors and data partitioning. I also review the current solutions to address the systematic errors. At the end, I discuss the rational of developing genome-scale nuclear gene markers for phylogenetic analysis.

1.2. Morphological *vs.* molecular data

Understanding phylogeny, the evolutionary relationships of life, is fundamentally important to many aspects of biological studies, such as taxonomy, comparative ecology, genome evolution, etc. Until the late 20th century, the majority data used to infer phylogeny were morphological characters. As the cost of collecting molecular data decreased and the computational capacity was improved, more and more phylogenetic studies included molecular data, especially DNA sequences as their primary data source. Although there is no question about the importance of morphology in understanding adaptation, life history, taxonomy, evolution, etc., the role of morphology in phylogenetic study is controversial (Jenner, 2004; Wiens, 2004; Wortley and Scotland, 2006). In a review of 26 recent studies using both molecular and morphological data, Wortley and Scotland (2006) found that adding morphological data into the analysis did not increase the support for the resulted phylogeny and improved little in the resolution, whereas adding molecular data into the analysis dramatically improved both the support and the resolution of the results.

Both morphological and molecular data have pros and cons as phylogenetic characters, but there are two shortcomings in morphological data constraining it from being the ultimate solution to phylogenetic studies. Firstly, few homologous characters can be found in a wide range of taxa. Many morphological characters are considered as “synapmorphies”, the characters defining a clade, but the homologous counterparts in more diverged taxa are hard to be established, resulting a lot of missing data. For example, it is difficult to find a set of morphological characters that can be used to score all ray-finned fish, a wide range of taxonomic group. Secondly, the total number of potential morphological characters is limited, so morphological characters alone are not enough to resolve many phylogenetic questions. Instead of adding more to the debate of whether one should use morphological data in phylogenetic analysis or not, I would like to point out that the imminent need is to include more informative data in the analysis. Because homologous genes exist in a wide taxonomic range of taxa and the number of potential molecular characters is enormous, developing more independent molecular markers should be the foremost task to facilitate phylogenetic studies, and it is the major goal of this dissertational study.

1.3. Parsimony *vs.* probabilistic methods

The analytical approaches commonly used in current phylogenetic inferences include maximum parsimony (MP) and probabilistic methods, such as maximum likelihood (ML) or Bayesian analysis. The important advantage of probabilistic methods over parsimony is statistically consistent. MP is not consistent, particularly in the case of unequal evolutionary rates between different lineages (Felsenstein, 1978).

Because no explicit models are used in MP method, it is claimed as a “model free” method and immune from model misspecification. But in fact, MP method have been shown always producing the same results as a parameter-rich ML model (Goldman, 1990; Steel and Penny, 2000). The “model freeness” of MP methods does not grant it less error from model misspecification, but rather they are less flexible to accommodate complex data signals. For example, nonstationarity can mislead both MP and probabilistic methods (Foster and Hickey, 1999; Lockhart et al., 1994). Using probabilistic methods, the misleading effects of nonstationarity can be avoided by explicit modeling (Blanquart and Lartillot, 2006; Foster, 2004), while nothing can be changed to rectify the misleading effect from nonstationarity when MP method is used. The relative performance and the connections between MP and ML methods have been hotly debated (Farris, 1983; Felsenstein and Sober, 1986; Goldman, 1990; Kolaczkowski and Thornton, 2004; Sanderson and Kim, 2000; Sober, 2004; Steel, 2005; Steel and Penny, 2000), and no consensus has been reached. In this dissertation, I use mainly the probabilistic method (both ML and Bayesian) and report the results from MP analyses just for comparisons, because probabilistic methods are consistent and flexible to accommodate complex signals in data.

1.4. Probabilistic methods and assumptions

The popular probabilistic methods include ML and Bayesian methods. ML method starts with a model of how the data evolve and calculates the probability of the observed data given the model. The parameters of the model, including the phylogenetic tree, can be optimized by maximizing the probability of the observed data. For a general introduction to ML, see Felsenstein (2004) or Bryant et al. (2005). Because of the large size of tree space and many nuisance parameters, the regular implementation of ML (Swofford, 2003) is not efficient enough to handle large data sets (30 taxa or more). New implementations of ML gain considerable efficiency by not optimizing all parts of each step (Guindon and Gascuel, 2003; Jobb et al., 2004) or by using genetic algorithm (Zwickl, 2006). The Bayesian method combines the prior of parameters with the data to generate the posterior distribution of parameters, upon which all inferences about the parameters are based. The development of Markov chain Monte Carlo (MCMC) algorithms was the computational breakthrough that made the Bayesian method tractable and generally faster than ML method. For a general introduction to Bayesian method, see Yang (2005) or Felsenstein (2004).

Both ML and Bayesian methods involve a hypothetical evolutionary model, which approximates the rules that the evolving sequence characters followed. For DNA sequence, the basic model is composed of the topology of the phylogenetic tree, the branch lengths, stationary nucleotide frequencies and substitution matrix. In reality, too many complicated forces and stochastic processes drive molecular evolution. It is impossible and unnecessary to determine the exact model of molecular evolution. The basic model used in phylogenetic analysis is simplified model based on many assumptions to make them computationally tractable and statistically efficient. There is always a trade-off for complex models. Complex models fit the data better, but it would also have higher sampling errors because more parameters need to be estimated from the data. The basic model works well when the assumptions are met. Below, I list most if not all assumptions made in the basic models:

1. The evolution of characters follows a Markov model with Poisson distribution, but some evidence suggested the overdispersed point process fits the data better (Gillespie, 1994).

2. Each site evolves independently and according to the identical process, so called “i.i.d.” process. This is an unrealistic assumption. Some sites interact functionally with each other may be correlated. Different sites do not necessarily evolve in the same way.

3. Molecular clock assumption describes the evolutionary rate as constant along the evolutionary process. Most implementations of probabilistic methods assume no molecular clock while some enforce strict molecular clock. In reality, the behavior of the evolutionary rate should be in between the two extremes.

4. Stationarity and time reversibility. Stationarity and time reversibility assure the expected frequencies of the nucleotides or amino acids are constant along the evolutionary pathway.

All these assumptions are made to facilitate the likelihood calculation and improve the efficiency of the models. However if the assumptions are violated, using these models will lead to inconsistency, so called model misspecification. Thus, more parameters need to be introduced into the models to reduce the systematic errors.

1.5. Analytical systematic errors and improved models

When the assumptions are not held and the model cannot account for the confounding signals in the data, the inferred results may become inconsistent and erroneous. I call this type of errors as analytical systematic errors, because the errors are caused by model misspecification. Below, I discuss the types of analytical systematic errors and the assumptions being violated. I also review the improved models that have been proposed to relax the assumptions (Fig. 1.1).

When the assumption of stationarity is not held, that is the nucleotide (or amino acid) frequencies changed along the evolutionary pathway, the phylogenetic inference could be misled (Foster, 2004; Foster and Hickey, 1999; Steel et al., 1993). For example, it was found that the high GC bias in the recombination activating protein 1 (RAG1) gene of Clupeiformes and Elopeiformes artifactually grouped them together (Orti et al., unpublished data) in spite of other molecular and morphological evidences indicating that they are not closely related (Lecointre and Nelson, 1996). One easy way to reduce the systematic error from GC bias is to recode the data. For example, RY coding (code A and G as R, C and T as Y) can homogenize the base composition and remove the GC bias (Phillips et al., 2004; Woese et al., 1991), but it cannot remove the more general base compositional bias and may also lose some phylogenetic information. The better way is to account the nonstationarity in the model explicitly. A series of models has been proposed including a distance method (Lockhart et al., 1994), likelihood methods assigning local base frequencies to each branch (Galtier and Gouy, 1998; Yang and Roberts, 1995), and Bayesian methods assigning different base frequencies to predefined number of clades (Foster, 2004). However, the methods assigning base frequencies to branches or clades associate the change of base frequencies with speciation events, which is not realistic. Blanquart et al. (2006) proposed a new model that employing a compound stochastic process, that is the variation of base frequencies also is driven by a stochastic process. Their method is more reasonable, because it decouples the change of base frequencies from speciation events and also reduces the number of parameters to estimate.

When the assumption of molecular clock is not held, that is, the substitution rates are varied along the tree, heterogeneity of the rates has to be considered in the model. In most common implementations, no molecular clock is enforced (Felsenstein, 2005; Ronquist and Huelsenbeck, 2003; Swofford, 2003), and each branch is allowed to have a different rate. However, the model would be overparameterized if no constrains are imposed on the rate variation. Hence, autocorrelated relaxed-clock models have been devised based on the assumption that the rate for a branch is correlated to its adjacent branches (Sanderson, 1997). Recently, an uncorrelated relaxed-clock model was proposed, which does not assume the rate correlation among different lineages, but the correlation can be detected from the data if it exists (Drummond et al., 2006). The other advantage of the uncorrelated relaxed-clock model is that it can optimize the rate and the phylogeny simultaneously, which cannot be done by using the autocorrelated models.

Until now, I only focus on how to model the molecular evolution at single site. The likelihood of observing the data would be the product of likelihoods of all individual sites calculated using the same model, if all sites follow the “i.i.d.” process. However, in reality, different sites could have different rates, substitution matrix and even different stationary frequencies. When the rate is heterogeneous among different sites, among site rate variation (ASRV) model (Yang, 1994) and invariable sites model (Churchill et al., 1992) often can increase the likelihood significantly. When the rates are not only varied among site but also along the tree, they can mislead both MP and ML inference and the process is called *covarion* (for Concomitantly VARiable codON), heterotachy or site-specific rate variation (Fitch, 1971; Lopez et al., 2002). Existing models addressing the conundrum of heterotachy are simple covarion models, which assume a compound process of evolution, so called *Markov-modulated Markov processes* or *Cox processes* (Fitch, 1971; Galtier, 2001; Galtier and Jean-Marie, 2004; Tuffley and Steel, 1998). In the covarion model, the rate of substitution is also modeled as Markov processes so that the rate can stochastically take values from a discrete rate space. The new uncorrelated relaxed-clock model (Drummond et al., 2006) mentioned above is also a promising direction to solve the problem of heterotachy (Pybus, 2006).

Besides the evolutionary rate, the substitution matrix and stationary frequencies can also vary among sites. For example, some sites of the molecule may have different base composition from other sites (Gowri-Shankar and Rattray, 2006). A Gaussian process model has been proposed to account for the compositional variation among sites (Gowri-Shankar and Rattray, 2006). Especially when multiple gene sequences are analyzed concatenately, each gene or codon position may have different evolutionary properties. In this case, dividing the data into partitions and allowing each data partition to has its own model would increase the likelihood (Brandley et al., 2005), and this kind of models are termed as mixed models. Naturally, concatenated multiple gene data can be partitioned by genes and by codon positions. However, if some partitions are similar to each other, assigning separate models for each partition may become overparameterized. In the other hand, if there is still heterogeneity within each “nature” partitions (by genes or codon positions), the mixed model is underparameterized. Another different strategy dealing with heterogeneity among sites is the mixture model (Lartillot and Philippe, 2004; Pagel and Meade, 2004). In the mixture model, no predefined partition is required. The likelihood for each site is calculated for a number of models and then summed up with a weight for each model. The mixture model does not need predefined partitions, because it can detect the heterogeneous evolutionary patterns from the data themselves. The mixture model also has no risk of overparameterizing, because the number of models can be chosen by the data (Pagel and Meade, 2005)..

1.6. Biological systematic errors

If the model used can sufficiently describe the data, there will be less error resulted from the model misspecification. However, phylogenetic inferences may still be confounded by another type of errors that are caused by the discrepancy between the gene genealogy and organismal phylogeny. I call them biological systematic errors. For example, paralogy (Maddison, 1997), incomplete lineage sorting (Funk and Omland, 2003; Maddison, 1997; Maddison and Knowles, 2006) and horizontal gene transfer (Kurland et al., 2003) can all led to inconsistent results. To identify the biological systematic errors, one can resolve the speciation and other confounding events simultaneously (Page and Cotton, 2002) or include data from more individuals or more gene markers to unveil the phylogenetic signals (Maddison and Knowles, 2006).

1.7. Genome-scale data and the "super model"

To reduce the random as well as systematic errors, data from many independent loci are needed. Genome-scale data, including complex genome-level characters (such as gene content and gene order) and sequences from many independent gene loci, provide great potential to sort out the nonphylogenetic noise and recover the true phylogenetic signals. With a large number of characters, the stochastic errors associated with the estimations should decrease (Delsuc et al., 2005). Using many independent nuclear genes can also reduce some systematic errors (Collins et al., 2005; Maddison and Knowles, 2006; Poe and Swofford, 1999). As discussed above, more complicated models would fit the data better and alleviate the misleading effects from analytical systematic errors. However, the complicated models are only useful when there are enough data to estimate the large number of parameters. Thus, including a large number of genome-scale data is not only beneficial but also necessary for using more realistic models. Genome-scale phylogenetics or phylogenomics was criticized as not immune from systematic errors (Kelchner and Thomas, 2006; Soltis et al., 2004), but these conclusions were based on analyses using underparameterized models.

To avoid the biological systematic error, using many independent genome-scale data is one of the solutions, such as inferring phylogeny despite incomplete lineage sorting (Maddison and Knowles, 2006). In the light of genome-scale sequence data, the future complex model, the “super model” should incorporate all complex data structure and confounding signals, such as the variation of base composition and rates among sites and along the tree (Fig. 1.1). The “super model” should be always tested as the null model. Then, the “super model” or reduced models can be selected by using AIC or BIC model selection approaches (Posada and Buckley, 2004).

In this dissertational work, I describe a new tool to develop genome-scale nuclear gene markers. I used the newly developed markers to infer the phylogeny of Ray-finned fish (Actinopterygii) and the interrelationships among clupeiforms. I discussed the potential base compositional bias in Chapter two, Chapter four and Chapter five. I explored the RY coding method to reduce the error form compositional bias. I tested different partitioning schemes and proposed a novel partitioning approach in Chapter three.

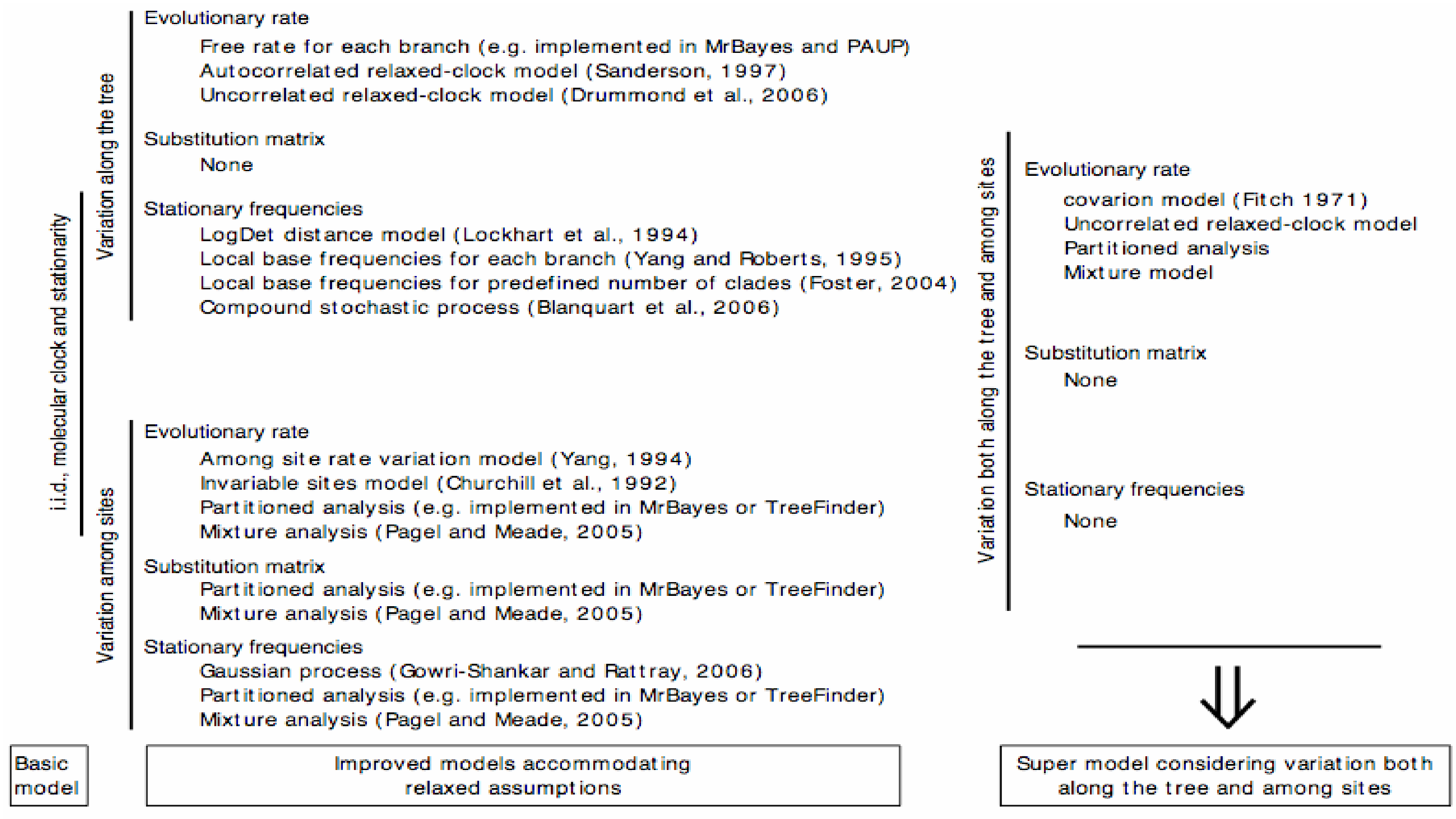
i.i.d., molecular clock and stationarity
Variation along the tree
Evolutionary rate
Free rate for each branch (e.g. implemented in MrBayes and PAUP)
Autocorrelated relaxed-clock model (Sanderson, 1997)
Uncorrelated relaxed-clock model (Drummond et al., 2006)
Substitution matrix
None
Stationary frequencies
LogDet distance model (Lockhart et al., 1994)
Local base frequencies for each branch (Yang and Roberts, 1995)
Local base frequencies for predefined number of clades (Foster, 2004)
Compound stochastic process (Blanquart et al., 2006)
Variation among sites
Evolutionary rate
Among site rate variation model (Yang, 1994)
Invariable sites model (Churchill et al., 1992)
Partitioned analysis (e.g. implemented in MrBayes or TreeFinder)
Mixture analysis (Pagel and Meade, 2005)
Substitution matrix
Partitioned analysis (e.g. implemented in MrBayes or TreeFinder)
Mixture analysis (Pagel and Meade, 2005)
Stationary frequencies
Gaussian process (Gowri-Shankar and Rattray, 2006)
Partitioned analysis (e.g. implemented in MrBayes or TreeFinder)
Mixture analysis (Pagel and Meade, 2005)
Variation both along the tree and among sites
Evolutionary rate
covarion model (Fitch 1971)
Uncorrelated relaxed-clock model
Partitioned analysis
Mixture model
Substitution matrix
None
Stationary frequencies
None
Basic model
Improved models accommodating relaxed assumptions
Super model considering variation both along the tree and among sites

Fig. 1.1 Complexity in molecular evolution and models proposed to accommodate it. The “super model” should consider the variation in rates, substitution matrices and stationary base frequencies both among sites and along the phylogenetic tree.

Chapter 2 - A practical approach to phylogenomics: the phylogeny of ray-finned fish (Actinopterygii) as a case study

2.1. Abstract

Molecular systematics occupies one of the central stages in biology in the genomic era, ushered in by unprecedented progress in DNA technology. The inference of organismal phylogeny is now based on many independent genetic loci, a widely accepted approach to assemble the tree of life. Surprisingly, this approach is hindered by lack of appropriate nuclear gene markers for many taxonomic groups especially at high taxonomic level, partially due to the lack of tools for efficiently developing new phylogenetic makers. I report here a genome-comparison strategy for identifying nuclear gene markers for phylogenetic inference and apply it to the ray-finned fishes - the largest vertebrate clade in need of phylogenetic resolution.

A total of 138 candidate markers were obtained by comparing whole genome sequences of two model organisms, zebrafish (*Danio rerio)* and Japanese pufferfish (*Takifugu rubripes*). Experimental tests of 15 randomly sampled markers on 50 taxa representing nearly all of the ray-finned fish orders demonstrate that ten of these candidates are easily amplified by PCR from whole genomic DNA extractions in a vast diversity of fish taxa. The phylogeny of 14 taxa inferred from concatenated sequences of ten markers (total of 7,872bp) showed large congruencies with the consensus view of the fish phylogeny except for two discrepancies.

I developed a practical approach that compares whole genome sequences to identify single-copy nuclear gene markers for inferring phylogeny. Compared to traditional approaches (manually picking genes for testing), my methods use genomic information and automate the process to identify larger number and genome-scale candidate makers. The approach shown here to be successful for fishes could be applied

to other groups of organisms for which two or more complete genome sequences exist, which has important implications for assembling the tree of life.

2.2. Background

The ultimate goal of obtaining a well-supported and accurate representation of the tree of life relies on the assembly of phylogenomic data sets for large numbers of taxa (Delsuc et al., 2005). Molecular phylogenies based on DNA sequences of a single locus or a few loci often suffer from low resolution and marginal statistical supports due to limited character sampling. Individual gene genealogies also may differ from each other and from the organismal phylogeny (gene-tree vs. species-tree issue) (Fitch, 1970; Pamilo and Nei, 1988), and in many cases this is due to systematic biases leading to statistical inconsistency in phylogenetic reconstruction (i.e., compositional bias, long-branch attraction, heterotachy) (Felsenstein, 1978; Foster and Hickey, 1999; Lopez et al., 2002; Weisburg et al., 1989). Phylogenomic data sets—using genome sequences to study evolutionary relationship—provide the best solution to these problems (Delsuc et al., 2005; Eisen and Fraser, 2003). This solution requires compilation of large data sets that include many independent nuclear loci for many species (Bapteste et al., 2002; Driskell et al., 2004; Murphy et al., 2001; Philippe et al., 2004; Rokas et al., 2003b; Takezaki et al., 2003). Such data sets are less likely to succumb to sampling and systematic errors (Rokas et al., 2003b) by offering the possibility to focus on more phylogenetically reliable characters and also of corroborating phylogenetic results by varying the species sampled. Most attempts to use this approach have been based either on available complete genomic sequence data (Chen et al., 2004; Rokas et al., 2005; Rokas et al., 2003b), or cDNA and ESTs sequences (Bapteste et al., 2002; Philippe et al., 2004; Rokas et al., 2005; Whittall et al., 2006) for relatively few taxa. Availability of complete genomes limits the number of taxa that can be analyzed (Chen et al., 2004; Rokas et al., 2003b), imposing known problems for phylogenetic inference associated with poor taxon sampling (Hillis et al., 2003; Soltis et al., 2004). On the other hand, methods based on ESTs or cDNA sequence data are not practical for many taxa because they require construction of DNA libraries and fresh tissue samples. In addition, some genes may not be expressed in certain tissues or developmental stages, leading to cases with undesirable amounts of missing data (Philippe et al., 2004). The most efficient way to collect nuclear gene sequences for many taxa is to directly amplify target sequences using “universal” PCR primers, an approach so far used for just a few widely-used nuclear genes (Groth and Barrowclough, 1999; Lovejoy and Collette, 2001; Mohammad-Ali et al., 1995; Saint et al., 1998), or selected taxonomic groups (e.g., placental mammals and land plants). Widespread use of this strategy in most taxonomic groups has been hindered by the paucity of available PCR-targeted gene markers.

Mining genomic data for phylogenetic studies requires stringent criteria, since not all loci are likely to carry desired levels of historical signal. The phylogenetic informativeness of characters has been extensively debated on theoretical grounds (Lyons-Weiler et al., 1996; Philippe et al., 2005b), as well as in empirical cases (Collins et al., 2005; Phillips et al., 2004; Steel et al., 1993). My study does not intend to contribute to this debate, but rather to focus on the practical issues involved in obtaining the raw data for analysis. What is the best strategy to select a few hundreds candidate loci from thousands of genes present in the genome? For practical purposes, a good phylogenetic nuclear gene marker must satisfy three criteria. First, orthologous genes should be easy to identify and amplify in all taxa of interest. One of the main problems associated with nuclear protein-coding genes used to infer phylogeny is uncertainty about their orthology (Fitch, 1970). This is especially true when multiple copies of a target gene are amplified by PCR from whole genomic DNA. To minimize the chance of sampling paralogous genes among taxa (the trap of “mistaken paralogy” that will lead to gene-tree-species-tree discordance), my approach is initiated by searches for single-copy nuclear genes in genomic databases. Under this criterion, even if gene duplication events may have occurred during evolution of the taxa of interest (e.g., the fish-specific whole-genome duplication event) (Amores et al., 1998; Meyer and Van de Peer, 2005), duplicated copies of a single-copy nuclear gene tend to be lost quickly, possibly due to dosage compensation (Ciccarelli et al., 2005). Some authors estimate that almost 80% of the paralogs have been secondarily lost following the genome-duplication event (Jaillon et al., 2004; Woods et al., 2005). Thus, if duplicated copies are lost before the relevant speciation events occur (Fig. 2.1a, b), no paralogous gene copies would be sampled. If the alternative situation occurs (Fig. 2.1c) paralogy will mislead phylogenetic inference. In the latter case, the distribution of this discordance is, however, not expected to influence all genes in the same way (i.e., it should not lead to systematic error when many genes are analyzed). The second criterion that will facilitate data collection is to identify protein-coding genes with long exons (longer than a practical threshold determined by current DNA sequencing technology, for example 800 bp). Most genes are fragmented into small exons and large introns. For high taxonomic-level phylogenetic inference (deep phylogeny), intron sequences evolve too fast and are usually not informative, becoming an obstacle for the amplification and sequencing of more informative exon coding sequences. The third criterion is to identify reasonably conserved genes. Genes with low rates of evolution are less prone to homoplasy, and also provide the practical advantage of facilitating the design of universal primers for PCR that will work on a diversity of taxa. Usually, conserved protein-coding genes also are easy to align for analysis, based on their amino acid sequences.

Sequence conservatism and long exonic regions have been used as the criteria to choose phylogenetics markers in the past (Friedlander et al., 1992). However, the probability of finding a reliable, easy-to-apply gene marker would be very small if genes are haphazardly selected for study. This complexity partially explains the scarcity of currently available nuclear gene markers in many taxonomic groups. To address this problem, I developed a simple approach to obtain nuclear gene markers based on the three aforementioned criteria using both bioinformatic and experimental methods. My method incorporates two improvements over the traditional way of manually picking genes and testing their phylogenetic utilities. These improvements include using genomic information and automating the process of searching for candidate makers. I apply the method to Actinopterygii (ray-finned fish), the largest vertebrate clade—they make up about half of all known vertebrate species—that has a poorly defined phylogenetic backbone (Arratia, 2000; Greenwood et al., 1973; Miya et al., 2003; Stiassny et al., 1996a; Stiassny et al., 2004).

2.3. Materials and Methods

2.3.1. Genome-scale mining for phylogenetic markers

Whole genomic sequences of *Danio rerio* and *Takifugu rubripes* were retrieved from the ENSEMBL database (http://www.ensembl.org/index.html). Exon sequences with length > 800 bp were then extracted from the genome databases. The exons extracted were compared in two steps: (1) within-genome sequence comparisons and (2) between genome comparisons. The first step is designed to generate a set of single-copy nuclear gene exons (length > 800 bp) within each genome, whereas the second step should identify single-copy, putatively orthologous exons between *D. rerio* and *T. rubripes* (Fig. 2.2). The BLAST algorithm was used for sequence similarity comparison. In addition to the parameters available in the BLAST program, I applied another parameter, coverage (C), to identify global sequence similarity between exons. The coverage was defined as the ratio of total length of locally aligned sequences over the length of query sequence. The similarity (S) was set to S < 50% for within-genome comparison, which means that only genes that have no counterpart more than 50% similar to themselves were kept. The similarity was set to Sx > 70% and the coverage was set to C > 30% in cross-genome comparison, which selected genes that are 70% similar and 30% aligned between *D. rerio* and *T. rubripes*. EST sequences from five additional species (*Gasterosteus aculeatus*, *Ictalurus punctatus*, *Oreochromis niloticus*, *Pagrus auriga* and *Tetraodon nigroviridis*) from the TIGR Gene Indices project (http://www.tigr.org/tdb/tgi/) were used to further select for markers that have no paralogous loci in any of these species (Sx > 70% and C = 30%). Note that this step may not identify all paralogs, since genomic sequences are not complete in these species. The pipelines were automated in PERL language with the help from Dr. Guoqing Lu at University of Nebraska at Omaha.

2.3.2. Experimental testing for candidate markers

PCR and sequencing primers were designed on aligned sequences of *D. rerio* and *T. rubripes* for 15 randomly selected genes. Primer3 was used to design the primers (Rozen and Skaletsky, 2000). Degenerate primers and a nested-PCR design were used to assure the amplification for each gene in most of the taxa. Ten of the 15 genes tested were amplified with single fragment in most of the 50 taxa examined. PCR primers for ten gene markers are listed in Table 2.1. The amplified fragments were directly sequenced, without cloning, using the BigDye system (Applied Biosystems). Sequences of the frequently used RAG1 gene were retrieved for the same taxa from GenBank for comparison to the newly developed markers [GenBank: AY430199, NM_131389, U15663, AB120889, DQ492511, AY308767, AF108420, EF033039 – EF033043]. When RAG1 sequences for the same taxa were not available, a taxon of the same family was used, *i.e. Nimbochromis* was used instead of *Oreochromis* and *Neobythites* was used instead of *Brotula*.

2.3.3. Phylogenetic analysis

In this Chapter, sequences of the ten new markers in 14 taxa were used to assess the performance of these markers for phylogenetic analysis. For analyses and discussions on the phylogeny of ray-finned fish using all 52 taxa with some missing data, see Chapter three. Sequences were aligned using ClustalX (Thompson et al., 1997) on the translated protein sequences. ML corrected genetic distances were calculated using PAUP (Swofford, 2003). Relative substitution rates for each marker was estimated using a Bayesian approach (Ronquist and Huelsenbeck, 2003). Relative composition variability (RCV) and treeness were calculated following Phillips and Penny (Phillips and Penny, 2003). Prottest (Abascal et al., 2005) was used to chose the best model for protein sequence data and the AIC criteria to determine the scheme of data partitioning. Bayesian analysis implemented in MrBayes v3.1.1 and maximum likelihood analysis implemented in TreeFinder (Jobb et al., 2004) were performed on the protein sequences. One million generation with 4 chains were run for Bayesian analysis and the trees sampled prior to reaching convergence were discarded (as burnin) before computing the consensus tree and posterior probabilities. Two independent runs were used to provide additional confirmation of convergence of posterior probability distribution. To reduce the potential effect of biased base composition to the resulted phylogeny, I also analyzed the nucleotide data under the RY-coding scheme (C and T = Y, A and G = R), partitioned by gene in TreeFinder, since RY-coded data are less sensitive to base compositional bias (Phillips and Penny, 2003). Alternative hypotheses were tested by one-tailed Shimodaira and Hasegawa (SH) test (Shimodaira and Hasegawa, 1999) with 1000 RELL bootstrap replicates implemented in TreeFinder.

2.4. Results

The bioinformatic pipeline used is shown in Fig. 2.2. Within-genome sequence comparison resulted in 2,797 putative single-copy exons (> 800 bp) in zebrafish (*D. rerio*) and 2,833 in torafugu (*T. rubripes*). Among them, 154 putative homologs were identified between zebrafish and torafugu by cross-genome comparison. Further comparison with EST sequences from other fish species reduced this number to 138 candidate markers (Appendix A). The candidate markers are distributed among 24 of the 25 chromosomes of zebrafish (Fig. 2.3), and a Chi-square test did not reject a Poisson distribution of the markers among chromosomes (χ^2=16.99, df=10, p=0.0746). The size of candidate markers identified by these search criteria ranged from 802 to 5811 bp (in *D. rerio*). Their GC content ranged from 41.6% to 63.9% (in *D. rerio*), and the average similarity of the DNA sequence of these markers between *D. rerio* and *T. rubripes* varied from 77.3% to 93.2% (determined by the search criteria).

To test the practical value of these candidate markers for phylogenetic inference, 15 candidate markers were randomly chosen and tested experimentally on 52 taxa, representing all ray-finned fish orders except for Saccopharyngiformes, Ateleopodiformes and Stephanoberyciformes (Nelson, 2006). Ten out of the 15 markers tested were successfully amplified by a nested PCR approach in 50 taxa (Table 2.2), and

83% PCR reactions resulted single fragment (see Appendix B). Fourteen representative taxa with all ten genes sequenced (*Amia calva*, *D. rerio, Semotilus atromaculatus, Ictalurus punctatus*, *Oncorhynchus mykiss*, *Brotula multibarbata*, *Fundulus heteroclitus, Oryzias latipes*, *Oreochromis niloticus*, *Gasterosteus aculeatus, Lycodes atlanticus, T. rubripes*, *Morone chrysops*, *Lutjanus mahogoni*) were used to evaluate the ten new markers [GenBank: EF032909 – EF033038]. The size of the sequenced fragments ranged from 666 to 987 bp, while the average genetic distances for DNA sequence (likelihood corrected) of the ten markers among the 14 taxa ranged from 28% to 41% (Table 2.2). Some parameters obtained by phylogenetic analysis of these sequences, such as the substitution rate, consistency index (CI), gamma shape parameter (α), relative composition variability (RCV) and treeness (Phillips and Penny, 2003) of the ten new markers are similar to a commonly used nuclear marker—recombination activating gene 1 (RAG-1, Table 2.2). For the newly obtained phylogenetic markers, the substitution rate is negatively correlated with CI (r = -0.84, P = 0.0026) and marginally correlated with α (r = -0.56, P = 0.095). In contrast, base composition heterogeneity (RCV) and the phylogenetic signal to noise index (treeness index) are not correlated with substitution rate (Fig. 2.4). Based on the treeness value, genes ENC1, plagl2, Ptc and tbr1 are especially recommend for phylogenetic studies at high taxonomic level among ray-finned fishes.

A phylogeny of the 14 taxa using concatenated sequences of all ten markers (total of 7,872 bp) was inferred on the basis of protein and DNA sequences. For the protein sequence data, a JTT model with gamma parameter accounting for rate heterogeneity was selected by Prottest (Abascal et al., 2005). The data were partitioned by gene, as this strategy was favored by the Akaike information criterion (AIC) over treating the concatenated sequences as a single partition. Maximum likelihood (ML) and Bayesian analysis (BA) resulted in the same tree (Fig. 2.5a). A similar topology to Fig. 2.5a was obtained by ML analysis of nucleotide sequences with RY-coded nucleotides to address potential mpractic due to base compositional bias (Phillips and Penny, 2003). The positions of *Brotula* and *Morone* remain somewhat unresolved, receiving low bootstrap support and conflicting resolution based on protein or RY-coded nucleotide data. When analyzed separately, all individual gene trees have low support in many branches and none of them has the same topology as the tree based on all ten genes (Fig. 2.6.). Alternative topologies recovered by individual gene markers were rejected by data combining all ten genes, based on a one-tailed SH test ($p<0.05$), except for the one supported by tbr1 ($p=0.162$) and plagl2 ($p=0.498$). Also, six individual genes (zic1, RYR3, Ptc, tbr1, ENC1 and SH3PX3) rejected the best tree supported by data concatenating ten genes, indicating conflicting signal in individual genes.

2.5. Discussion

The bioinformatic approach implemented in this study resulted in a large set (138 loci) of candidate genes to infer high-level phylogeny of ray-finned fishes. Experimental tests of a smaller subset (15 loci) demonstrate that a large fraction (2/3) of these candidates are easily amplified by PCR from whole genomic DNA extractions in a vast diversity of fish taxa. The assumption that these loci are represented by a single copy in the fish genomes could not be rejected by the PCR assays in the species tested (all amplifications resulted in a single product), increasing the likelihood that the genetic markers are orthologous and suitable to infer organismal phylogeny. My method is based on searching the available complete genomic databases of organisms closely related to the taxa of interest under specific criteria. Therefore, the same approach that is shown to be successful for fishes could be applied to other groups of organisms for which two or more complete genome sequences exist. Parameter values (L, S, and C) used for the search (Fig. 2.2) may be altered to obtain fragments of different size or with different levels of conservation (i.e., less conserved for phylogenies of more closely related organisms).

An alternative way to develop nuclear gene markers for phylogenetic studies is to construct a cDNA library or sequence several ESTs for a small pilot group of taxa, and then to design specific PCR primers to amplify the orthologous gene copies in all the other taxa of interest (Small et al., 2004; Whittall et al., 2006). The major potential problem with this approach stems from the fact that the method starts with a cDNA library or a set of EST sequences, with no prior knowledge of how many copies a gene has in each genome. As discussed above, this condition may lead to mistaken paralogy. In my approach, I search the genomic sequence to find single-copy candidates so no duplicate gene copies, if present, would be missed.

Recent studies have proposed whole genome duplication events during vertebrate evolution and also genome duplications restricted to ray-finned fishes (Amores et al., 1998; Meyer and Van de Peer, 2005; Taylor et al., 2003; Van de Peer et al., 2003). My results indicate that many single-copy genes still exist in a wide diversity of fish taxa (representing 41 orders of actinopterygian fishes), in agreement with previous estimates that a vast majority of duplicated genes are secondarily lost (Jaillon et al., 2004; Woods et al., 2005). All 138 candidates were identified as single-copy genes in *D. rerio* and *T. rubripes*, and out of the 15 tested experimentally, ten were found in single-copy condition in all successful amplifications, including the tetraploid species, *O. mykiss*. My results also show the 138 candidate genes are randomly distributed in the fish genome (at least among chromosomes of *D. rerio*). The existence and identification of genome-scale single-copy nuclear markers should facilitate the construction of the tree of life, even if the evolutionary mechanism responsible for maintaining single-copy genes is poorly known (Ciccarelli et al., 2005).

The molecular evolutionary profiles of the ten newly developed markers are in the same range as RAG-1, a widely used gene marker in vertebrates. The genes with high treeness values have intermediate substitution rate, suggesting that optimal rate and base composition stationarity are important factors that determine the suitability of a phylogenetic marker. The phylogeny based on individual markers revealed incongruent phylogenetic signal among individual genes. This incongruence suggests that systematic error might overrun the true phylogenetic signal in some individual genes, but the direction of the bias is hardly shared among genes (Fig. 2.6), justifying the use of genome-scale gene makers to infer organismal phylogeny.

Finally, with respect to the phylogenetic results *per se*, there are two main discrepancies between the phylogeny obtained in this study (Fig. 2.5a) and a consensus view of fish phylogeny (Fig 2.5b) (Nelson, 2006). Although these differences could be due to poor taxonomic sampling (see Chapter 3), I discuss them briefly. First, the traditional tree groups *O. niloticus* with other perciformes, whereas my results showed the *O. niloticus* is more closely related to Cyprinodontiformes + Beloniformes. This latter result also was supported by two recent studies analyzing multiple nuclear genes (Chen et al., 2004; Steinke et al., 2006). The second difference is that the traditional tree groups *Lycodes* with other Perciformes, while *Lycodes* was found closely related to *Gasterosteus* (Gasterosteiformes) in my results. My observation was supported by the one-tailed Shimodaira-Hasegawa (SH) test (p=0.000) (Shimodaira and Hasegawa, 1999).

2.6. Conclusions

I developed a genome comparison approach that compares whole genome sequences to identify nuclear gene markers that are single copy copies, contain large exons, and are conserved across extensive taxonomic distance for phylogeny inference. I showed that my approach is viable through direct experimentation on a representative sample of ray-finned fish, the largest vertebrate clade in need of phylogenetic resolution. The same approach, therefore, could be applied to other groups of organisms as long as two or more complete genome sequences are available. This research may have important implications for assembling the tree of life.

2.7. Acknowledgement

The bioinformatic pipeline and computer scripts were developed with the help from Dr. Guoqing Lu and his student Gong Zhang at University of Nebraska at Omaha. I am very grateful for their input.

Table 2.1. PCR primers and annealing temperatures used to amplify ten new markers.

Gene*	Primers	Sequences	Annealing temp	PCR steps
zic1	zic1_F9	5’ GGACGCAGGACCGCARTAYC 3’	57	1st PCR
	zic1_R967	5’ CTGTGTGTGTCCTTTTGTGRATYTT 3’		
	zic1_F16	5’ GGACCGCAGTATCCCACYMT 3’	57	2nd PCR
	zic1_R963	5’ GTGTGTCCTTTTGTGAATTTTYAGRT 3’		
myh6	myh6_F459	5’ CATMTTYTCCATCTCAGATAATGC 3’	53	1st PCR
	myh6_R1325	5’ ATTCTCACCACCATCCAGTTGAA 3’		
	myh6_F507	5’ GGAGAATCARTCKGTGCTCATCA 3’	62	2nd PCR
	myh6_R1322	5’ CTCACCACCATCCAGTTGAACAT 3’		
RYR3	RYR3_F15	5’ GGAACTATYGGTAAGCARATGG 3’	55	1st PCR
	RYR3_R968	5’ TGGAAGAAKCCAAAKATGATGC 3’		
	RYR3_F22	5’ TCGGTAAGCARATGGTGGACA 3’	62	2nd PCR
	RYR3_R931	5’ AGAATCCRGTGAAGAGCATCCA 3’		
Ptc	Ptc_F458	5’ AGAATGGATWACCAACACYTACG 3’	55	1st PCR
	Pct_R1248	5’ TAAGGCACAGGATTGAGATGCT 3’		
	Ptc_F463	5’ GGATAACCAACACYTACGTCAA 3’	62	2nd PCR
	Pct_R1242	5’ ACAGGATTGAGATGCTGTCCA 3’		
tbr1	tbr1_F1	5’ TGTCTACACAGGCTGCGACAT 3’	57	1st PCR
	tbr1_R820	5’ GATGTCCTTRGWGCAGTTTTT 3’		
	tbr1_F86	5’ GCCATGMCTGGYTCTTTCCT 3’	62	2nd PCR
	tbr1_R811	5’ GGAGCAGTTTTTCTCRCATTC 3’		
ENC1	ENC1_F85	5’ GACATGCTGGAGTTTCAGGA 3’	53	1st PCR
	ENC1_R982	5’ ACTTGTTRGCMACTGGGTCAAA 3’		
	ENC1_F88	5’ ATGCTGGAGTTTCAGGACAT 3’	62	2nd PCR
	ENC1_R975	5’ AGCMACTGGGTCAAACTGCTC 3’		
Gylt	Glyt_F559	5’ GGACTGTCMAAGATGACCACMT 3’	55	1st PCR
	Glyt_R1562	5’ CCCAAGAGGTTCTTGTTRAAGAT 3’		
	Glyt_F577	5’ ACATGGTACCAGTATGGCTTTGT 3’	62	2nd PCR
	Glyt_R1464	5’ GTAAGGCATATASGTGTTCTCTCC 3’		
SH3PX3	SH3PX3_F461	5’ GTATGGTSGGCAGGAACYTGAA 3’	55	1st PCR
	SH3PX3_R1303	5’ CAAACAKCTCYCCGATGTTCTC 3’		
	SH3PX3_F532	5’ GACGTTCCCATGATGGCWAAAAT 3’	62	2nd PCR
	SH3PX3_R1299	5’ CATCTCYCCGATGTTCTCGTA 3’		
plagl2	plagl2_F9	5’ CCACACACTCYCCACAGAA 3’	55	1st PCR
	plagl2_R930	5’ TTCTCAAGCAGGTATGAGGTAGA 3’		

Table 2.1. PCR primers and annealing temperatures used to amplify ten new markers (cont.).

Gene*	Primers	Sequences	Annealing temp	PCR steps
	plagl2_F51	5’ AAAAGATGTTTCACCGMAAAGA 3’	62	2nd PCR
	plagl2_R920	5’ GGTATGAGGTAGATCCSAGCTG 3’		
sreb2	sreb2_F10	5’ ATGGCGAACTAYAGCCATGC 3’	55	1st PCR
	sreb2_R1094	5’ CTGGATTTTCTGCAGTASAGGAG 3’		
	sreb2_F27	5’ TGCAGGGGACCACAMCAT 3’	62	2nd PCR
	sreb2_R1082	5’ CAGTASAGGAGCGTGGTGCT 3’		

[*] Gene markers are named following annotations in ENSEMBLE. Zic1, zic family member 1; myh6, myosin, heavy polypeptide 6; RYR3, ovel protein similar to vertebrate ryanodine receptor 3; Ptc, hypothetical protein LOC564097; tbr1, T-box brain 1; ENC1, similar to ectodermal-neural cortex 1; Glyt, glycosyltransferase; SH3PX3, SH3 and PX domain containing 3; plagl2, pleiomorphic adenoma gene-like 2; sreb2, super conserved receptor expressed in brain 2.

Table 2.2. Summary information of the ten gene markers amplified in 14 taxa.

Gene	Exon ID	No. of bp	No. of var.	No. of PI	Genetic distance (%)	Sub. rate	CI-MP	α	RCV	Treeness
zic1	ENSDARE00000015655	894	296	210	28(2.6-65.8)	0.64	0.61	1.64	0.13	0.23
myh6	ENSDARE00000025410	735	323	235	36(10.1-59.5)	1.35	0.54	0.68	0.11	0.22
RYR3	ENSDARE00000465292	825	389	258	36(10.1-58.1)	1.25	0.56	0.67	0.11	0.21
Ptc	ENSDARE00000145053	705	304	234	41(6.1-93.6)	1.03	0.57	1.64	0.12	0.29
tbr1	ENSDARE00000055502	666	256	170	28(3.1-79.1)	0.65	0.67	2.91	0.10	0.28
ENC1	ENSDARE00000367269	810	312	248	38(8.4-78.0)	1.13	0.55	1.10	0.16	0.33
Gylt	ENSDARE00000039808	870	463	335	41(7.6-77.0)	1.18	0.60	1.70	0.12	0.27
SH3PX3	ENSDARE00000117872	705	290	226	30(7.5-60.0)	1.11	0.55	1.53	0.14	0.22
plagl2	ENSDARE00000136964	675	250	184	29(6.0-60.6)	0.81	0.61	0.92	0.10	0.33
sreb2	ENSDARE00000029022	987	344	225	30(4.6-75.5)	0.85	0.61	0.88	0.11	0.23
RAG1	-	1344	684	514	38(9.8-75.0)	1.28	0.57	1.68	0.05	0.23

bp, base pairs; var., variable sites; PI, parsimony informative sites; Genetic distance, average ML-corrected distance, number in parenthesis are range of the distances; Sub. rate, relative substitution rate estimated using Bayesian approach; CI-MP, consistency index; α, gamma distribution shape parameter; RCV, relative composition variability.

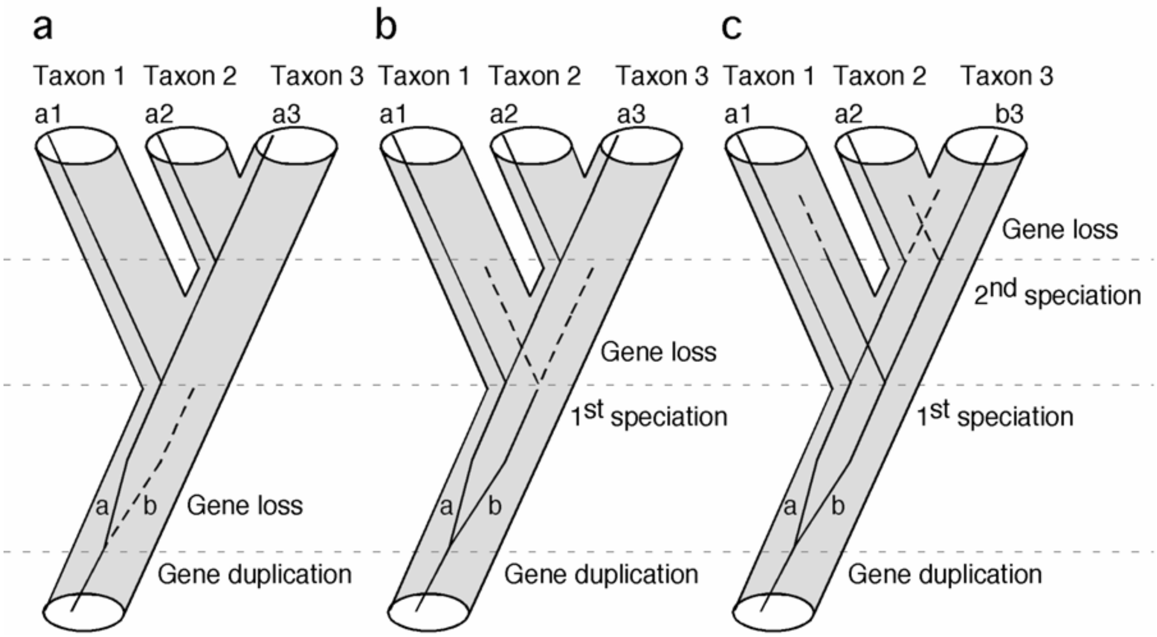

Fig. 2.1 Single-copy genes are useful markers for phylogeny inference. Gene duplication and subsequent loss may not cause incongruence between gene tree and species tree if gene loss occurs before the first speciation event (a), or before the second speciation event (b). The only case that would cause incongruence is when the gene survived both speciation events and is asymmetrically lost in taxon 2 and taxon 3 (c).

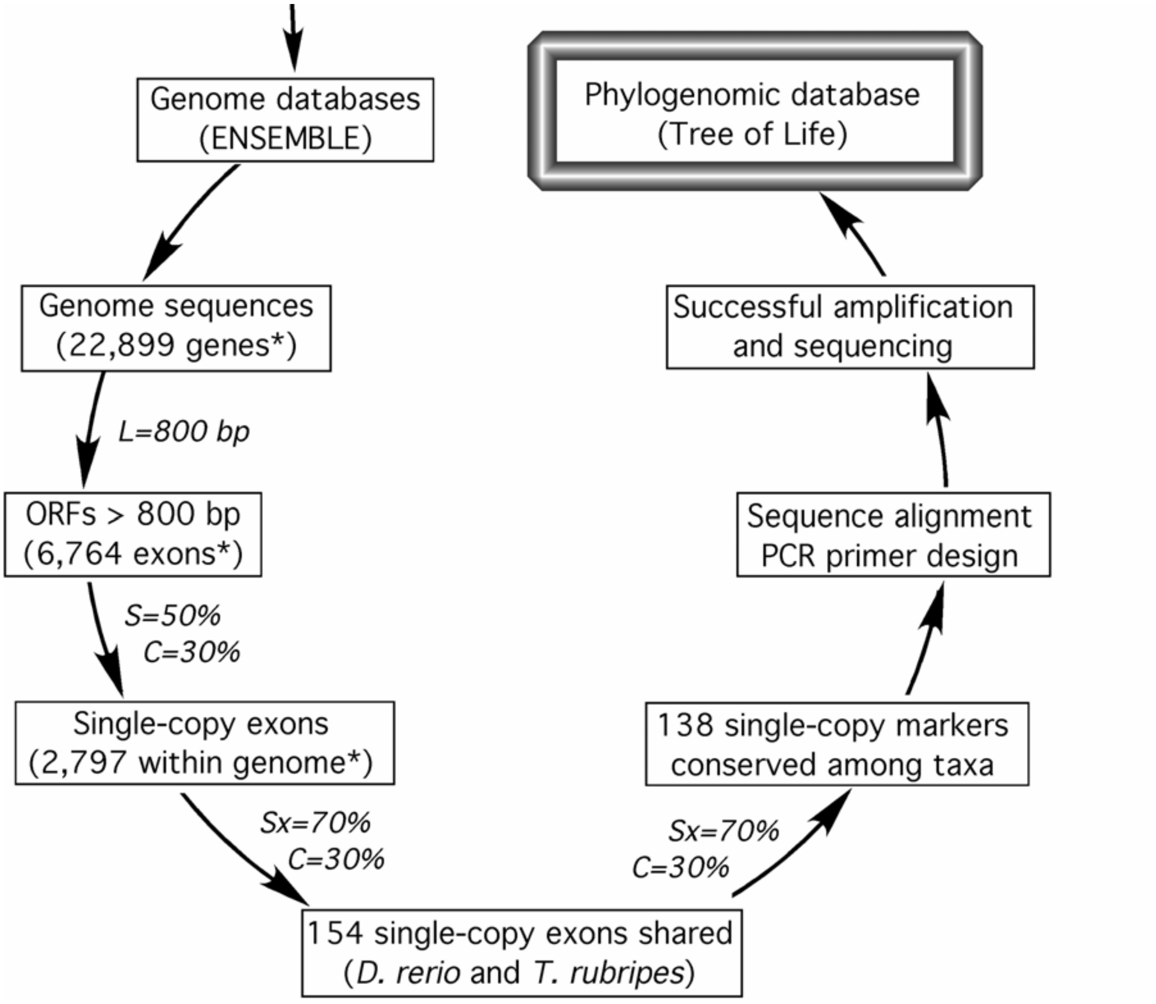

Fig. 2.2 The bioinformatic pipeline for phylogenetic markers development. It involves within- and across-genome sequences comparison, *in silico* test with sequences in other species, and experimental validation. Numbers of genes and exons identified for D. rerio are indicated by the asterisk. Exon length (L), within-genome similarity (S), between-genome similarity (Sx), and coverage I are adjustable parameters (see methods).

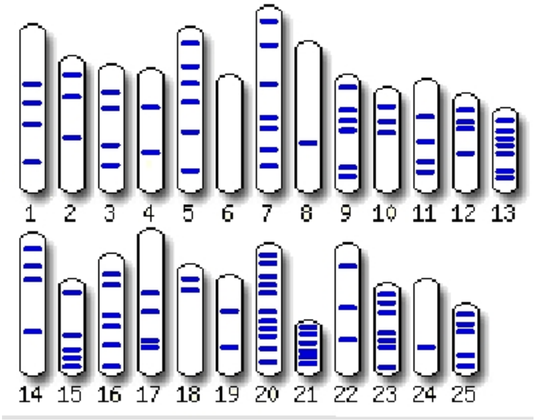

Fig. 2.3. Distribution of the candidate markers on *Danio rerio* chromosomes

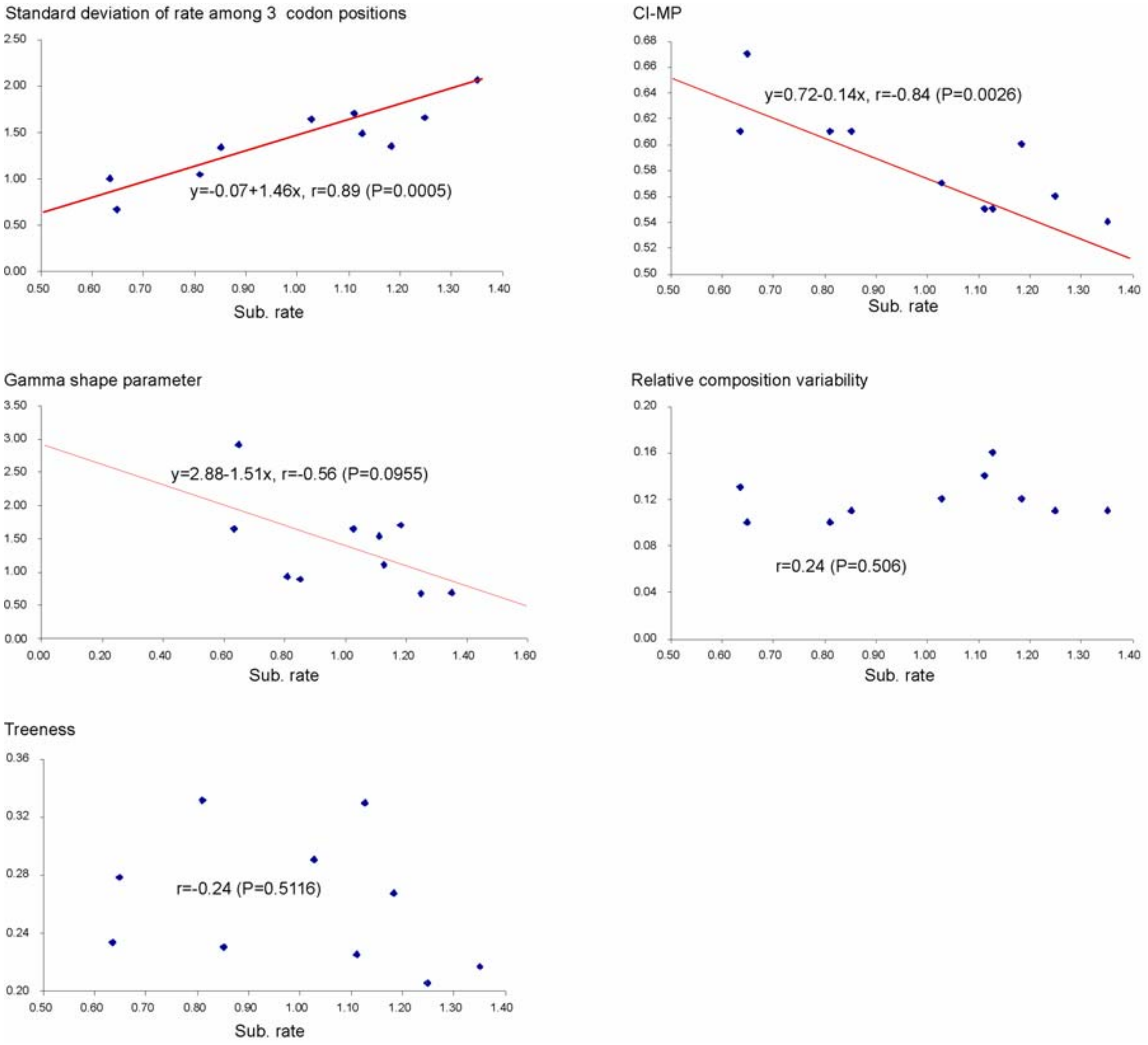

Fig. 2.4. Correlation between gamma shape parameter, SDR, consistency index, relative composition variability, treeness and substitution rate

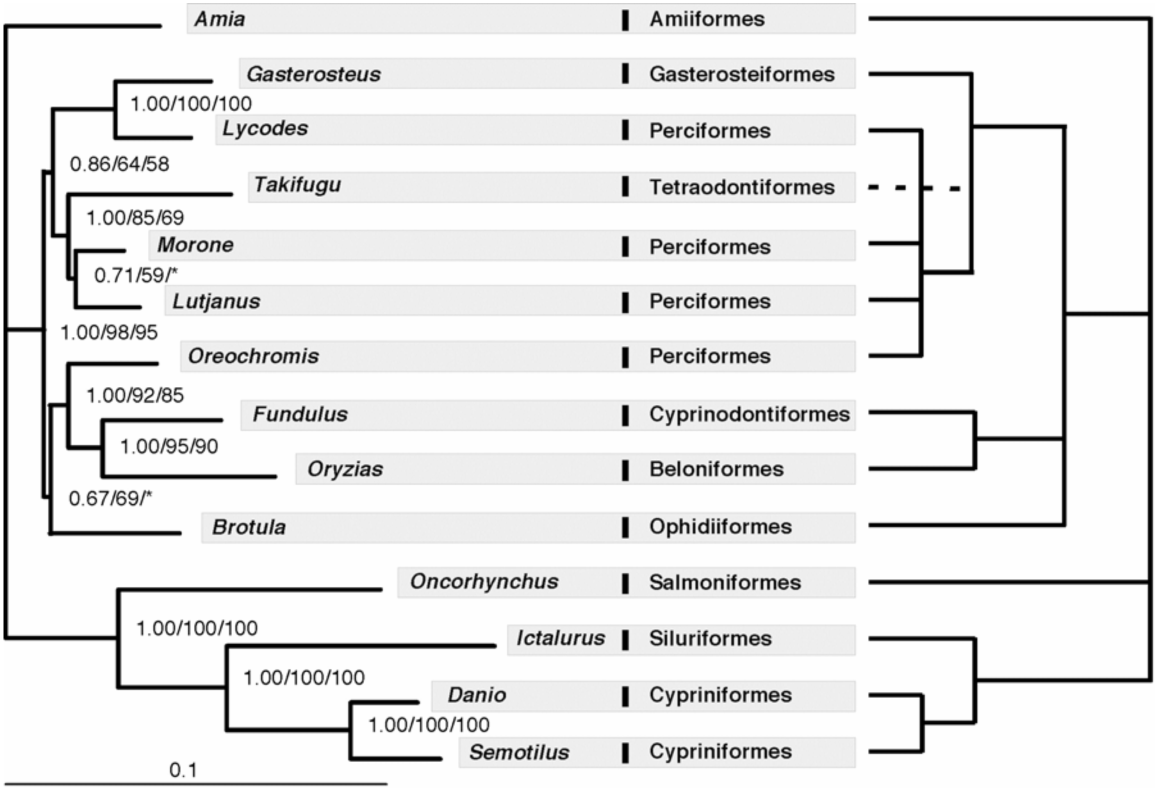

Fig. 2.5. A comparison of the maximum likelihood phylogram inferred in this study with the conventional phylogeny. Right panel – the phylogram of 14 taxa inferred from protein sequences of ten genes; left panel – a “consensus” phylogeny following Nelson (Nelson, 2006). The numbers on the branches are Bayesian posterior probability, ML bootstrap values estimated from protein sequences and ML bootstrap values estimated from RY-coded nucleotide sequence. Asterisks indicate bootstrap supports less than 50.

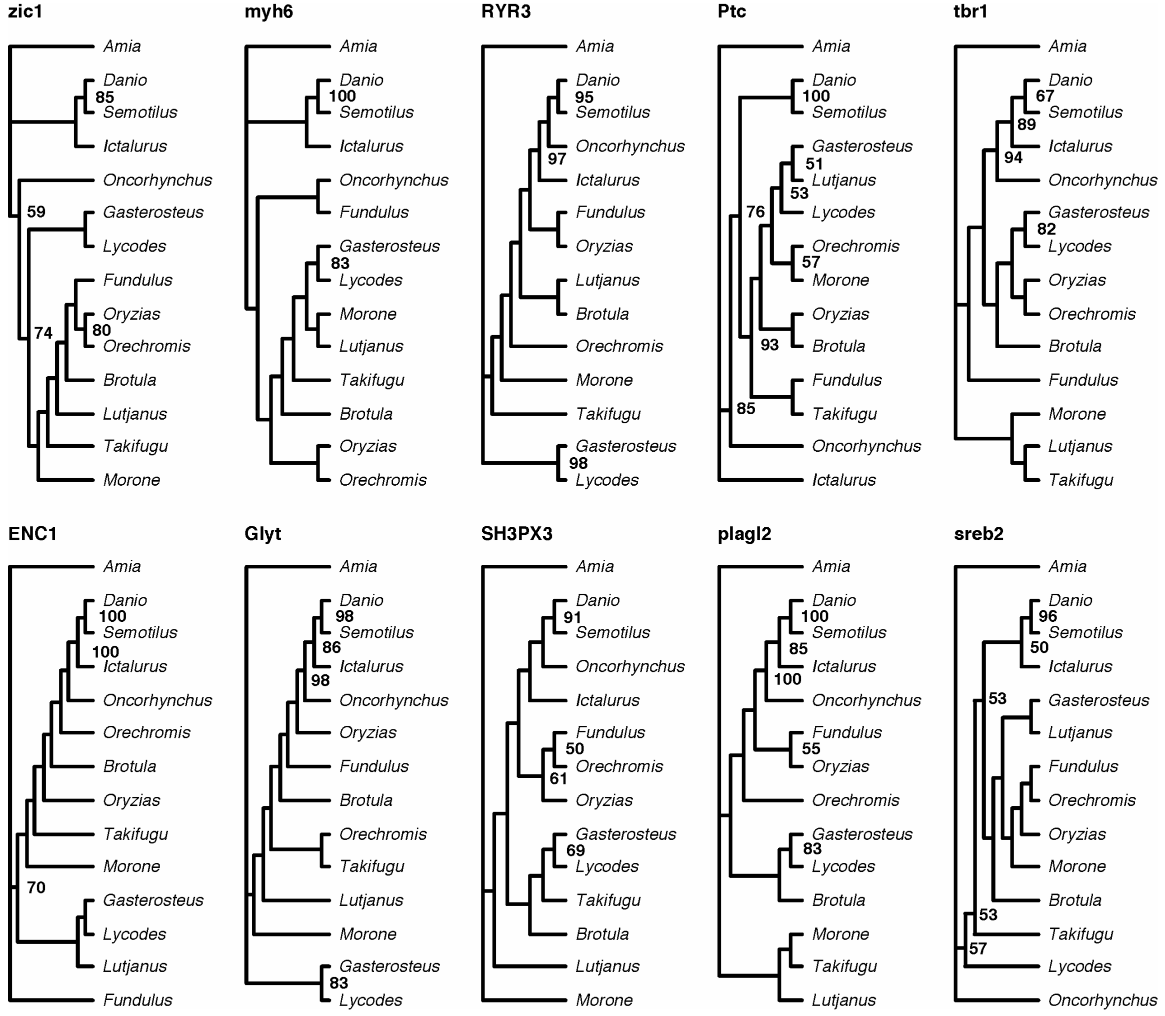

Fig. 2.6 Maximum likelihood phylogeny based on protein sequences of individual genes, zic1, myh6, RYR3, Ptc, tbr1, ENC1, Gylt, SH3PX3, plagl2, and sreb2. Bootstrap value higher than 50% were mapped on branches.

Chapter 3 – Data Partitioning Guided by Cluster Analysis and Phylogeny of Ray-finned Fish (Actinopterygii) Based on Ten Nuclear Loci

3.1. Abstract

Partitioned analysis is one of the best ways to accommodate heterogeneities in evolutionary rates and patterns among sites in molecular phylogenetic analysis. The common ways of data partitioning are dividing data by genes, codon positions, or by both. Partitioning by both genes and codons has high risk of over-parameterizing, although it often result in better likelihood. Reducing the number of partitions by grouping similar data partitions should increase the efficiency of the models. I propose using cluster analysis on model parameters to guide the procedure of data grouping. I tested this strategy using sequence data of ten nuclear genes collected from 52 ray-finned fish (Actinopterygii) and four tetrapods. Concatenating sequences of exons of ten nuclear genes resulted 7995 nucleotide sites. The results showed that most of heterogeneities exist among three codon positions. Reduced number of partitions guided by the cluster analysis performed better than the full 30 partitions by both genes and codon positions indicated by AIC values and Bayes factors. Data partitioning not only affected the fit of the models but also changed the topologies inferred from my data, particularly when Bayesian analysis method was used. The phylogenetic relationships among the major clades of ray-finned fish were assessed using the best data partitioning schemes selected by AIC values and Bayesian factors. Some significant results include the monophyly of "Chondrostei" (polypteriforms + acipenseriforms), the monophyly of "Holostei", elopmorphs as the sister-group to all other extant teleosts, the sister-taxa relationship between esociformes and salmoniforms, a sister-taxa relationship between osmeriforms and stomiforms, a close relationship between lophiiforms and tetraodontiforms, the non-monophyly of protacanthopterygians, the non-monophyly of paracanthopterygians and the non-monophyly of perciforms.

3.2. Background

In the light of genomic era, phylogenetic studies using multilocus sequence data become increasingly popular (e.g. Baurain et al., 2007; Comas et al., 2007; McMahon and Sanderson, 2006; Rokas et al., 2005; Rokas et al., 2003b). The large number of characters and the independent phylogenetic evidences from the multilocus data often resulted in well-resolved and highly supported phylogenies (e.g. Comas et al., 2007; Philippe et al., 2005a; Rokas et al., 2003a). In spite of these successes and the initial optimism about “genome-scale” approach (Gee, 2003; Rokas et al., 2003b), cautions have been called for phylogenetic analysis even when “genome-scale” data were used, in the case of sparse taxon-sampling (Soltis et al., 2004), base compositional bias (Collins et al., 2005; Phillips et al., 2004) or incompleted lineage sorting (Kubatko and Degnan, 2007). Models accommodating these complexities in real molecular evolution should be developed to avoid the inconsistency resulted from analyzing multilocus data. One of these complexities is the heterogeneity in evolutionary rates and patterns among sites (Buckley et al., 2001; Bull et al., 1993). A common way to explicitly model the heterogeneous rates and patterns among sites is to partition the data — using different model for each data partition. Data partitioning should be the obvious choice when analyzing multilocus data, because each locus may have different evolutionary properties (Nylander et al., 2004; Reed and Sperling, 1999). Simulation and empirical studies have shown that analyzing each partition with its own model can significantly improve the likelihood, often increase the nodal supports and may also result in different topologies (Brandley et al., 2005; Castoe et al., 2004; Caterino et al., 2001; Pupko et al., 2002).

The common partitioning strategy is to divide the concatenated sequences by genes, codon positions or both, because this probably captures the most heterogeneity in the sequences. Many studies indeed found out that partitioning by both genes and codon positions resulted in the best fit of the data (Brandley et al., 2005; Caterino et al., 2001). However, over partitioning — dividing the data into too many partitions could result in high sampling errors, because too many parameters associated with excess data partitions need to be estimated from the data. Instead, combining predefined partitions (e.g. by codon positions or genes) that have similar patterns may improve the overall efficiency of the model. For example, first codon positions of two similar genes might be better fitted with one model than two separate models. To choose the best partitioning strategy, ideally, all possible combinations of predefined data partitions should be compared, but the number of combinations becomes astronomically large and mpractical to evaluate when many genes are used. “Background information” or model parameters of each partition have been used to guide the combination of data partitions (Brandley et al., 2005; Poux et al., 2005). For example, the first codon positions were grouped with second condon positions but not the third (Brandley et al., 2005), or partitions with no model parameters differed by more than 100% were grouped together (Poux et al., 2005). These strategies were good attempts for grouping similar data partitions, but they failed to provide a systematic and objective way to explore potential combinations. A better way to group similar data into categories is cluster analysis (Hartigan, 1975). In this study, we proposed using cluster analysis to group the predefined partitioins (by genes and codon positions) into fewer number of data partitions. The model parameters estimated from each predefined partitions were used as the raw data for cluster analysis. We tested whether the reduced number of partitions fit the data better or not by comparing the AIC values and Bayes factors. Partitioned analysis were implemented in both maximum likelihood (ML) method (Jobb, 2006) and in Bayesian approach (Ronquist and Huelsenbeck, 2003).

Ray-finned fish (Actinopterygii) comprises near 27,000 described species, recognized as three subclasses, 44 orders and 453 families (Nelson, 2006). It is the most speciose vertebrate group with high diversity in morphology, ecology, behavior and physiology (see Helfman et al., 1997). Ray-finned fish dates as far back as the Late Silurian (Burrow and Turner, 2000). Understanding the phylogeny of ray-finned fish would help us in studies, such as comparative anatomy, adaptation, taxonomy, vertebrate evolution, biogeography and etc. Because ray-finned fish has the largest diversity in vertebrates, thus high comparative values, knowing the phylogenetic relationships of ray-finned fishes also helps in study of vertebrate genome evolution (Crollius and Weissenbach, 2005). The phylogenetic relationships of ray-finned fish have been the interest of ichthyologists and systematists for many years, yet many parts of the phylogeny are still controversial and unresolved (e.g. Cloutier and Arratia, 2004; Greenwood et al., 1973; Kocher and Stepien, 1997; Lauder and Liem, 1983; Meyer and Zardoya, 2003; Miya et al., 2003; Springer and Johnson, 2004; Stiassny et al., 1996b).

Because the wide range of taxa involved and the lack of synapmorphies, it is difficult to resolve higher-level phylogenies of ray-finned fish by morphological characters alone. To better address the phylogenetic relationships using morphological characters, we still have a lot to learn about the homologies of various characters (Cloutier and Arratia, 2004). Alternatively, molecular data have been used to uncover the phylogenies of ray-finned fish. (Chen et al., 2003; Kocher and Stepien, 1997; Lopez et al., 2004; Miya et al., 2005; Miya et al., 2003; Wiley et al., 2000). Many of the early molecular studies used short sequences and a few loci. Because of the stochastic nature of molecular evolution and insufficient data in short sequences, nodes supported by strong signal can be recovered, whereas some difficult nodes, such as the deep and short internal branches, are hard to be resolved (Weisrock et al., 2005). Collecting data from long sequences or concatenating sequences from many loci would increase the signal to noise ratio and improve the resolution of phylogenetic inference.

One strategy to collect more data is to sequence whole mitochondrial genome, which has the advantage of easy amplification and no difficulty in identifying homologs in contrast to using nuclear genes (Curole and Kocher, 1999; Miya and Nishida, 2000). Impressive works have been done on ray-finned fish phylogenies using mitochondrial genomic data (Inoue et al., 2003; Ishiguro et al., 2003; Miya et al., 2001; Miya et al., 2005; Miya et al., 2003; Saitoh et al., 2003). Novel phylogenetic hypotheses have been proposed, and the resolutions of many parts of the ray-finned fish phylogeny have been improved by these studies. However, one major problem with mitochondrial genomic data is that all genes are usually linked in mitochondrial of vertebrates, thus the whole mitochondrial genome is essentially a single locus. While the large number of characters in mitochondrial genomes can reduce the sampling errors, the linkage of all mitochondrial genes will increase the risk of systematic errors. In fact, independent evidences from nuclear genes have been called to investigate the discrepancies between the results based on mitochondrial loci and morphological data (Curole and Kocher, 1999; Hurley et al., 2007; Meyer and Zardoya, 2003). Here we collected DNA sequences for ten newly developed nuclear gene markers (see Chapter two) in 52 ray-finned fish taxa and four outgroups to assess the hypotheses of ray-finned fish phylogenies.

3.3. Materials and methods

3.3.1. Taxon Sampling, Amplification and Sequencing

We sampled 52 ray-finned fish taxa representing 41 of 44 ray-finned fish orders, except for Saccopharyngiformes, Ateleopodiformes and Stephanoberyciformes due to the short of tissue samples (see Appendix B). Four tetrapods *Xenopus tropicalis*, *Monodelphis deomestica*, *Mus musculus* and *Homo sapiens* were used as outgroups to root the ray-finned fish phylogeny. Certainly the taxon sampling in the present paper is not enough to represent the most diversity of ray-finned fish, even the 41 order, because the delineation of the orders is still an open question (Nelson, 1976, 1984, 1994, 2006). Nevertheless, this is the first attempt to address the phylogenetic relationships among ray-finned fishes using sequences of multiple nuclear genes in a large taxonomic scale.

The nuclear gene makers used were 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 super conserved receptor 2 (serb2) gene (see Chapter two). Sequences of these ten loci for the four tetrapods and the two tetraodontiforms were retrieved from the ENSEMBL genome browser (http://www.ensembl.org, see Appendix B). Sequences for the rest of taxa were determined in this study. The primers used for PCR and sequencing and the reaction conditions followed Chapter 2.

3.3.2. Alignment and Homology Assessment

Because the ten loci used are exons of protein-coding genes, the alignments were done on translated protein sequences using ClustalW (Thompson et al., 1994) implemented in MEGA3.1 (Kumar et al., 2004). Then the aligned protein sequences were translated back into nucleotides for phylogenetic analysis. The ten nuclear genes used are "practical single-copy" gene, which have no duplicates that are more than 50% similar to themselves. Nonetheless, to test whether or not the sequences collected for each locus have paralogs resulted from the fish specific genome duplication events (Taylor et al., 2003; Van de Peer et al., 2003), the most similar fragments, putative "paralogs" in the genome other than the locus itself were download from ENSEMBL for zebrafish, stickleback, medaka, torafugu and spotted green pufferfish. The putative "paralogs" were aligned with all sequences collected in the present study and Neighbor-joining (NJ) trees were constructed for each locus (Saitou and Nei, 1987). If all sequences collected are homologous to each other, the "paralogs" are expected to be positioned at the base of the common ancestor of ray-finned fishes.

3.3.3. Parameters Estimation, Cluster Analysis and Data Partitioning

At first, data matrix for ten nuclear genes was partitioned as the common ways — by genes, by codon positions or by both genes and codons. The most thorough partitioning scheme was by both genes and codons, resulting in 30 blocks of data. Reduced number of partitions may exist that can better explain the data because some of the 30 partitions could have similar evolutionary properties. To reduce the number of partitions from the full 30, I used cluster analysis to group partitions based on parameters estimated from each partitions using GTR + Gamma model. The parameters, including five substitution rates, three base compositional proportions, one gamma parameter and one relative rate for each data partition were estimated using both ML method implemented in TreeFinder (Jobb, 2006) and Bayesian method implemented in MrBayes (Nylander et al., 2004). The ten parameters estimated were then used in a hierarchical cluster analysis with centroid distance to join the partitions into reduced number of groups. The cluster analysis was carried as PROC CLUSTER in SAS program. The tree resulted from the cluster analysis was used to guide the grouping process that reducing the number of partitions. All different partitioning schemes, from one to 30 partitions were compared for their effects in phylogenetic analysis using AIC values and Bayes factors. The effects of different partitioning on resulted topology were also examined.

3.3.4. Phylogenetic Analysis

The basic summary information for each loci, such as the number of parsimony informative site, average genetic distance and consistence index were calculated using PAUP (Swofford, 2003). All data partitioning schemes were tested use both ML and Bayesian methods. The best partitioning scheme was chosen by AIC values or Bayes factors. Bayesian analyses implemented in MrBayes v3.1.1 and ML analyses implemented in TreeFinder (Jobb, 2006) were performed on the nucleotide sequences. GTR + G model was used for all data partitions, and the model parameters were estimated for each partition. Three million generations with 4 chains were run for Bayesian analysis. The tree sampling frequency used was one in a hundred. The last 1/6 trees sampled were used to compute the consensus tree and posterior probabilities. Two independent runs were used to provide additional confirmation of convergence of posterior probability distribution. Two hundreds bootstraps was carried for ML analysis for the best partitioning scheme. Alternative hypotheses were tested by one-tailed Shimodaira and Hasegawa (SH) test (Shimodaira and Hasegawa, 1999) with 1000 RELL bootstrap replicates implemented in TreeFinder.

3.4. RESULTS

3.4.1. Characteristics of the Ten Nuclear Loci Amplified in Ray-finned Fishes

The aligned sequences concatenating all ten loci produced 7995 nucleotides. Sequences were collected for most taxa and loci with about 16% missing data (see Appendix B). The summary information for each locus is listed in Table 3.1. NJ analyses on putative “paralogs” and sequences collected showed that the “paralogs” sequences are all positioned at the root of ray-finned fish tree or join the root as polytomies, suggesting the sequences collected are homologous fragment (results not shown).

3.4.2. Comparison among Partitioning by Genes and Codons and Its Reduced Forms

To analyze the concatenated sequences, data were traditionally partitioned by genes, codon positions or by both genes and codons. Partitioning by both genes and codons resulted in 30 blocks of data in the present study. Hierarchical cluster analysis was carried to join the 30 blocks into smaller number of groups. Cluster analyses were performed on the model parameters (results not shown) estimated using both ML and Bayesian approaches. Clusterings based on parameters estimated from ML or Bayesian method have similar patterns except for minor differences exist within the major clades (Fig. 3.1). The most significant clustering indicated by the PST2 values (data not show) for both ML and Bayesian approach are two clusters and three clusters. The two clusters include a clade of first and second codon positions and a clade of third codon positions of all ten genes, while the three clusters include three clades grouped by codon positions (Fig. 3.1).

All different partitioning schemes, from 1 partition (no partitions) to 30 paritioins guided by the tree resulted from cluster analysis as well as the traditional partitioned by genes strategy were compared for their effects on phylogenetic analysis. The performances of different partitioning schemes were evaluated under both ML and Bayesian context (Table 3.2). The AIC value decreases dramatically when the data were partitioned by (1^{st} + 2^{nd}) and 3^{rd} codon position, while the AIC value decreases slowly in subsequent further dividing the data. Nonetheless, partitioning by both genes and codons has the lowest the AIC value (Table 3.2, Fig. 3.2a). Because there were very little improvements after more than 21 partitions were used indicated by the value of AIC_i-$AIC_{(I-1)}$, I chose 21 partitions as our best scheme for phylogenetic analysis (Table

3.2).The Bayesian analysis for different partitioning schemes resulted in the similar patterns (Table 3.2, Fig. 3.2b). However, partitioning the data by 17 groups yielded the best likelihood instead of using the full 30 partitions by both genes and codons (Table 3.2, Fig. 3.2b). When more partitions are used, less data are available to estimate the increased number of parameters, which can lead to higher sampling errors and the slower convergence of MCMC runs in MrBayes. I found that higher number of partitions resulted in slower convergence of two MrBayes runs suggested by the average standard deviation of split frequencies (Table 3.2). Considering both the likelihoods and the standard deviation of split frequencies, I chose 16 partitions instead of 17 partitions for the best partitioning scheme (Table 3.2). In both of the ML and Bayesian context, partitioning by 10 genes produced much worse likelihood than the 10 partitions selected by cluster analysis (Table 3.2, Fig. 3.2). Data partitioning not only changed the likelihood, but also changed topology of the resulted phylogeny (Fig. 3.2,).

3.4.3. Interrelationships among Ray-finned Fishes

Considering both AIC values and Bayes factors, the reduced number of partitions preduced better results than the tranditional partitioning by both genes and codon positions. ML analysis and Bayesian analysis based on their best partitioning schemes yielded almost the same topology (Fig. 3.3). The only difference between the results from ML methods and Bayesian approach is the branching order among Aulopiformes, Percopsiformes, and Gadiformes, which is depicted as a polytomy in Fig. 3.3.

3.5. DISCUSSION

3.5.1. Effects of Different Partitioning Schemes

When data from multiple loci are used in phylogenetic analysis, partitioned analysis is one of the best ways to accommodate the heterogeneous molecular evolution among different parts of the concatenated sequences. The most common ways of partitioning multiple loci data are by secondary structures, by genes or by codons (Brandley et al., 2005; Castoe et al., 2004). Using more partitions should increase the likelihood of the data, but it also loses statistic power because more parameters need to be estimated for more partitions. Therefore, combining partitions into smaller groups should be considered and evaluated by their AIC values or Bayes factors to optimize the best strategy of partitioning. However, no systematic and objective ways of combining partitions have been proposed other than using “background information” (Brandley et al., 2005) or similarity between model parameters (Poux et al., 2005). In this paper, parameters estimated from the smallest block of partitions (by genes and codons) were used in cluster analyses to determine the way of grouping data. My results show that partitioning by codons resulted in the biggest improvement in AIC values and Bayes factor, indicating the most heterogeneity is between different codon sites, especially between the first and second codon and the third codon. The cluster analysis has been shown as an effective way to group the small partitions. Although, the improvement of partitioning became smaller when a larger number of partitions used, the largest number of partitions is still the best strategy according to AIC values (Table 3.2 and Fig. 3.2). However, the Bayes factors suggest that reduced number of partitions is better than the full 30 partitions by genes and codons. Nylander et al. (2004) also found Bayes factor preferred simple partitioning model than complex ones in comparison of non-nested models. Because Bayes factors choose the reduced number of partitions other than the full 30 partitions and the AIC values indicates small gains after more than 21 partitions, we think reduced number of partitions obtained from cluster analysis is more efficient than fully partitioned by both genes and codon positions.

Data partitioning not only improves the likelihood of the data, but also increases sampling error due to too many parameters introduced. Therefore, when selecting the partitioning scheme, we prefer a conservative rule — picking the model with less number of partitions if there is no significant improvement for the more complex model. If a partition has only a few characters, there would be just not enough data to estimate the model parameters, which could lead to no convergency of MCMC process. The slower convergency rate when data were analyzed with higher number of partitioins were observed in my Bayesian analysis indicated by the standard deviation of split frequencies (Table 3.2). The high standard deviation of split frequencies can be used as a good indicator of excessive number of data partitions.

In the contrary of the large change in the likelihood, the topology usually remain similar among different partitioned analysis (Buckley et al., 2001). However, I observed many changes in topology when the data were analyzed with different number of partitions for both ML and Bayesian methods (Fig. 3.2). First topology changes happened when the partitioning used switched from no partition to two partitions and to three partitions (Fig. 3.2). Then the topology remained the same as the number of partitions increased. When the number of partitions kept rising, more topological changes were resulted (Fig. 3.2). This pattern of topological changes may suggust that when a few reasonable partitions were introduced into the model, it would reveal the true topology by fitting the data better. When too many partitions were used, it many change the topology again just because the high random errors being introduced into the model along with more parameters. These later topological changes were more conspicuous in Bayesian analysis than in ML methods (Fig. 3.2), which is consistant with that Bayesian approach account for model uncertainty more than ML methods does. The failure of covergency of MCMC runs indicated by the standard deviation of split frequencies also predicted the unstable topology inferred using Bayesian method when too many partitions were used.

3.5.2 Lower Actinopterygians

The extant actinopterygians belong to five major clades, polypteriforms, acipenseriforms, lepisosteiforms, amiiforms and teleosts. Lower actinopterygians are the basal ray-finned fishes, including two extant lineages, polypteriforms and acipenseriforms and about 270 fossil genera (Gardiner, 1993; Grande and Bemis, 1996). Lower actinopterygians were sometimes referred to as “Chondrostei” (Nelson, 1994; Schaeffer, 1973), but recent evidences from both morphological (Gardiner et al., 2005; Grande and Bemis, 1996) and molecular (Inoue et al., 2003; Kikugawa et al., 2004; Venkatesh et al., 2001) data all pointed out that “Chondrostei” is actually a paraphyletic group. The most consensus view place polypteriforms as the basal group to all other actinoterygians while putting acipenseriforms as the sister group to neopterygians (*Lepisosteus*, *Amia* and teleosts) (Nelson, 2006). Interestingly, my results support the old "Chondrostei" hypothesis, grouping the polypteriforms together with acipenseriforms as a monophyletic group with a bootstrap value of 64% and a posterior probability of 0.86. However, the SH-test cannot reject polypteriforms as the basal clade to all other ray-finned fishes (p=0.823, Table 3.3).

3.5.3. Basal Neopterygians

Most morphological (Patterson, 1973; Regan, 1923) and molecular (Crow and Wagner, 2006; Hurley et al., 2007; Kikugawa et al., 2004; Lê et al., 1993) evidences support 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 (Jessen, 1972; Nelson, 1969). More recent morphological hypotheses suggest that either Amiiformes (Grande and Bemis, 1996; Patterson, 1973) or Lepisosteiformes (Olsen, 1984) is the sister-group to teleosts. However, mitogenome data and a insertion in nuclear RAG2 gene support a very different view, that is the Acipenseriformes, *Lepisosteidae* and *Amia* form a monophyletic "ancient fish" group, and together join to teleost as a sister-group (Inoue et al., 2003; Venkatesh et al., 2001). My data support the "Holostei" hypothesis with a 100% bootstrap value and a 1.0 posterior probability. SH-tests using our data could not reject the "ancient fish" (p=0.225) hypothesis, but rejected either the *Amia* and teleosts sister-group (p=0.028) or the *Lepisosteidae* and teleost sister-group hypotheses (p=0.023) (Table 3.3). The "Holostei" hypothesis was also recovered in a study using multiple nuclear genes (Kikugawa et al., 2004) and in a re-analysis of morphological characters using both extant and fossil species (Hurley et al., 2007). The discrepancies between my results and the "ancient fish" theory could be explained by the artifacts in data analysis of mitogenome data (Kikugawa et al., 2004) or parallel insertion events in the RAG2 gene. However, to settle this controversy, I should collect more molecular and morphological data and understand better about the evolution of molecular and morphological characters. Rare genomic changes (RGCs), such as the insertions in coding region of RAG2 are good phylogenetic characters (Rokas and Holland, 2000), but they are not immune from homoplasy. Only one insertion in RAG2 gene support the “ancient fish” hypothesis (Venkatesh et al., 2001), therefore more RGCs characters should be pursued to test the competing hypotheses.

3.5.4. Basal Teleosts

The monophyly of Teleostei is supported by many morphological characters (Arratia, 2000; de Pinna, 1996). There are four major teleostean lineages, Elopmorpha, Osteoglossomorpha, Ostarioclupeomorpha and Euteleostei (Nelson, 2006). After strong evidences grouped the Clupeomorpha and Ostariophysi into Ostarioclupeomorpha (Arratia, 1997; Lê et al., 1993; Lecointre and Nelson, 1996), ostarioclupeomorphs are generally placed as the sister-group to euteleosts (Arratia, 1997; Inoue et al., 2001; Lê et al., 1993). However, the interrelationships among elopmorphs, osteoglossomorphs and more advanced teleosts are still controversial. Both morphological (Patterson and Rosen, 1977) and molecular (Inoue et al., 2001) studies support that osteoglossomorphs are more primitive than elopmorphs, but this view was challenged by a hypothesis suggesting that elopmorphs is the living sister-group of all other extant teleosts (Arratia, 1991, 1997, 2000; Shen, 1996). Base on weak support from 28S gene, Lê et al. (1993) proposed another different hypothesis that osteoglossomorphs and elopmorphs are more close to each other than to the rest teleosts. Our data support elopforms as the basal teleost, although with very low node support (Fig. 3.3). This result is the first evidence from molecular data that confirmed the view of Arratia (1997) that elopmorphs are the most primitive living teleost.

As I mentioned above, sister-group relationship of clupeomorphs and ostariophysans are well established (Arratia, 1997; Lê et al., 1993; Lecointre and Nelson, 1996). My results are consistant with the Ostarioclupeomorpha hypothesis. Ostariophysi has five major lineages, gonorynchiforms, cypriniforms, characiforms, siluriforms and gymonotiforms (Fink and Fink, 1981; Nelson, 2006). Because the lack of otophysic connection, gonorynchiforms were named as Anotophysi and constantly placed as the basal group to all the rest ostariphysans (Fink and Fink, 1981; Nelson, 1994, 2006; Rosen and Greenwood, 1970). A recent study using mitogenomic data challenged this view and proposed that gonorynchiforms are more closely related to clupeomorphs (Saitoh et al., 2003). However, my results support the classic view that gonorynchiforms are the basal ostariophysans. Within Otophysi (ostariophysans minus gonorynchiforms), different phylogenetic hypotheses exist. Recent morphological studies highly support a phylogeny of (Cypriniformes, (Characiformes, (Gymnotiformes, Siluriformes))) (Dimmick and Larson, 1996; Fink and Fink, 1981, 1996), while molecular data support a phylogeny of (Cypriniformes, (Siluriformes, (Gymnotiformes, Characiformes))) (Dimmick and Larson, 1996; Ortí and Meyer, 1996; Saitoh et al., 2003). My results give a different phylogeny, (Cypriniformes, (Gymnotiformes, (Characiformes, Siluriformes))) (Fig. 3.3), and this hypothesis also is supported by RAG1 gene sequences (Ortí et al., unpublished data). All of the three hypotheses agree in placing the Cypriniformes as the basal clade, but contradict each other in the relationships among the other three lineages. In our study, only one taxon was used to represent each of the three families, thus more taxa should be sampled in the future to test the three alternative hypotheses.

3.5.5. Protacanthopterygians

The classification of protacanthopterygians has been changed drastically since Greenword et al. (1966) use it to define a group of primitive teleosts of their division III (Arratia, 1997; Fink, 1984b; Lopez et al., 2004; Williams, 1987). The compositions of protacanthopterygians are still varying in many different hypotheses. Four major clades, argentiniforms, osmeriforms, salmoniforms and esociformes usually are included in Protacanthopterygii (Nelson, 2006), but esociformes sometimes are regarded as the sister-group to neoteleosts (Johnson and Patterson, 1996). However, many recent studies, especially in molecular data, support that esociformes and salmoniforms are sister taxa (Arratia, 1997; Ishiguro et al., 2003; Lopez et al., 2004; Williams, 1987). Besides confirming the sister relationship between esociformes and salmoniforms, Lopez et al. (2004) also suggested a novel sister-taxa relationship between osmeriforms and stomiforms (Neoteleostei). My data corroborate both findings of Lopez et al., the sister-taxa between esociformes and salmoniforms and the close relationship between osmeriforms and stomiforms (Fig. 3.3). More data and complete taxa sampling should be used in the future to test the sister-taxa relationship between osmeriforms and stomiforms, since it suggests potential needs of redefining the Protacanthopterygii and Neoteleostei.

3.5.6. Neoteleostei

Neoteleostei is a monophyletic group supported by a few morphological characters (Johnson, 1992; Nelson, 1994). Monophyly of Neoteleostei is also supported by my data with a 92% bootstrap value and a 1.0 Bayesian posterior probability, if osmeriforms is also included in Neoteleostei as the sister group to stomiforms. Neoteleostei has eight major lineages, Stenopterygii, Ateleopodomorpha, Cyclosquamata, Scopelomorpha, Lampriomorpha, Polymiciomorpha, Paracanthopterygii and Acanthopterygii in the sequence of branching order (Nelson, 2006), although the composition of some lineage is continually changing, e.g. Paracanthopterygii (Greenwood et al., 1966; Miya et al., 2005; Miya et al., 2003; Patterson and Rosen, 1989). There are no representing taxa sampled for Ateleopodomorpha in the present study, and taxa sampled for the rest of lineages are also sparse. So, I have no ambition to resolve the interrelationships among these groups, but instead to show some of the classic patterns supported by our data and some novel relationships which worth more investigation. Stomiiforms (Stenopterygii) together with osmeriforms were found as the basal group to the rest of neoteleosts in our results. The next clade suggested by my results is a polytomy of percopsiforms, gadiforms, aulopiforms and the rest of teleosts (Fig. 3.3). The next group supported is myctophiforms (Scopelomorpha), a clade grouping *Polymixia* (Polymixiomorpha) with *Zeus* (Zeiformes), lampriforms (Lampriomorpha) and Acanthopterygii (Fig. 3.3). The major different between my results and the classic view (Nelson, 2006) or the mitogenomic hypothesis (Miya et al., 2005; Miya et al., 2003) is the treatment of “Paracanthopterygii”. In agree with the results of mitogenomic studies (Miya et al., 2005; Miya et al., 2003), our results suggested that the former “Paracanthopterygii” members, ophidiiforms and batrachoidiforms are actually basal acanthopterygians, while the lophiiformes are close to the more derived acanthopterygian, Tetreodontiformes (Fig. 3.3). Different from the mitogenomic results (Miya et al., 2005; Miya et al., 2003), the other putative paracanthopterygians, Polymixiidae and Zeioidei were not found in the same clade with the rest paracanthopterygians in my results (Fig. 3.3).

3.5.7. Acanthopterygii

If ophidiiforms, batrachoidiforms and lophiiforms are included, Acanthopterygii also is supported as monophyletic by my data with a 100% bootstrap value. Beryciforms, ophidiiforms and batrachoidifroms were found as the basal acanthopterygians in the sequence of branching order. The rest acanthopterygians were grouped as a monophyletic clade with a 100% bootstrap value, and this clade corresponds to Percomorpha by Johnson and Patterson (Johnson and Patterson, 1993). Within Percomorpha, two major clades were supported. One clade includes highly supported sister-taxa of atherinomorphs, mugiliomorphs and perciforms (Cichlids). The other highly supported clade includes tetraodontiforms, lophiiforms, perciforms, gasterosteiforms and scorpaeniforms, grouped with pleuronectiforms with low support. My results within Percomorpha corroborate the finding of mitogenomic studies (Miya et al., 2005; Miya et al., 2003), but not fully agree with the “Smegmamorpha” hypothesis suggested by Johnson and Patterson (1993), which group Gasterosteiformes with Atherinomorpha, Mugiloidei, Elassomatidae and Synbranchiformes. One of the significant indications from the interrelationships of percomorphs is that members belong to Perciformes are paraphyletic, and this result also was showed up in mitogenomic studies (Miya et al., 2005; Miya et al., 2003). Because acanthopterygians have 267 families, 2,422 genera (Nelson, 2006), more taxa should be sampled before major revisions can be made in this group.

Table 3.1. Characteristics of the ten nuclear loci amplified in ray-finned fishes.

Genes	No. of bp	No. of var. sites	No. of PI sites	Average p-distance	CI-MP	No. of species sequenced
zic1	927	395	345	0.158	0.232	54
myh6	735	369	325	0.174	0.232	48
RYR3	834	497	425	0.215	0.280	41
Ptr	705	426	375	0.206	0.272	51
tbr1	720	410	328	0.196	0.367	42
ENC1	810	405	359	0.180	0.242	50
Gylt	888	589	509	0.215	0.291	44
SH3PX3	705	373	319	0.168	0.270	45
plagl2	684	410	344	0.179	0.316	44
sreb2	987	431	387	0.149	0.254	51

*bp, base pairs; var., variable sites; PI, parsimony informative sites; CI-MP, consistency index on the maximum parsimonious tree.

Table 3.2. Comparison of log likelihood, AIC and Bayes factors among different partitioning schemes.

	Maximum likelihood				Bayesian analysis		
Number of partitions	Likelihood (-L)[a]	Number of parameters	AIC	$\text{AIC}_i - \text{AIC}_{(I-1)}$	Likelihood (-L)[b]	Bayes factor[c]	Split deviation[d]
1 partition	130936	9	261890		131050		0.005943
2 partitions	127075	19	254188	7702	127095	3955	0.004624
3 partitions	126686	29	253431	758	126720	375	0.007499
4 partitions	126654	39	253387	44	126694	26	0.005629
5 partitions	126484	49	253066	321	126542	152	0.006435
6 partitions	126421	59	252961	105	126474	68	0.008284
7 partitions	126373	69	252885	76	126364	110	0.008371
8 partitions	126324	79	252806	79	126327	37	0.008377
9 partitions	126237	89	252652	154	126282	45	0.009426
10 partitions	126190	99	252579	73	126261	20	0.010901
11 partitions	126160	109	252538	41	126178	84	0.008561
12 partitions	126119	119	252475	63	126136	41	0.014122
13 partitions	126068	129	252393	82	126126	10	0.008394
14 partitions	126038	139	252353	40	126114	12	0.015416
15 partitions	125988	149	252275	79	126086	28	0.016578
16 partitions	125966	159	252249	25	125947	138	0.015155
17 partitions	125913	169	252165	85	125857	91	0.031614
18 partitions	125861	179	252079	85	125907	-51	0.020992
19 partitions	125829	189	252036	44	125881	26	0.028444
20 partitions	125816	199	252030	5	125865	17	0.039061
21 partitions	125718	209	251855	176	125921	-57	0.025118
22 partitions	125703	219	251844	11	125840	81	0.035717
23 partitions	125691	229	251841	3	125893	-52	0.023924
24 partitions	125678	239	251834	7	125885	8	0.048132
25 partitions	125650	249	251798	36	125935	-50	0.034249
26 partitions	125630	259	251777	20	125903	32	0.035437
27 partitions	125607	269	251752	25	125897	5	0.096736
28 partitions	125600	279	251759	-6	125897	0	0.064801

Table 3.2. (cont.).

	Maximum	likelihood			Bayesian analysis		
Number of partitions	Likelihood (-L)[a]	Number of parameters	AIC	$AIC_i - AIC_{(I-1)}$	Likelihood (-L)[b]	Bayes factor[c]	Split deviation[d]
29 partitions	125569	289	251716	42	126032	-135	0.051778
30 partitions	125551	299	251699	17	125937	96	0.132187
By genes	130509	99	261216	-	130570	-	0.021610

[a] -Log likelihood calculated using TreeFinder (Jobb, 2006).

[b] Harmonic mean of -log likelihood calculated using MrBayes (Nylander et al., 2004).

[c] Bayes factor calculated as comparing model i to i-1.

[d] Average standard deviation of split frequencies of two MCMC runs.

Table 3.3. SH-test on hypothesis of interrelationships among basal actinopterygians

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

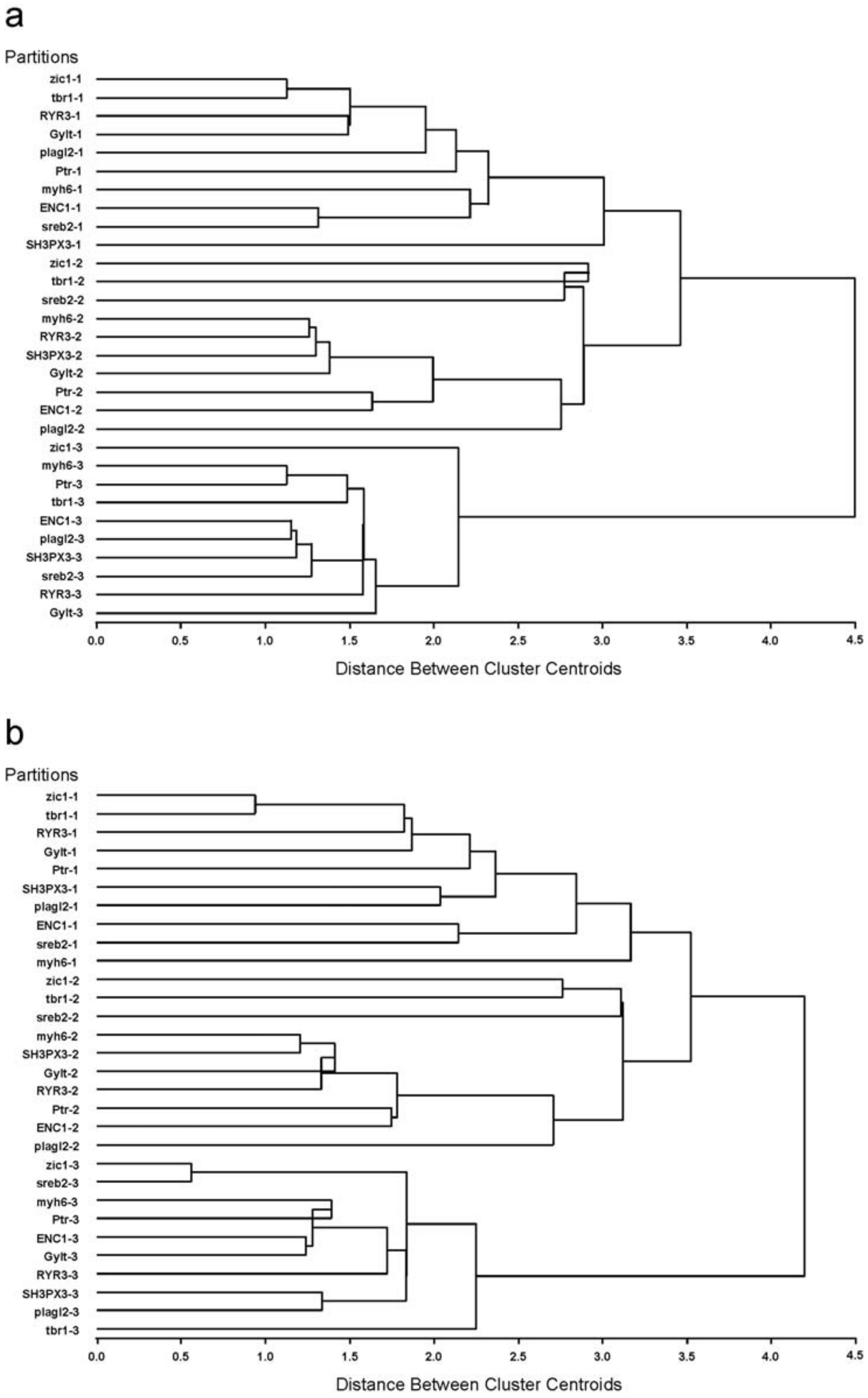

Fig. 3.1. Cluster analysis on 30 blocks of data partitioned by genes and codons. Ten model parameters estimated from each block were used as the raw data for cluster analysis. a. cluster analysis on parameters estimated using ML method in TreeFinder. b. cluster analysis on parameters estimated using Bayesian approach in MrBayes.

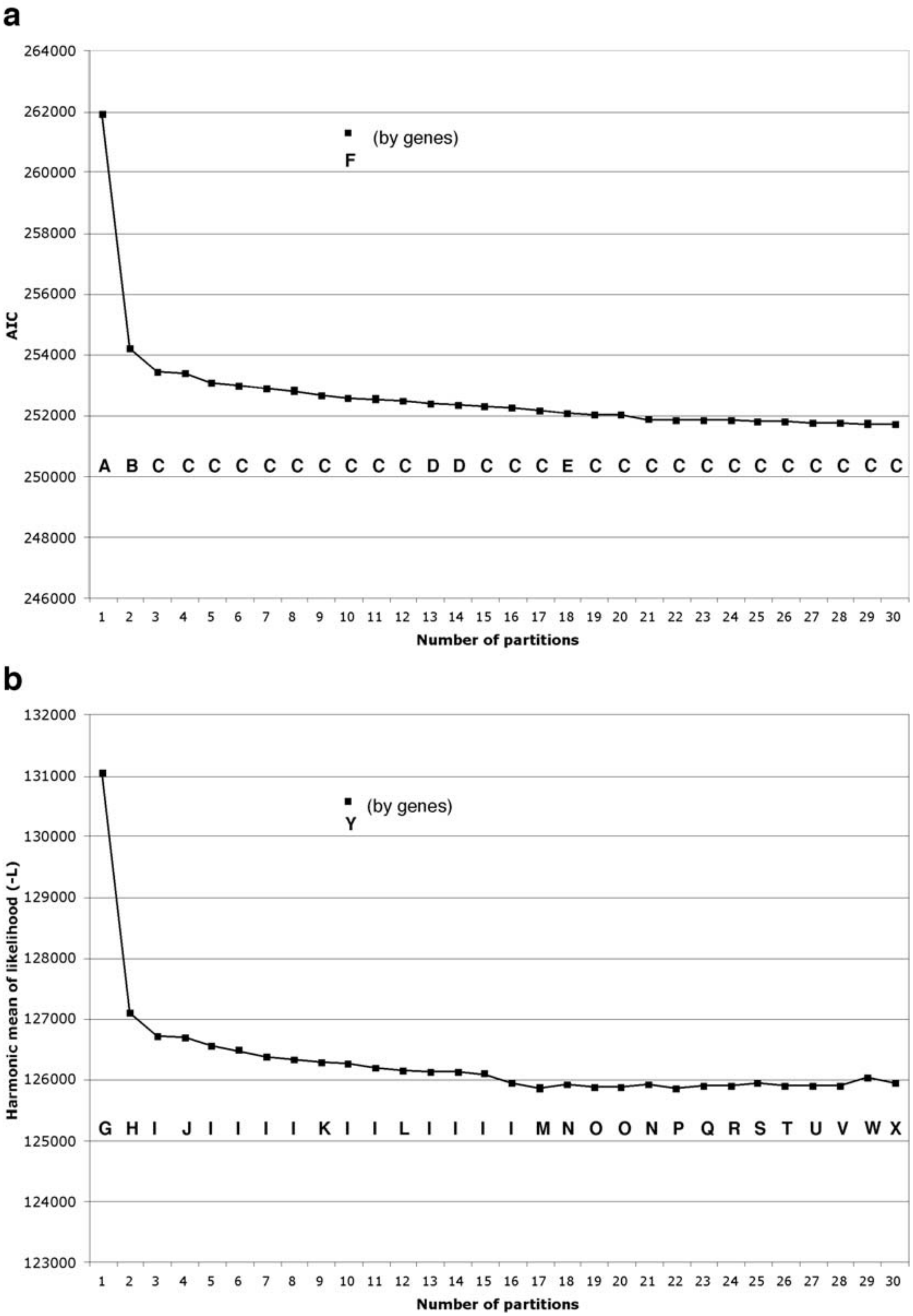

Fig. 3.2. (a) AIC values and (b) Bayesian posterial likelihood for analyses under different data partitioning schemes. A to Y indicate the different topologies resulted from different partitioning schemes, the topology of A to Y can be found in supplemental Fig. 1.

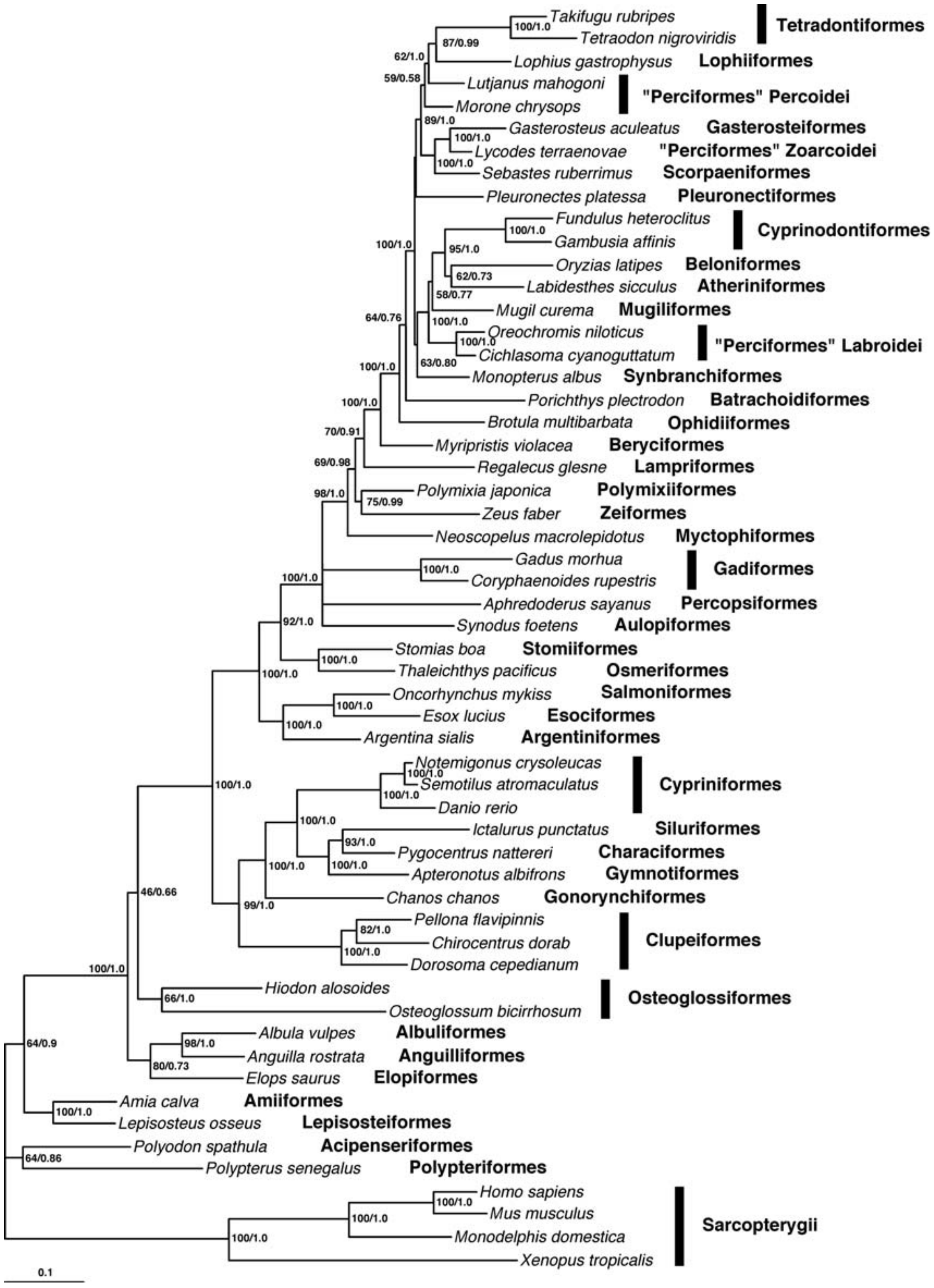

Fig. 3.3. Phylogeny of ray-finned fish based on partitioned analyses of ten nuclear genes. Data were partitioned into 21 groups in ML analysis and 16 groups in Bayesian analysis. The numbers on branches are ML bootstrap values and Bayes posterior probabilities. The only difference between ML and Bayesian methods is the branching order among Aulopiformes, Percopsiformes and Gadiformes, which is depicted as polytomy here.

Chapter 4 – Molecular phylogeny of Clupeiformes inferred from nuclear RAG genes and mitochondrial ribosomal rRNA

4.1. Abstract

The taxonomy of clupeiforms has been extensively studied, yet phylogenetic relationships among component taxa remain controversial or unresolved. Here I test current and new hypotheses of relationships among clupeiforms using mitochondrial rRNA genes (12S and 16S) and nuclear RAG1 and RAG2 sequences (total of 4749 bp) for 37 clupeiform taxa representing all 5 extant families and all subfamilies of Clupeiformes, except Pristigasterinae, plus 7 outgroups. My results, based on maximum parsimony, maximum likelihood, and Bayesian analyses of these data, show that some traditional hypotheses are supported. These include the monophyly of the families Engraulidae, consisting of two monophyletic subfamilies, Engraulinae (*Engraulis* and *Anchoa*) and Coilinae (*Coilia* and *Setipinna*), and Pristigasteridae (here represented only by *Ilisha* and *Pellona*). The basal position of *Denticeps* among clupeiforms is consistent with the molecular data when base compositional biases are accounted for. However, the monophyly of Clupeidae was not supported. Some clupeids were more closely related to taxa assigned to Pristigasteridae and Chirocentridae (*Chirocentrus*). These results suggest that a major revision in the classification of clupeiform fishes may be necessary, but should await a more complete taxonomic sampling and additional data.

4.2. Background

Clupeiform fishes include well-known species such a herrings, sardines, and anchovies, and most of them are important components of global commercial fisheries. They have a worldwide distribution, but the majority of the 402 described species occur in the Indo-West Pacific Ocean, inhabiting marine and brackish waters, and only around

70 species are primarily freshwater fishes (Froese and Pauly, 2004; Whitehead, 1985). Clupeiforms are classified in the Clupeomorpha, a group that includes many Cretaceous fossil taxa (Grande, 1985), and the order Clupeiformes that includes all living species. Several well-defined synapomorphies support the monophyly of Clupeiformes, such as the presence of the *recessus lateralis*, a unique structure associated with the connection of the anterior portion of the swim bladder with the inner ear. Currently, the order is divided into 5 families: Denticipitidae, Pristigasteridae, Engraulidae, Clupeidae and Chirocentridae (Froese and Pauly, 2004; Nelson, 1994). The taxonomic composition of clupeiforms has been extensively studied (Greenwood, 1968; Whitehead, 1972, 1985; Wongratana, 1987), as the group was perceived to be a distinctive subdivision of basal teleosts. The interrelationships of clupeiforms with other basal teleosts have been hotly debated, but their current placement as the sister-group of ostariophysans (gonorhynchiforms, cypriniforms, siluriforms, characiforms, and gymnotiforms), initially proposed on the basis of molecular characters, is now generally accepted (Di Dario, 2002, 2004; Grande, 1985; Lecointre and Nelson, 1996; Nelson, 1967, 1970); relationships among and within families of Clupeiformes, however, remain controversial and unresolved.

The family Denticipitidae, a group known from some West African coastal rivers, is distinguished by unique characters among clupeiforms such as the possession of denticles on roofing bones of the skull and a complete lateral line. *Denticeps*, is the only living representative of this family that has been considered to represent the basal taxon in the order (Greenwood, 1968). Chirocentridae and Clupeidae have been united under the superfamily Clupeoidea on the basis of two synapomorphies, an increase in rib to pleural vertebrae ratio and the fusion between epicentrals and ribs (Grande, 1985; Patterson and Johnson, 1995). However, a recent morphological study suggested that Chirocentridae is in fact more closely related to Engraulidae (Di Dario, 2005). The interrelationships among Pristigasteridae, Engraulidae and Clupeidae also are controversial. Clupeidae and Engraulidae were proposed to be more closely related to each other based on the presence of cartilage chevrons at the tips of epicentrals, a similar orientation of parapophyses of the second vertebra, and the complex pattern of interzygapophysal articulation (Di Dario, 2002; Patterson and Johnson, 1995). However, there is also morphological evidence (e.g., the gongyloid cartilage) supporting a close relationship between Pristigasteridae and Engraulidae (Di Dario, 2002).

Below the family level, the main problem remains in the definition of subfamilies within Clupeidae, the largest assemblage in the order. The subfamilies Dorosomatinae, Alosinae, and Clupeinae were considered as "groups of convenience", since no osteological characters were found to support the monophyletic status of these subfamilies (Grande, 1985). Pellonulinae and Dussumieriinae were proposed as a monophyletic group (Grande, 1985), but Dussumieriinae was also thought to be polyphyletic (Bill Eschmeyer, personal comm.). Nelson (1970) proposed that Clupeidae should be divided into two groups, Clupeinae and Dorosomatinae based on the characters in gill arches.

Another intriguing case involves the taxonomic placement of *Sundasalanx*, which is a miniature, transparent, and highly paedomorphic freshwater fish distributed in Southeast Asia (Borneo, Laos and Thailand). Seven species have been described for this genus (Britz and Kottelat, 1999; Roberts, 1981; Siebert, 1997). Originally, *Sundasalanx* was placed in its own family and considered most closely related to Salangidae (Roberts, 1981), and was indeed placed in the Salangidae later (Roberts, 1984), and Osmeridae subsequently (Fink, 1984a). Based on non-paedomorphic features, such as an ossified prootic bulla and the *recessus lateralis*, Siebert (1997) proposed that *Sundasalanx* are in fact clupeids, and went even further to suggest that *Sundasalanx* is possibly the sister-group of the Caribbean endemic genus *Jenkinsia*. Britz and Kottelat (1999) criticized Siebert's conclusion and called for additional evidence to demonstrate that *Sundasalanx* shares synapomorphies with Clupeidae, advocating the need for a broader phylogenetic study. New evidence came in the form of mitochondrial genomic data when Ishiguro et al. (2005) confirmed that *Sundasalanx* grouped with clupeiforms in their phylogeny with high bootstrap support; given the sparse taxon sampling with the order, however, their study could not resolve the phylogenetic place of *Sundasalanx* within Clupeiformes.

The current hypotheses on the phylogeny of Clupeiformes are summarized in Fig.4.1. All significant studies published are based solely on morphological characters (Di Dario, 2002; Grande, 1985; Nelson, 1970; Patterson and Johnson, 1995). Mitochondrial genomic data are being used in a study on the patterns of diversification of clupeiform fishes (Sébastien Lavoué, personal comm.). In this study, I use both mitochondrial DNA (mtDNA, 12S and 16S ribosomal genes) and nuclear DNA sequence data (recombination activating genes, RAG-1 and 2) to study the phylogenetic relationships among families of Clupeiformes and among subfamilies of Clupeidae, based on a representative taxonomic sampling. Because nuclear genes have low evolutionary rate and less likely to succumb to saturation, they should provide important information complementary to the mitochondrial genes for this study.

4.3. Materials and Methods

4.3.1 Taxon sampling

Thirty-seven clupeiform species and 7 outgroup taxa were sampled for this study (Table 4.1). The outgroups included 5 ostariophysan fishes, and 2 more distantly related taxa, *Oncorhynchus mykiss* and *Salangichthys microdon*. Ostariophysi was chosen as an outgroup for this study because it has been proposed as the sister group to Clupeomorpha (Lê et al., 1993; Lecointre and Nelson, 1996). *O. mykiss* (Salmoniformes) and *S. microdon* (Osmeriformes) were used to assess the affinities of *Sundasalanx*. The 37 clupeiform fishes examined are distributed among 22 genera. They represent all 5 extant families and all subfamilies of Clupeiformes, except Pristigasterinae (Fig. 4.1).

4.3.2. DNA extraction, polymerase chain reaction and sequencing

DNA samples were extracted from ethanol-preserved muscle or gill tissues using the DNeasy tissue kit (Qiagen). Fragments of two mitochondrial ribosomal genes (12S and 16S) were sequenced for this study. To design PCR and sequencing primers for the mtDNA genes, sequences for each gene were retrieved from the mitochondrial genome data of *Engraulis japonicus* (accession numbers: NC_003097), *Chirocentrus dorab* (NC_006913), *Sundasalanx mekongensis* (NC_006919) and *Danio rerio* (NC_002333). The primers were designed based on the alignments of these sequences. The PCR target fragment for 12S (size = 726 bp) spans sites 1249 through 1974 in the *D. rerio* mitochondrial genome. The target fragment for 16S (size = 1408 bp) spans most of this gene, from position 2178 through 3586 in the *D. rerio* mitochondrial genome. Two primers were designed for 12S while three were designed for 16S, since an additional internal primer for sequencing was necessary (Table 4.2). The same thermo-cycler profiles for PCR were used for 12S and 16S gene, with 31 cycles and annealing temperature of 57°C.

In addition to the mtDNA sequences, nuclear DNA (nucDNA) fragments of two single-copy genes were sequenced. Recombination activating genes 1 and 2 (RAG1 and RAG2) are closely linked genes coding for components of recombinase, an enzyme involved in immunoglobin function (Hansen and Kaattari, 1996; Oettinger, 1992; Willett et al., 1997). Fragments of these genes are being used increasingly in phylogenetic studies of fishes and other vertebrates (Lopez et al., 2004; Lovejoy and Collette, 2001). Published PCR primers for RAG1 (Lopez et al., 2004) and RAG2 genes (Lovejoy and Collette, 2001) are available, but new primers also were designed for this study (Table 4.2) based on alignments of published sequences from *O. mykiss* (AF137176 and U31670) and *D. rerio* (U71093 and U71094). Nested-PCR was used to amplify RAG1 and RAG2 genes for taxa that failed to be amplified in a single PCR. In these cases, products of the first-round PCR were diluted 100 times and used as the template for a second PCR with a set of primers inferred to nest within the fragment amplified in the first PCR. Conditions for amplification of the RAG1 fragment for both rounds of PCR used 15 cycles with annealing temperature at 52°C followed by 15 cycles with annealing temperature at 51°C. PCR and nested-PCR conditions for RAG2 used 15 cycles with annealing temperature at 62°C followed by 15 cycles with annealing temperature at 60°C. Primers for PCR, nested-PCR, and sequencing are listed in Table 4.2.

4.3.3. Sequence alignment, insertions/deletions (indels), substitution saturation test, and base compositional stationarity test

Alignments of mitochondrial gene sequences were performed using the program Clustal X (Thompson et al., 1997) with default parameters. The protein-coding regions of RAG1 and RAG2 DNA sequences were aligned based on their putative amino acid sequences (genetic code = universal) using ClustalW (under default conditions) integrated with MEGA 3.1 (Kumar et al., 2004). For some analyses, aligned amino acid sequences were back-translated to their original DNA sequences while keeping the amino acid alignment.

Inferred indels (insertion/deletion events) resulting from the alignment procedure may harbor valuable phylogenetic information in the form of shared characters, and can be used in phylogenetic analysis by coding the alignment gaps as additional characters in the data matrix (Simmons and Ochoterena, 2000). Gapped regions generated in the alignment were coded for phylogenetic analysis using the modified complex coding method (Müller, 2006) implemented in the software SeqState (Müller, 2005). The coded indels were used as additional characters together with the nucleotide sequences in parsimony and Bayesian analysis. Because the alignment was based on the amino acid sequences, the coded gap characters were given a weight = 3 times higher than single nucleotide substitutions for parsimony analyses.

In order to check the degree of saturation for substitutions at each gene, I plotted the pair-wise absolute number of substitutions against maximum likelihood corrected sequence divergence for each data partition. A linear relationship would be expected if there is no saturation. Best-fit models and parameters used to calculate the corrected distance were chosen by Modeltest (Posada and Crandall, 1998) (Table 4.3)

To detect potential systematic errors in phylogenetic inference that may result from heterogeneous base composition among taxa, I estimated the base composition (%G+C) at variable sites for each gene; stationarity of base composition was tested further with the Chi-square test implemented in PAUP* v4.0 (Swofford, 2003). Phillips *et al.* (2004) showed that RY-coding (coding purines as R, and pyrimidines as Y) could effectively reduce misleading signal from biased base composition. I used this method to compare support for alternative topologies (splits) based on the raw nucleotide sequences and the RY-coded data to test whether a critical node in the tree could be an artifact produced by convergent base composition. The amount of phylogenetic signal supporting alternative branching splits was calculated using Spectronet (Huber et al., 2002).

4.3.4. Data partitions and phylogenetic analysis

Nuclear and mtDNA were treated as separate data partitions for phylogenetic analyses. Each genetic fragment (12S, 16S, RAG1, and RAG2) was, however, considered separately for the alignment and to characterize their evolutionary dynamics and properties. Given that both mtDNA fragments have similar function/structure (coding for ribosomal RNA) and are also tightly linked, they were considered as a single partition for phylogenetic analyses. Similarly, RAG1 and RAG2 fragments were also treated as a single partition. Congruence among partitions was assessed by the ILD test (Farris et al., 1995a, b) implemented in PAUP*. All data (mtDNA and nucDNA partitions) were subsequently combined for a total evidence analyses.

Maximum parsimony analysis (MP) was applied on mtDNA, nucDNA and combined sequence data using PAUP*. In all cases, heuristic searches were replicated 100 times (with random addition of taxa) using tree-bisection-reconnection (TBR) branch swapping. In the parsimony analysis of mtDNA data, equal weighting of all characters was initially used and also a transversion/transition weight of 3:1 was applied, based on the known substitution dynamics of this molecule (Ortí and Meyer, 1997). To estimate statistical support for branches, bootstrap analysis with 1000 replicates was conducted in MP analysis. To test alternative hypothesis using parsimony, I used Wilcoxon signed-ranks (WS-R) tests as implemented in PAUP*. A number of *a priori* phylogenetic hypotheses were tested (Table 4.4). To generate trees for these hypotheses, I edited the most parsimonious tree to construct topological constraints following each of the alternative hypotheses using Treeview 1.6.6. (Page, 1996), and then used parsimony searches implemented in PAUP* v4.0 to obtain the best tree that satisfied each constraint.

Maximum-likelihood analyses (ML) were performed on mtDNA, nucDNA, and concatenated data using PAUP* v4.0 with a heuristic search option, stepwise addition, 20 replications and TBR swapping. Likelihood-ratio tests implemented in the program Modeltest v3.07 (Posada and Crandall, 1998) were used to chose the best-fit model and estimate parameters for each data partition (Table 4.3). Alternative hypotheses were tested by using one-tailed Shimodaira and Hasegawa (SH) tests with 1000 RELL bootstrap replicates (Shimodaira and Hasegawa, 1999). Alternative topologies were generated using maximum likelihood by a similar process described in Wilcoxon signed-ranks tests (above). ML analyses also were implemented with a mixed model using TreeFinder (Jobb et al., 2004), in which the data were partitioned into four parts: mtDNA, 1st, 2nd, and 3rd codon positions of RAG genes, each with its own optimized parameters for GTR+ I+Γ model. Bootstrap support values for ML analyses also were calculated using TreeFinder.

Bayesian analyses were implemented on mtDNA, nucDNA, RAG protein sequences, and coded indels in several combinations. DNA sequence data were partitioned in the same way as in TreeFinder ML analyses (4 partitions: mtDNA, 1st, 2nd, and 3rd codon positions of RAG genes). The substitution model used was the general time reversible model with invariant sites and among-site variation (GTR+ I+Γ) for each partition. For RAG amino acid sequences the JTT model (Jones et al., 1992), allowing for invariant sites and among-site variation (JTT+ I+Γ) was used. This model was selected by ProtTest under the AIC and BIC criteria (Abascal et al., 2005). The coded indel data partition was analyzed using the standard discrete model (Lewis, 2001), allowing for among-site rate heterogeneity (Std+Γ). All parameters were optimized for each data partition. Bayesian analysis was run in MrBayes v3.1.2 (Ronquist and Huelsenbeck, 2003) with 4 chains. One million generations wer run with a sample frequency of 100 generations. The trees sampled before reaching stationarity of the MCMC chain were discarded for computing the consensus tree and posterior probabilities. Two independent runs were used to provide additional confirmation for the convergence of posterior probability distribution.

4.4. Results

4.4.1. Sequence variation and data partitions

Mitochondrial ribosomal 12S and 16S genes (rRNA) were sequenced for most of the 44 taxa, with a few sequences retrieved from GenBank (Table 4.1). Alignment of the 12S fragments resulted in 602 sites, corresponding to sites 1304 through 1880 of *D. rerio* mitochondrial genome (NC_002333). The alignment of 16S sequences resulted in 1384 sites, corresponding to sites 2207 through 3517 of *D. rerio* mitochondrial genome (NC_002333). The combined mtDNA data has a total of 1986 characters, with 268 characters that include alignment gaps, 823 constant characters, 158 parsimony-uninformative characters, and 737 parsimony-informative characters. The null hypothesis of base composition stationarity of variable sites in the combined data was rejected by Chi-square test (p<0.0001). The GC content at variable sites ranges from a low of 35% (*Denticeps clupeoides*) and 38% (*Spratelloides delicatulus*) to 60% (*Brevoortia tyrannus*). Plots of absolute numbers of substitutions against ML corrected sequence divergence reveal a substitution saturation pattern in mtDNA data (non-linear relationship, Fig. 4.2).

The RAG1 fragment was sequenced for all taxa examined, except for *Sundasalanx mekongensis*, *Ilisha elongata*, *Sardina pilchardus* and *Alosa aestivalis* that did not amplify, perhaps due to mutations at the priming sites in these species. Alignment of all 40 taxa includes 1734 nucleotides, spanning most of exon 3 of the RAG1 gene, corresponding to site 1540 through 3006 in *D. rerio* (U71093). One previously undescribed intron was discovered in *Spratelloides gracilis*. The intron is 228 bp long and located at site 1684 (in *D. rerio*). The hypothesis of base composition homogeneity at variable sites of RAG1 gene was rejected by a Chi-square test (p<0.0001). The percent of

G and C (GC content) at all variable sites in the taxa examined ranged from a low of 46% in *Ictalurus punctatus* to a high of 85% in *Coilia mystus*. Most of the clupeiform fishes have high GC content, close or above 70%, except for *Denticeps clupeoides* (61%) and *Spratelloides delicatulus* (59%). Base composition homogeneity for only clupeiforms and clupeiforms without *D. clupeoides* and *S. delicatulus* also is rejected (p<0.0001). The consequence of relative low GC content in *Denticeps* and *Spratelloides* on inferring their phylogenetic position will be addressed in the discussion. Plots of absolute number of substitutions against ML corrected sequence divergence reveal an almost linear relationship, indicating that there is little substitution saturation in RAG1 data (Fig. 4.2).

The RAG2 fragment was sequenced for all clupeiforms studied except for *Sundasalanx mekongensis*. Alignment of the 43 taxa yielded a length of 1647 bp, including sites 162 through 1383 corresponding to the *D. rerio* RAG2 gene (U71094). One undescribed intron was found for *Anchoa lyolepis*, spanning 390 nucleotides, located at position 1055 of the *D. rerio* gene. At the same position, another intron was found in *Spratelloides gracilis*, spanning only 226 bp and very different in sequence to the one found in *Anchoa*. Base compositional stationarity was tested for all variable sites. The null hypothesis of stationary base composition was rejected by a Chi-square test (p<0.0001). The GC content of RAG2 sequences ranged from 51% in *Hepsetus odoe* to 78% in *Sardina pilchardus*. Similar to RAG1, *S. delicatulus* (58%) and *D. clupeoides* (59%) have the lowest GC content in clupeiforms. Base composition homogeneity for only clupeiforms and clupeiforms without *D. clupeoides* and *S. delicatulus* also is rejected (p<0.0001). Plots of absolute substitutions against ML corrected sequence divergence show a linear relationship, suggesting that there is little saturation in RAG2 data (Fig 4.2).

In subsequent phylogenetic analyses, RAG1 and RAG2 were combined and analyzed together as they occur immediately adjacent to each other (Peixoto et al., 2000) and are highly correlated in GC content among taxa with a correlation coefficient R = 0.885 for my data. The homogeneity partition test also indicated they harbor congruent phylogenetic signal (p>0.05). The combined RAG data include 2763 characters without the intron sites, consisting of 1179 invariable characters, 276 parsimony-uninformative characters and 1308 parsimony-informative characters. All alignment files are available upon request.

4.4.2 Phylogenies of mtDNA and nucDNA data

Analysis of mtDNA data under unweighted parsimony resulted in 4 equally short trees with L = 5595 steps, while the 3:1 transvertion/transition ratio recovered a similar topology, with minor difference in bootstrap support value (results not shown). A phylogeny with lnL = -24921 was obtained under maximum likelihood (Fig. 4.3 left). Three shortest trees with L = 6595 steps were recovered under parsimony using RAG DNA sequences (results not shown). ML analysis resulted in a phylogeny with lnL = -34606 (Fig 4.3. right). In analyses using either mtDNA or RAG DNA data, the ML topologies shown are very similar to the MP and Bayesian trees, so only the ML phylograms are presented, but bootstrap values from MP, partitioned ML analyses (from TreeFinder), and Bayesian posterior probabilities are indicated for all nodes to show the degree of congruence among results (Fig. 4.3).

Most parts of the mtDNA tree are consistent with the RAG tree but mtDNA provides higher resolution for relationships at intermediate levels. In both trees, Ostariophysi was found as the sister group to all clupeiforms, except for *Denticeps* (Fig. 4.3). *Denticeps* formed a clade with Ostariophysi to the exclusion of clupeiforms; this unexpected result may be an artifact due to shared low GC content in *Denticeps* and ostariophysans (see discussion). Both mtDNA and RAG data supported the monophyly of Engraulidae. Within Engraulidae, monophyly of subfamily Engraulinae was highly supported, while monophyletic subfamily Coilinae was supported by mtDNA data but not by the ML analysis of RAG data (Fig. 4.3). However, MP analysis of RAG DNA data supported the monophyly of Coilinae with a bootstrap value of 59% (Table 4.5). One clade of Clupeidae, denoted as “Clupeidae I” in Fig. 4.2, was unanimously supported by MP, ML and Bayesian analysis using both mtDNA and RAG sequences. This clade includes a monophyletic group of Alosinae (*Alosa* and *Brevoortia*) joined with a

Clupeinae (*Sardina*), a monophyletic group of Pellonulinae (*Pellonula* and *Odaxothrissa*) and a clade including three genera of Clupeinae (*Opisthonema*, *Sardinella*, *Harengula*) joined with subfamily Dorosomatinae (*Dorosoma*). Both mtDNA and RAG data suggested that the other clupeids included in this study form part of a paraphyletic group, which is denoted as “Clupeidae II” (Fig. 4.3). A sister-group relationship of two Dussumieriinae genera, *Jenkinsia* and *Spratelloides* was supported (Fig. 4.3). However, the other Dussumieriinae genus *Etrumeus* was not closely related to them (Fig. 4.3). There is low resolution for nominal clupeiform families, and this is where the major discrepancies between mtDNA data and RAG data are observed. Within Clupeoidei, mtDNA data supported Engraulidae as the basal clade, while taxa assigned to Clupeidae, Pristigasteridae, and Chirocentridae were grouped together as a polytomy (Fig. 4.3, left). Analyses of RAG DNA data resulted in lower resolution among nominal families: Engraulidae, Clupeidae, and Pristigasteridae were grouped as a polytomy, while Chirocentridae and Dussumieriinae (without *Etrumeus*) were placed as a basal clade to them (Fig. 4.3, right). The discrepancies between mitochondrial mtDNA and nuclear RAG gene data are underscored by a significant result of the homogeneity partition test ($p<0.01$) and SH test. In SH tests, the RAG topology was rejected by mtDNA data ($p<0.001$), while mtDNA topology was also rejected by RAG DNA data ($p<0.001$).

4.4.3. Analysis of combined data and a priori hypothesis tests

In spite of conflicting phylogenetic signal between mtDNA and RAG DNA data, both data partitions were combined to explore further the resolution of clupeiform phylogeny. A total of 4749 nucleotide sites were concatenated from the RAG and mtDNA data partitions. The gene sequences that were unavailable for a few taxa were coded as missing data.

A single shortest tree with L = 12630 steps was found under parsimony. One tree with lnL = - 60136 was recovered under ML analysis. The Bayesian analyses produced consensus topologies highly congruent with those obtained by ML and MP analysis. The shallow clades inferred using combined data are similar to those supported by individual genes, but with higher bootstrap support (Table 4.5). The consensus tree obtained by mixed-model Bayesian analysis of three data partitions (mtDNA, RAG protein sequences, and coded indels) is shown in Figure 4.4. ML models and parameters estimated are listed in Table 4.3; a summary of support values from the diverse analyses performed on individual and combined data partitions are shown in Table 4.3. A number of a *priori* hypotheses were tested (Table 4.4). Both WS-R test and SH failed to reject the sister group relationship of Pristigasteridae + Engraulidae ($p>0.05$) and sister-group relationship of Engraulidae + Clupeidae ($p>0.05$). WS-R test rejected the sister-group relationship between Chirocentridae and Engraulidae ($p<0.05$), but the more conservative SH test failed to reject it ($p>0.05$). WS-R test rejected the monophyly of Clupeidae and monophyly of Dussumieriinae ($p<0.01$), but SH test failed to reject them. Both WS-R and SH tests rejected the monophyly of Clupeinae ($p<0.01$), the sister group relation between *Sundasalanx* and *Salangidae* ($p<0.01$) and the sister-group relationship between *Sundasalanx* and *Jenkinsia* ($p<0.01$).

4.5. Discussion

4.5.1. Compositional bias and the phylogenetic position of Denticeps

Denticeps is a small herring-like fish found in small freshwater streams from southeast Benin to northwest Cameroon. Because this fish has some rare features as a teleost, such as small tooth-like structure (odontodes) on the exposed surface of most skull roofing bones, a new family, Denticipitidae was erected for it (Clausen, 1959). In spite of some peculiar characters, *Denticeps* was thought to be a clupeomorph based on several apomorphic characters shared with clupeomorph fishes (Grande, 1985; Greenwood, 1968). For example: (i) the presence of intracranial swim bladder diverticula encased in bony bullae, (ii) Hypural 2 fused with the first ural centrum at all stages of development, and an autogenous first hypural, (iii) the presence of one or more abdominal scutes (including a pelvic scute), which are composed of a single element (Grande, 1985; Greenwood, 1968). Other characters: (iv) the presence of *recessus* *lateralis,* (v) loss of the beryciform foramen, further suggest that Denticipitidae specifically has strong affinities to Clupeiformes of Clupeomorpha (Grande, 1982, 1985). Some clupeomorph-like characters are missing in Denticipitidae, such as a well-defined pre-epiotic fossa, dorsal scutes with a median keel, but these were thought to be secondary loss (Grande, 1985; Greenwood, 1968). Other characters, such as the *recessus lateralis,* are very different in Denticipitidae compared to other clupeiforms (Di Dario, 2004; Grande, 1985; Greenwood, 1968). With no controversy, the modern taxonomy places *Denticeps* as in its own suborder within Clupeiformes (Fig. 4.1), unambiguously as the sister group to all other clupeiforms (Grande, 1985; Nelson, 1994). My phylogenetic results, grouping *Denticeps* with Ostariophysi are, therefore, surprising (Figs. 4.3 and 4.4).

As noted above, the GC content at variable sites of RAG1 and RAG2 genes for *Denticeps clupeoides* and *Spratelloides delicatulus* were the lowest two among all clupeiforms examined, and close to the low value observed among ostariophysans. The other clupeiform taxa studied have significantly higher GC content (much higher than the average actinopterygian fish, Ortí *et al*., unpublished data). This pattern is repeated, albeit at a lesser degree, in the mitochondrial genes. The stationarity of nucleotide frequencies is clearly not met by the data sets used in this study. It is well-known that biased GC content can obscure the true phylogenetic signal by erroneously joining taxa with similar GC content rather than true evolutionary relationship (Foster and Hickey, 1999; Weisburg et al., 1989). RY coding has been proposed to effectively reduce the influence of biased GC content (Phillips et al., 2004); thus, support for *Denticeps* + Ostariophysi should decrease significantly when analyzing RY-coded data, if this relationship is artificially obtained due to non-stationarity. To test this hypothesis, I calculated the branch weight (absolute number of characters that support the branch) under nucleotide-coded data (NT-coded) and RY-coded data using Spectronet (Fig. 4.5). The two competing hypotheses tested were: (*Denticeps* + Ostariophysi) vs. (*Denticeps* + clupeiforms). Under the NT-coded data, the former hypothesis has higher support, while under RY-coded data and the alternative wins, suggesting that a significant proportion of signal for the position of *Denticeps* shown in Figures 4.3 and 4.4 is due to the biased (low) GC content shared between *Denticeps* and Ostariophysi. Thus, the morphology-based hypothesis of relationships for *Denticeps* is consistent with the DNA sequence data—when the analyses correctly account for base composition bias.

4.5.2. Engraulidae, Clupeidae, and the other clupeoid taxa

Of the four families recognized for the suborder Clupeoidei (Fig. 4.1), only the monophyly of Engraulidae and Pristigasteridae (in part) are well supported by the molecular data and taxa sampled in this study (Figs. 4.3 and 4.4; Table 3). The relationships of pristigasterids (*Ilisha* and *Pellona,* 2 genera of the subfamily Pelloninae) to the other clupeoids, and among the other clupeoid taxa inferred from the molecular data are significantly different from those implied by the currently accepted classification (Fig. 4.1, Table 4.4). The most important difference is a total lack of support for the monophyly of Clupeidae, as currently recognized. The two representative species from the genus *Clupea* are weakly related to *Etrumeus*, to the exclusion of all other taxa. Elements assigned to Pristigasteridae (*Ilisha* and *Pellona*) and to Chirocentridae (*Chirocentrus*) are closely related to other taxa assigned to Clupeidae. Engraulidae (*Engraulis*, *Anchoa*, *Coilia* and *Setipinna*) is well supported as a monophyletic group by all analyses on every dataset (Figs. 4.3 and 4.4; Table 4.5). Within Engraulidae, the subfamilies Engraulinae (*Engraulis* + *Anchoa*) and Coilinae (*Coilia* + *Setipinna*) were also shown as monophyletic groups by all analysis except for ML and Bayesian trees using RAG data alone (Table 4.5). Chirocentridae and Clupeidae have been united under the superfamily Clupeiodea by an increase in rib to pleural vertebrae ratio and fusion between epicentrals and ribs (Grande, 1985; Patterson and Johnson, 1995). However, results from a new morphological study placed Chirocentridae closer to Engraulidae (Di Dario, 2005). My results show *Chirocentrus* closely related to the clupeids *Jenkinsia* and *Spratelloides* (Figs. 4.3 and 4.4, Table 4.5). This relationship also was supported by mitogenomic data (Sébastien Lavoué, personal comm.). None of My analysis placed Engraulidae as the sister taxon to Chirocentridae, however the topology tests failed to reject this hypothesis (Table 4.4). Clupeidae and Engraulidae were proposed to be more closely related to each other than to Pristigasteridae based on the presence of cartilage chevrons at the tips of epicentrals (Patterson and Johnson, 1995). Di Dario (2002) added two more characters (the orientation of parapophyses of the second vertebra and the complex pattern of interzygapophysal articulation) to support this hypothesis. However, there is also morphological evidence (the gongyloid cartilage) to support a close relationship between Pristigasteridae and Engraulidae (Di Dario, 2002). In most of my results, Pristigasteridae was closely related to taxa currently assigned to Clupeidae, with Engraulidae as a sister group to them (Figs. 4.3 and 4.4).

4.5.3. Relationships within Clupeidae

The monophyly of Clupeidae, as currently recognized, was not recovered in analyses of the molecular data sampled in this study (Figs. 4.3 and 4.4; tables 4.4 and 4.5). Similar to results obtained with mitogenomic data, my study suggests that *Chirocentrus* is nested within Clupeidae (Sébastien Lavoué, personal comm.). One group of clupeid taxa, identified as “Clupeidae I” (Figs. 4.3 and 4.4), was strongly supported by my data. This clade includes *Dorosoma*, closely related to three representatives of the currently recognized Clupeinae (*Sardinella*, *Opisthonema* and *Harengula*), and the Pellonulinae (*Pellonula* and *Odaxothrissa*). The second component of “Clupeidae I” are taxa currently assigned to Alosinae (*Alosa* and *Brevoortia*) plus *Sardina*. This group (Clupeidae I) also is supported by morphology of the gill arches (Nelson, 1970) and the results from mitogenomic data (Sébastien Lavoué, personal comm.). The other clupeids sampled in my study (“Clupeidae II” in Fig. 4.3), which include *Etrumeus*, *Jenkinsia*, *Spratelloides* and *Clupea*, are closely related to Pristigasteridae and Chirocentridae, (Figs. 4.3 and 4.4), but do not form a monophyletic group. A close relationship among *Etrumeus*, *Jenkinsia*, *Spratelloides* and *Clupea* was proposed by Nelson (Nelson, 1967; 1970) based on the foramen in the fourth epibranchial. My results show clearly that *Etrumeus* is not in the same clade as *Jenkinsia* and *Spratelloides* (Figs. 4.3 and 4.4), but it groups instead with *Clupea*, albeit with low support (Fig. 4.4, Table 4.5). Polyphyly of Dussumieriinae was proposed earlier based on the shape of the hymandibular bone (Eschmeyer, personal comm.).

4.5.4. Phylogenetic position of Sundasalanx

Sundasalanx are miniature, transparent and highly paedomorphic freshwater fish distributed in Southeast Asia. Because they have unusual characters among teleosts, such as a pectoral girdle with a median cartilaginous scapulocoracoid, Roberts (1981) erected a new family, Sundasalangidae. This family was thought to be closely related to Salangidae because they share some features, such as a single cartilaginous jaw suspension, well-developed separate fourth hypobranchials, pedunculate pectoral fins, no symplectics, no circumorbital bones, and muscles failing to meet at the ventral midline (Roberts, 1981; 1984). Fink (1984a) further included Sundasalanx in the family Salangidae, while Nelson (1994) listed the Sundasalangidae as an osmerid family. Siebert (1997) described four new species of *Sundasalanx* from Borneo, and proposed a new radical hypothesis of relationships. By closer examination of the evidence presented by Roberts, Siebert (1997) found that these characters were all paedomorphic and also plesiomorphic, being features found in larvae of lower teleosts and some euteleosts. The only few non-paedomorphic features of *Sundasalanx* include ossified prootic bulla (apomorphic for clupeomorphs) and *recessus lateralis* (apomorphic for clupeiforms), indicating a relationship of *Sundasalanx* to clupeomorph fishes. Siebert (1997) went even further to suggest that *Sundasalanx* is a spratelloidin, and possibly the sister-group of the Caribbean endemic genus *Jenkinsia*, since they both exhibit a derived, highly consolidated, caudal skeleton. Britz and Kottelat (1999) suggested that Siebert's conclusion was premature because there was not enough evidence to demonstrate shared derived characters of *Sundasalanx* and Clupeidae. In my results, *Sundasalanx* was not supported as the sister taxon to *Salangichthys* (Table 4.4), but was highly supported as a clupeiform with bootstrap 100% for MP analysis, 100% for ML analysis and 1.0 for Bayesian posterior probability. *Sundasalanx* was found closely related to my “Clupeidae” I (Figs. 4.3 and 4.4) with marginal support (only mtDNA supported this relationship, RAG data were not possible to obtain in this study). A close relationship between *Sundasalanx* and *Jenkinsia* was rejected by both the WS-R test and SH test ($p<0.01$, Table 4.4), against the hypothesis of Siebert (1997). Although closely related to clupeids, the precise phylogenetic position of *Sundasalanx* within Clupeidae remains uncertain and requires further study.

Table 4.1. Taxon sampling for clupeiforms.

Taxa used	RAG1	RAG2	12S	16S	Museum/ tissue no.
Outgroup					
Oncorhynchus mykiss	AF137176[*]	U31670[*]	NC_001717[*]	NC_001717[*]	-
Salangichthys microdon	AY380539[*]	-	NC_004599[*]	NC_004599[*]	-
Danio rerio	U71093[*]	U71094[*]	NC_002333[*]	NC_002333[*]	-
Cyprinus carpio	AY787040[*]	AY787041[*]	NC_001606[*]	NC_001606[*]	-
Ictalurus punctatus	AY423859[*]	AY184245[*]	NC_003489[*]	NC_003489[*]	-
Hepsetus odoe	DQ912097	AY804086[*]	U33825[*]	AY788030[*]	GO126
Distichodus sp.	DQ912098	AY804071[*]	U33827[*]	AY788012[*]	GO196
Denticipitidae (1 genus)					
Denticeps clupeoides	DQ912100	DQ912133	DQ912028	DQ912063	NSMT-P68224
Engraulidae					
Engraulinae (12 genera)					
Anchoa delicatissima	DQ912108	DQ912141	DQ912036	DQ912071	T510
Anchoa hepsetus	DQ912112	DQ912145	DQ912040	DQ912075	T1212
Anchoa mitchilli	DQ912113	DQ912147	DQ912042	DQ912077	C1507
Anchoa choerostoma	DQ912119	DQ912153	DQ912048	DQ912083	T3895
Anchoa lyolepis	DQ912120	DQ912154	DQ912049	DQ912084	T5152
Engraulis encrasicolus	DQ912103	DQ912136	DQ912031	DQ912066	No voucher[a]
Engraulis mordax	DQ912109	DQ912142	DQ912037	DQ912072	T550
Engraulis eurystole	DQ912121	DQ912155	DQ912050	DQ912085	T5153
Coilinae (5 genera)					
Coilia nasus	DQ912123	DQ912157	DQ912052	DQ912087	No voucher[b]
Coilia brachygnathus	DQ912124	DQ912159	DQ912054	DQ912089	No voucher[b]
Coilia mystus	DQ912126	DQ912162	DQ912057	DQ912092	No voucher[b]
Setipinna taty	DQ912125	DQ912161	DQ912056	DQ912091	No voucher[b]
Clupeidae					
Alosinae (7 genera)					
Alosa aestivalis	-	DQ912146	DQ912041	DQ912076	T1504
Alosa pseudoharengus	DQ912115	DQ912149	DQ912044	DQ912079	T1585
Alosa sapidissima	DQ912116	DQ912150	DQ912045	DQ912080	T1586
Alosa chrysochloris	DQ912117	DQ912151	DQ912046	DQ912081	T1910
Brevoortia patronus	DQ912105	DQ912138	DQ912033	DQ912068	GO602

Table 4.1. Taxon sampling for clupeiforms (cont.).

Taxa used	RAG1	RAG2	12S	16S	Museum/ tissue no.
Brevoortia tyrannus	DQ912106	DQ912139	DQ912034	DQ912069	GO676
Clupeinae (16 genera)					
Clupea harengus	DQ912114	DQ912148	DQ912043	DQ912078	T1583
Clupea pallasii	DQ912118	DQ912152	DQ912047	DQ912082	T3204
Harengula jaguana	DQ912122	DQ912156	DQ912051	DQ912086	T6543
Opisthonema oglinum	DQ912111	DQ912144	DQ912039	DQ912074	T1192
Sardina pilchardus	-	DQ912158	DQ912053	DQ912088	No voucher[a]
Sardinella aurita	DQ912104	DQ912137	DQ912032	DQ912067	GO598
Dorosomatinae (6 genera)					
Dorosoma cepedianum	DQ912099	DQ912132	DQ912027	DQ912062	No voucher[c]
Dussumieriinae (4 genera)					
Etrumeus teres	DQ912110	DQ912143	DQ912038	DQ912073	T1052
Jenkinsia lamprotaenia	DQ912107	DQ912140	DQ912035	DQ912070	T216
Spratelloides delicatulus	DQ912128	DQ912164	DQ912058	DQ912093	No voucher
Spratelloides gracilis	DQ912129	DQ912165	DQ912059	DQ912094	No voucher
Pellonulinae (23 genera)					
Pellonula leonensis	DQ912130	DQ912166	DQ912060	DQ912095	No voucher
Odaxothrissa vittata	DQ912131	DQ912167	DQ912061	DQ912096	No voucher
Sundasalangidae (1 genus)					
Sundasalanx mekongensis	-	-	AP006232*	AP006232*	No voucher
Chirocentridae (1 genus)					
Chirocentrus dorab	DQ912127	DQ912163	AP006229*	AP006229*	No voucher
Pristigasteridae (9 genera)					
Ilisha elongata	-	DQ912160	DQ912055	DQ912090	No voucher[b]
Pellona flavipinnis	DQ912101	DQ912134	DQ912029	DQ912064	GO309
Pellona castelnaeana	DQ912102	DQ912135	DQ912030	DQ912065	GO325

*Sequences taken from GenBank.

NSMT number: Vouchers from the National Science Museum, Tokyo; C and T number: Vouchers from The University of Kansas Natural History Museum, KS, USA; GO number: Tissue collection of G. Ortí, University of Nebraska, Lincoln, NE, USA; [a]Tissue samples provided by W. Chen, Saint Louis University, MO, USA; [b]Collected from Pudong, Shanghai, China; [c]Collected from Hershey, Nebraska, USA.

Table 4.2. Primers for PCR and sequencing for Clupeiformes in Chapter four.

Primers	Sequences	Reference
For 12S		
12S229F[a]	5’ GYCGGTAAAAYTCGTGCCAG 3’	This study
12S954R[a]	5’ YCCAAGYGCACCTTCCGGTA 3’	This study
For 16S		
16S135F[a]	5’ GCAATAGAVAWAGTACCGCAAGG 3’	This study
16S964F[a]	5’ YTCGCCTGTTTACCAAAAAC 3’	This study
16S1072R	5’ CCTTYGCACGGTYARAATAC 3’	This study
For RAG1		
RAG1-2510F[a]	5’ TGGCCATCCGGGTMAACAC 3’	This study
RAG1-2533F[b]	5’ CTGAGCTGCAGTCAGTACCATAAGATGT 3’	(Lopez et al., 2004)
RAG1-3098F	5’ TGTGCCTGATGYTYGTDGAYGART 3’	This study
RAG1-3222F	5’ TCYTTCCGCTTYCACTTCCG 3’	This study
RAG1-3261R	5’ CCCTCCATYTCNCGMACCATCTT 3’	This study
RAG1-3543R	5’ GTRGCRTTGCCRATRTCRCAGT 3’	This study
RAG1-4063R	5’ TTCTGNARRTACTTGGARGTGTAWAGCCA 3’	This study
RAG1-4078R[b]	5’ TGAGCCTCCATGAACTTCTGAAGRTAYTT 3’	(Lopez et al., 2004)
RAG1-4090R[a]	5’ CTGAGTCCTTGTGAGCTTCCATRAAYTT 3’	(Lopez et al., 2004)
For RAG2		
RAG2-F1[a]	5’ TTYGGNCARAARGGVTGGC 3’	This study
RAG2-F2[b]	5’ AARCGCTCMTGTCCMACTGG 3’	(Lovejoy and Collette, 2001)
RAG2-526F	5’ GTGGACTGCCCCCCKMAGGTSTT 3’	This study
RAG2-1096F	5’ CAGGGCTRCAGCCAGGARTC 3’	This study
RAG2-514R	5’ CAGTCCACCAYRCTGTTCCA 3’	This study
RAG2-1145R	5’ AAGTAGAGCTCCTCNGAGTCC 3’	This study
RAG2-R6	5’ TGRTCCARGCAGAAGTACTTG 3’	(Lovejoy and Collette, 2001)
RAG2-1466R[b]	5’CCRTGRTCCARGCAGAAGTACTT 3’	This study
RAG2-1453R[a]	5’CCRTGRTCCARGCAGAAGTA 3’	This study

[a] Primers for first-round PCR, [b] Primers for second round nested-PCR; the other primers were used for sequencing only. All PCR and nested-PCR primers also were used for sequencing.

Table 4.3. Best-fit models selected by likelihood-ratio tests or the AIC implemented in Modeltest v3.07 (for DNA sequences) or ProtTest (protein sequences). Parameters for DNA models estimated by Modeltest (PAUP*), and for protein and indel models by MrBayes.

Data Partition	ML model	Estimated base frequencies	Substitution rate matrix	Invariable sites (%) (I)	Gamma-shape parameter (α)
mtDNA	GTR+I+Γ	A = 0.3666 C = 0.2585 G = 0.1796 T = 0.1952	r_{A-C} = 2.1352 r_{A-G} = 6.5955 r_{A-T} = 2.9311 r_{C-G} = 0.4985 r_{C-T} = 16.553 r_{G-T} = 1.0000	0.2922	0.5599
RAG (DNA)	GTR+ I+Γ	A = 0.2199 C = 0.2945 G = 0.2711 T = 0.2146	r_{A-C} = 1.3537 r_{A-G} = 4.0185 r_{A-T} = 1.7633 r_{C-G} = 1.0440 r_{C-T} = 5.0255 r_{G-T} = 1.0000	0.3401	1.1813
RAG (protein)	JTT	fixed	fixed	0.29	0.98
Coded Indels	Standard	fixed	fixed	0	0.81

Table 4.4. Maximum parsimony Wilcoxon signed-ranks test and maximum likelihood-based Shimodaira-Hasegawa test of priori hypotheses. Using combined mtDNA and RAG DNA sequences.

Hypotheses tested	References	WS-R[a]	SH[b]
Pristigasteridae + Engraulidae	(Di Dario, 2002)	0.155	0.480
Clupeidae + Engraulidae	(Di Dario, 2002; Patterson and Johnson, 1995)	0.121	0.650
Chirocentridae +Engraulidae	(Di Dario, 2005)	0.034*	0.302
Clupeidae monophyly	(Grande, 1985; Nelson, 2006)	0.001*	0.221
Clupeinae monophyly	(Nelson, 2006)	0.000*	0.000*
Dussumieriinae monophyly	(Grande, 1985; Nelson, 2006)	0.000*	0.462
Sundasalanx + Salangidae	(Roberts, 1981)	0.000*	0.002*
Sundasalanx + *Jenkinsia*	(Siebert, 1997)	0.000*	0.003*

[a] Parsimony-based Wilcoxon signed-ranks test using a one-tailed probability (Templeton, 1983).

[b] Likelihood-based SH test using a one-tailed probability (Shimodaira and Hasegawa, 1999).

*Significant difference at $p<0.05$.

Table 4.5. Support values for major clades recovered in phylogenetic analyses on mtDNA and RAG gene sequences.

Taxon	mtDNA[1]	RAG DNA[1]	mtDNA + RAG DNA[3]	mtDNA + RAGprot[3]	mtDNA + RAGprot+ indels[2]	mtDNA + RAG RYcoding[3]
	MP/ML/MB	MP/ML/MB	MP/ML/MB	MB	MB	MP/ML
Engraulidae	100/100 /1.0	99/100/1.0	100/100/1.0	1.0	1.0	99/100
Engraulinae	100/100/1.0	100/100/1.0	100/100/1.0	1.0	1.0	100/100
Colilinae	100/100/1.0	59 / * / *	100/95/1.0	1.0	1.0	97/70
Clupeidae + Pristigasteridae + Chirocentridae	52/78/1.0	* / * / *	51/ * /0.95	0.94	1.0	*/63
Clupea + *Etrumeus*	* / * / *	* / * / 1.0	* /63/ *	1.0	0.95	*/*
Pristigasteridae	100/100/1.0	100/ * /1.0	100/100/1.0	1.0	1.0	100/100
Chirocentrus + *Jenkinsia* + *Spratelloides*	100/100/1.0	* / * / *	80/54/0.99	1.0	1.0	60/77
"Clupeidae I"	59/68/1.0	64/ * /1.0	86/55/1.0	1.0	0.99	94/97
Alosinae + *Sardina*	91/99/0.98	100/55/1.0	100/100/1.0	1.0	1.0	99/100
Dorosomatinae + Pellonulinae + *Sardinella* + *Harengula* + *Opisthonema*	88/99/1.0	100/100/1.0	98/100/1.0	1.0	1.0	95/100

MP: bootstrap values from MP analysis; ML: bootstrap values from ML analysis; MB: posterior probabilities from Bayesian analysis. 1 values from Figure 2; 2 values from Figure 3. 3 phylogenetic trees not shown. The asterisks indicate the nodes with bootstrap support lower than 50%, posterior probability <0.9, or nodes that were not recovered in that analysis.

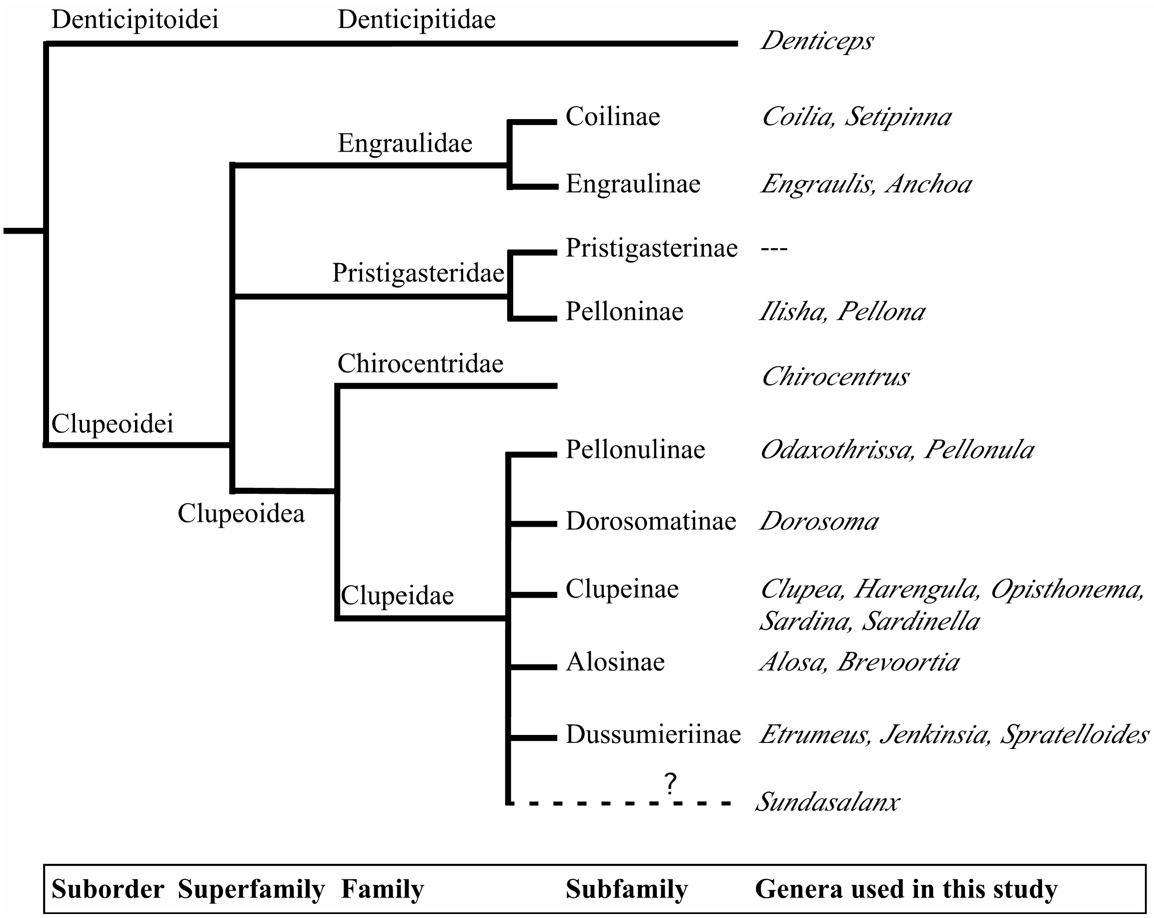

Fig. 4.1. Current classification of Clupeiformes (Froese and Pauly, 2004; Grande, 1985; Nelson, 1994; Whitehead, 1985).

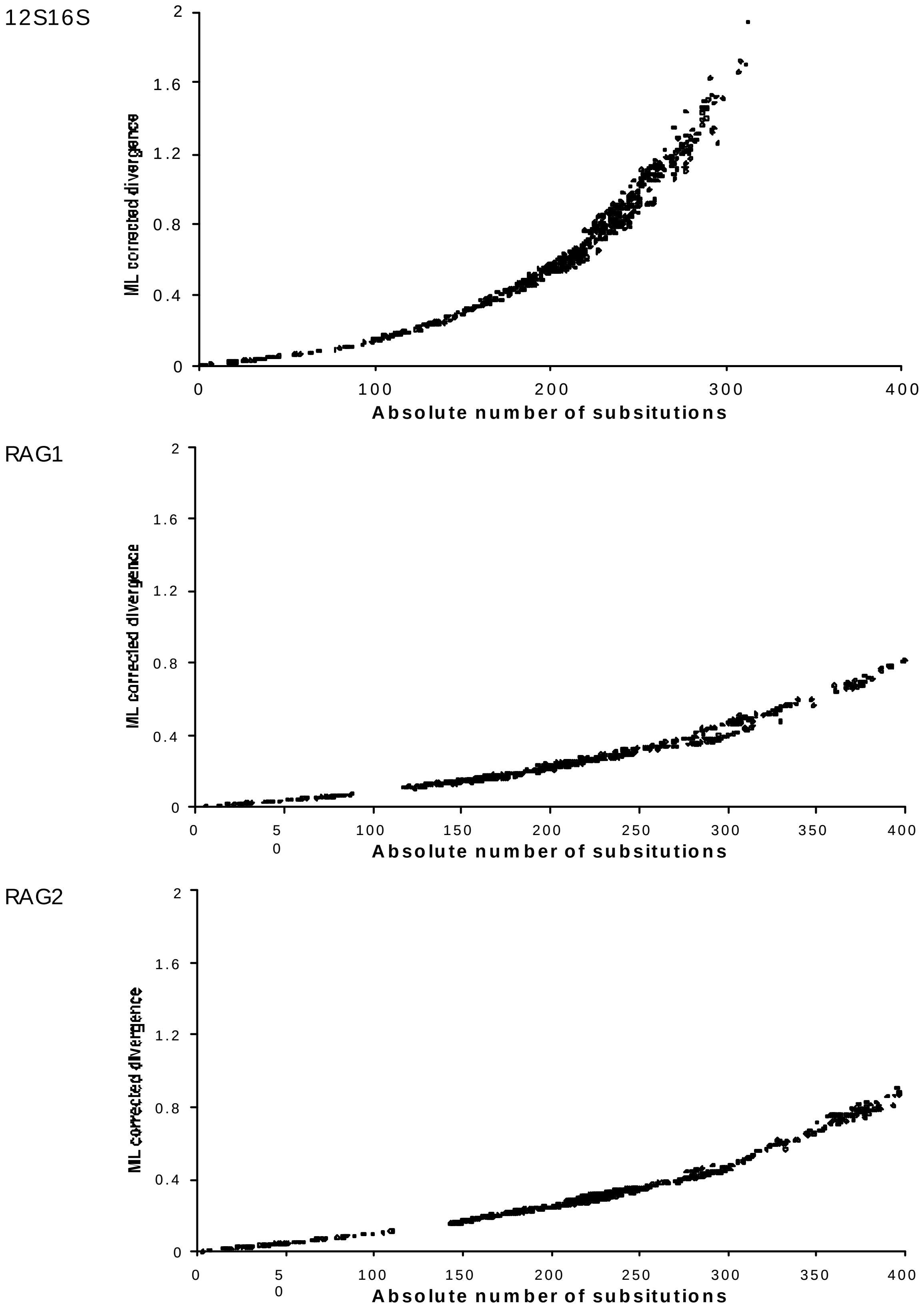

Fig. 4.2 Plots of absolute subsitutions against ML corrected divergences.

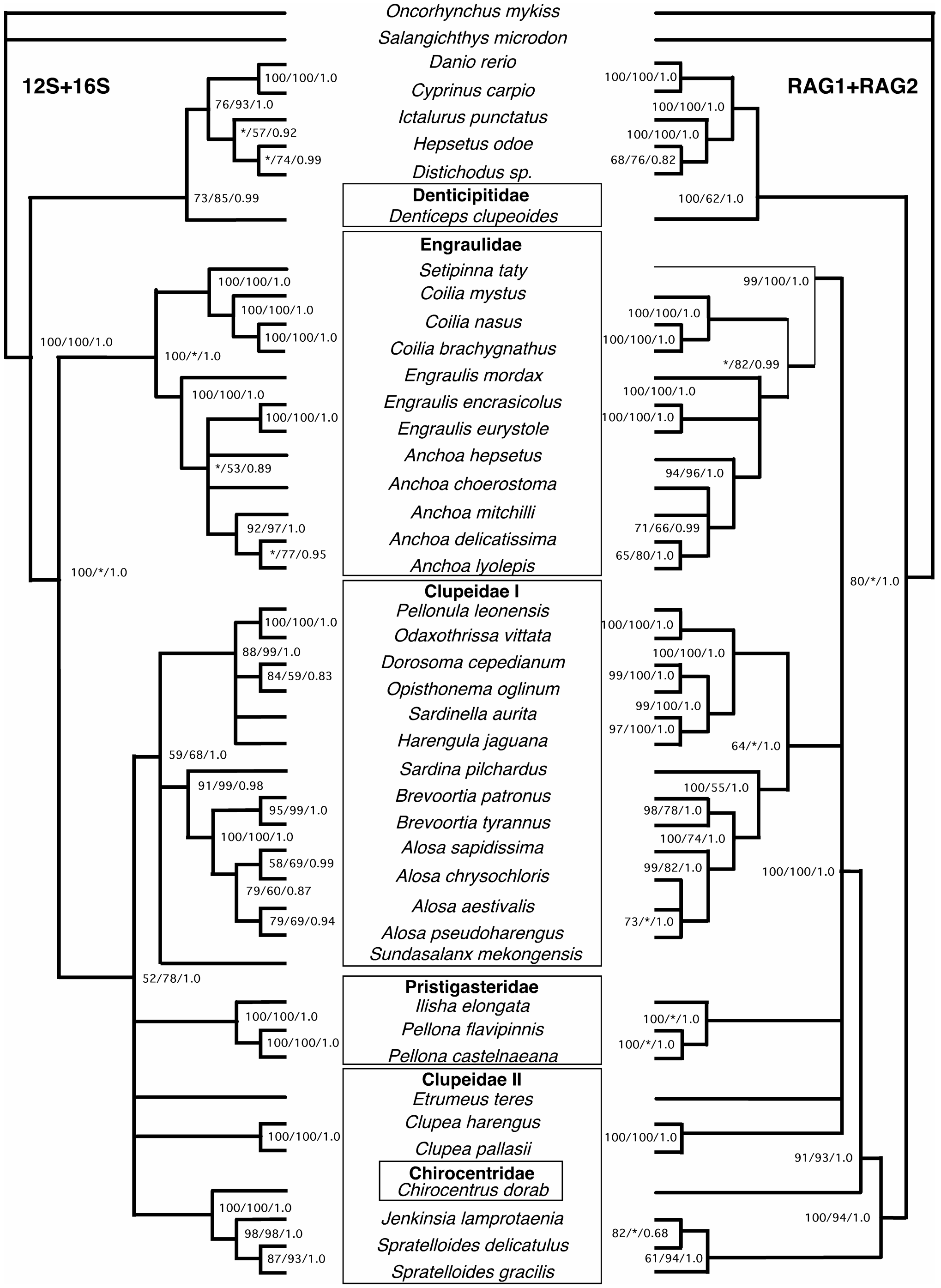

Fig. 4.3. Maximum likelihood trees from the analysis on mtDNA sequences (1986 bp, left) and RAG nucDNA (2763 bp, right). The numbers on branches are MP bootstrap values (>50%), partitioned ML bootstrap values from TreeFinder (>50%), and Bayesian posterior probabilities, from left to right, respectively. Branches with low support (<50% for bootstrap or <0.9 for Bayesian posterior probabilities) in more than two of the three analyses (MP, ML and Bayesian analysis) were collapsed. The asterisks indicate bootstrap values smaller than 50% or posterior probability <0.9.

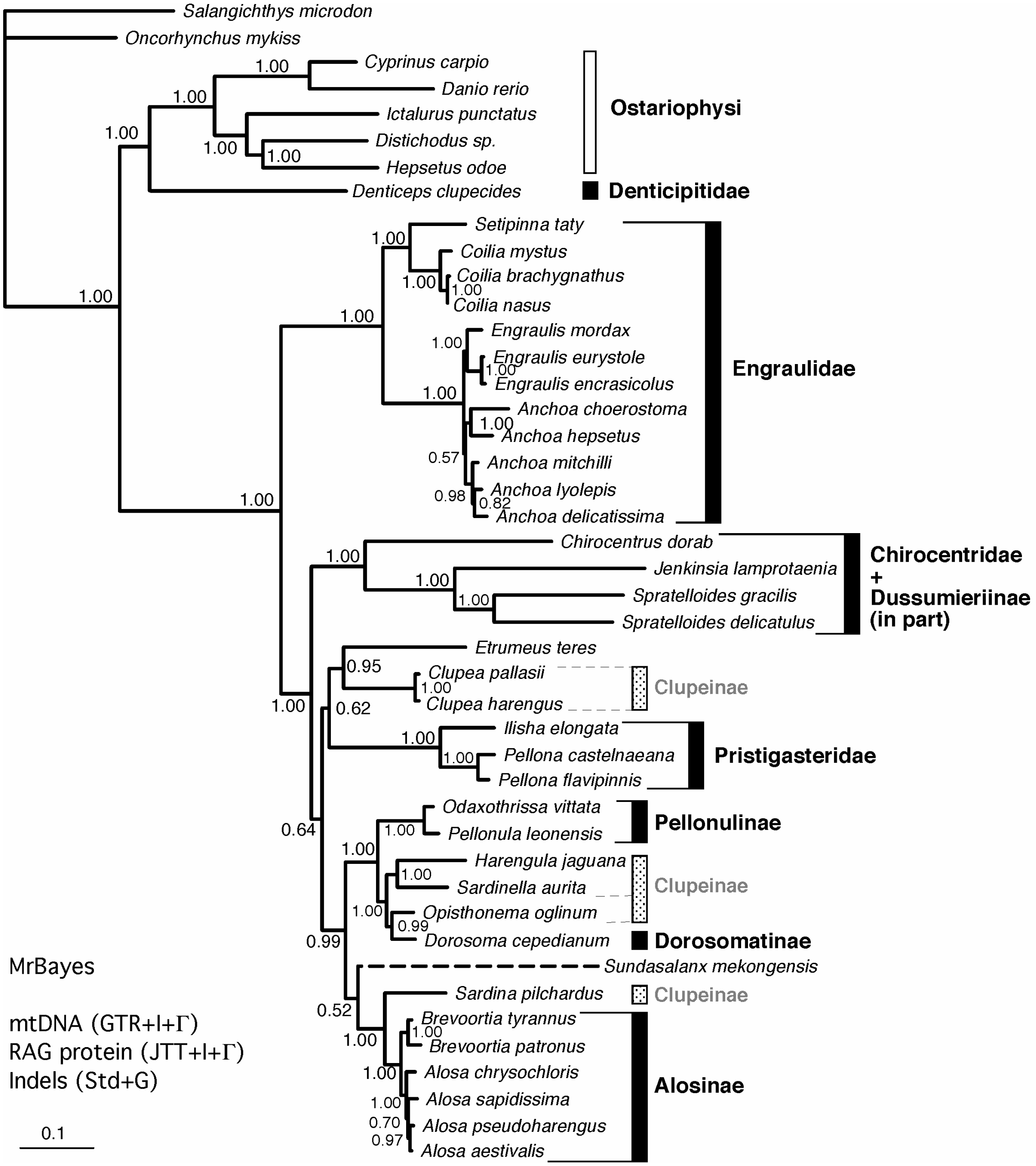

Fig. 4.4. Consensus phylogram of 3000 post-burnin trees obtained with mixed-model Bayesian analysis. MtDNA sequences (1986 bp) were analyzed under the GTR+I+Γmodel, RAG protein sequences (921 amino acids) under the JTT+I+G model, and coded indels (137 characters) under the Std+G model (for parameters see Table 4.2). Posterior probabilities are indicated next to the nodes.

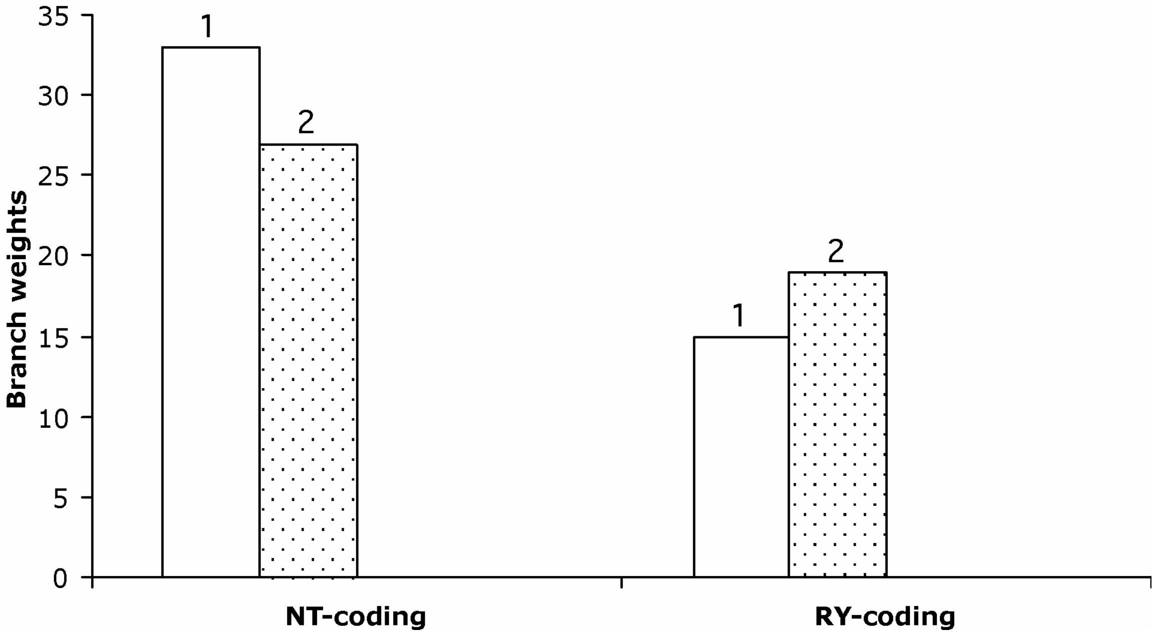

Figure 4.5. Signal supporting competitive branches under NT-coding and RY-coding. Branch weights are number of characters supporting the splits calculated in Spectronet (Huber et al., 2002). The competing hypotheses are: 1. *Denticeps* + Ostariophysi; 2. *Denticeps* + Clupeiformes.

Chapter 5 – The interrelationships of Clupeiformes: improved resolution based on ten loci

5.1. Abstract

As a sequel of Chapter four, eight more species including six more genera were sampled and six newly developed loci were sequenced to improve the resolution of phylogeny of clupeiforms recovered in Chapter four. With 25% missing data, the concatenated sequences resulted in 9963 sites. Adding these new data increased the resolution of the phylogeny of clupeiforms. The major changes in the new results include the basal position of dussumieriids to all the rest of clupeioids (clupeiforms excluding *Denticeps*) and the sister-group relationships between Engraulidae and a clade composed of pristigasterids, clupeids and *chirocentrus*. The basal position of dussumieriids has been shown not resulted from artifacts because of biased GC composition. The difference between the results based on ten loci and the results based on rDNA and RAGs along maybe due to the overwhelming signal in rDNA. However, the missing data should be determined before the discrepancies can be confidently resolved. The phylogenetic positions of all newly added taxa in this study were clearly identified

5.2. Background

One well-recognized difficult situation in phylogenetic inference is when there are short internal branches buried deeply in the tree and followed by subsequent long terminal branches (Rokas and Carroll, 2006; Rokas et al., 2005; Weisrock et al., 2005). Because of the short period of time corresponding to the short internal branches, there were few synapmorphies accumulated, while the subsequent long terminal branches may introduced parallel substitutions or multiple substitutions, further blurring the phylogenetic signal (Rokas and Carroll, 2006; Weisrock et al., 2005). This short-internal-branch situation often led to long branch attraction (Felsenstein, 1978) or low resolutions (Rokas et al., 2005), which was observed in the phylogeny of clupeiforms obtained in Chapter four. To improve the resolutions and avoid the miss-leading effects, a large sequence matrix of many nuclear loci should be assembled.

As a following up study, I collected sequences from six newly developed nuclear loci to address two questions that have not been resolved in Chapter four. First, the resolution of interrelationships among families of clupeiforms was low in the results of Chapter four. The phylogenies among some families were either not resolved in separate analyses of mtDNA or nuclear genes or received very low support in the combined analysis (see Chapter four). Including more nuclear loci is expected to improve the resolution. The second question is the discrepancy between the results from mtDNA and the nuclear genes. The mtDNA 12S and 16S data supported engraulids as the basal group to all the rest of clupeiods (clupeiforms excluding *Denticeps*), while the nuclear RAG1 and RAG2 gene suggested the dussumieriids as the basal group. The combined analysis using mtDNA, RAG proteins and indels yielded a phylogeny similar to the phylogeny based on mtDNA alone but a with better resolution (see Chapter four). The phylogeny resulted from the RAG genes could be explained by the biased GC composition of RAG genes in dussumieriids. However, the results of combined analysis could also be overwhelmed by the fast evolving mtDNA genes. In the present study, more nuclear loci were sequenced and the RY-coding method, which is the less sensitive to composition bias, was carried to test the two alternative hypotheses supported by mtDNA and nuclear genes. In addition to the two major questions asked above, eight more taxa were sampled and their phylogenetic positions were examined.

5.3. Materials and methods

5.3.1. Taxon sampling

Taxon sampling of this study was expanded from Chapter four. Eight new taxa including six new genera were included, which were not sampled in Chapter four. The new taxa sampled in this study are *Dorosoma petenense*, *Anodontostoma chacunda*, *Nematalosa japonica*, *Ethmalosa fimbriata*, *Sardinella maderensis*, *Clupeonella cultriventris*, *Ilisha elongata* and *Sprattus sprattus*. When there are more than two species available for certain genera, only two species were used. A total of 44 taxa, including six outgroups were used in this study (Table 5.1).

5.3.2. PCR amplification, sequencing and alignment

Ten gene markers were used in this study, including 12S, 16S, RAG1, RAG2 and six newly developed nuclear loci. The six nuclear loci used are zic1, RYR3, ENC1, Gylt, plagl2 and Sreb2. The primers for PCR and sequencing and the conditions for PCR reactions followed Chapter two and Chapter four. Ribosomal DNA sequences were aligned directly using ClustalW (Thompson et al., 1994), whereas the nuclear gene sequences were translated into amino acids before alignment.

5.3.3. Sequence descriptions and phylogenetic analysis

Aligned sequences were examined and the average p-distance, consistency index were calculated using PAUP (Swofford, 2003). The relative evolutionary rate for each loci were estimated using ML method implemented in TreeFinder (Jobb, 2006).

Partitioned ML analysis and Bayesian analysis were performed on concatenated DNA sequences using TreeFinder (Jobb, 2006) and MrBayes (Ronquist and Huelsenbeck, 2003). Because there are 25% percent missing data, the common way of data partitioning was followed, that is by genes for the ten loci and also by codons for the protein coding genes. The GTR + G + I model was chosen according the AIC values. Two hundreds bootstrap analysis were executed in ML analysis using TreeFinder. Two millions of iterations with four chains were run in the Bayesian analysis. The consensus tree and posterior estimations for parameters were calculated after the non-stationary samples were discarded using burnin step. Two independent runs were done for Bayesian analysis to ensure the convergence of the MCMC chains. To test the effect of biased base composition, RY-coding, the method has been shown less sensitive to GC bias (Phillips et al., 2004; Phillips and Penny, 2003) was performed in the ML analysis implemented in TreeFinder.

5.4. Results

5.4.1. Characteristics of the ten loci sequenced

All 44 taxa were sequenced for 12S and 16S. The next gene with most of the taxa sequenced is zic1, followed by RAG1 (Table 2). The gene with the least number of taxa sequenced is plagl2, with only 20 from the 44 taxa sequenced. The average percentage of missing data is 25% (Table 2). Most of the missing data are probably due to the mutations in the priming sites. New primers should be designed to amplify the missing fragments in the future. The evolutionary rates are faster in mtDNA than in nuclear loci, while the consistency index are higher in the nuclear genes than in mtDNA genes (Table 2). The other general characteristics of every locus are listed in Table 5.2.

5.4.2. Interrelationships among clupeiforms

Both ML and Bayesian analysis produce similar phylogeny (Fig. 5.1). *Denticeps* is grouped with ostariophysans. Dussmieriids are placed as the basal group to the rest of clupeiods (Fig 5.1). Within clupeiods, three monophyly groups are well supported: monophyly of engraulids, monophyly of “clupeids I” (see Chapter four) and monophyly of a clade composing “clupeids II” (see Chapter four), prestigasterids and *Etrumeus*. The relationships among these three major clades are resolved using ML method (Fig 5.2), but the relationships have low supports from Bayesian approach (results not shown), so they are described as a polytomy in Fig 5.1. When only rDNA and RAG genes were used to construct the phylogeny, no resolution was obtained among major clades of clupeiforms (Fig. 5.2).

5.5. Discussion

5.5.1. Phylogenetic positions of new samples

The phylogenetic positions of all eight species added in this study are clearly identified with high bootstrap values. *Dorosoma petenense* is grouped with the other *Dorosoma*, while *Anodontostoma chacunda* and *Nematalosa japonica* form a sister-group and together join the clade composing *Dorosoma* and *Opisthonema* (Fig. 5.1). *Ethmalosa fimbriata* and *Sardinella maderensis* join at the basal of the clade, “Dorosomatinae” (Fig. 5.1). Surprisingly, *Sardinella maderensis* and *S. aurita* do not form a monophylytic group, thus samples from more individuals and more species of *Sardinella* should be used to examine the relationships within this genus. The Caspian *Clupeonella cultriventris* form a group with *Sundasalanx mekongensis* with a 95% bootstrap value and a 1.0 posterior probability, and they are grouped with Alosinae (Fig. 5.1). In Chapter four, *S. mekongenesis* also was found as the basal taxa to Alosinae but only with a 0.52 posterior probability (see Chapter four). *Illisha africana* is supported as the sister taxa to *Illisha elongata* and *Pellona*, but it does not form a monophylytic group with *Illisha elongata* (Fig. 5.1). Grande (1985) also proposed that the genus *Illisha* might not be a monophylytic group. More samples and data need to be collected before a revision for *Illisha* can be done. *Sprattus sprattus* is found closely related to *Clupea* and together they form a group with *Etrumeus teres* (Fig. 5.1).

5.5.2. Basal position of dussumieriids, GC content and RY-coding analysis

The major difference between the results of mtDNA and nuclear RAGs DNA is the position of dussumieriids (*Spratelloides* and *Jenkinsia*). The mtDNA data supported engraulids as the basal group to the rest of clupeiods, while the RAG gene sequences supported dussumieriids as the basal group (see Chapter 4). With six new nuclear loci added, the data with 9963 sites highly support dussumieriids as the basal group to the rest of clupeiods with a bootstrap value of 99% and a Bayesian posterior probability of 1.0 (Fig. 5.1). The GC contents of RAGs in *Jenkinsia* and *Spratelloides delicatulus* are lower than the average of clupeiforms, which could misled the phylogenetic inference (Table 5.3, also see Chapter four). However, the GC contents of the other six nuclear genes used in the present study do not show much difference between dussumieriids and the other clupeiforms (Table 5.3). To further test the potential effects from biased GC content, I also analyzed RY-coded data using ML method implemented in TreeFinder. The only difference between the results of RY-coding and regular nucleotide coding is that *Denticeps* swaps to the basal of clupeiforms instead of grouping with ostariophysans. However, dussumieriids still are highly supported as the basal clupeiods (results not shown). The results from RY-coded data suggest that the basal position of dussmieriids is not an artifact from the biased GC content in RAG genes. Resolving the discrepancies between the results of mtDNA and nuclear DNA should await determining the missing data in nuclear loci.

5.5.3. Improved resolution in phylogeny of clupeiforms

Short internal branches buried in deep time causes a dilemma in phylogenetic inference (Rokas and Carroll, 2006). Because of the short time between speciation events around the short branches, fast-evolving markers are preferred to obtain enough synapmorphies to construct a significant non-zero branch. At the same time, slow-evolving markers are better choices to avoid the noise being introduced along the subsequent long terminal branches. One solution to this problem is to use many of the slow markers, like the protein coding nuclear genes. Because these markers have a slow evolutionary rate, they would have less problem with saturation and homoplasy than fast-evolving markers, such as mtDNA. If the number of characters is large enough, a good number of phylogenetic informative characters should be found even on these short branches. In the present study, I test this approach by comparing the phylogeny constructed using mtDNA and RAGs alone and the phylogeny based on all ten loci. The results show that when six more nuclear loci were added into the data matrix, a better resolution was achieved (Fig 5.2). When the tree was built with mtDNA and RAGs alone, all four major lineages of clupeiods formed a polytomy (Fig 5.2 right). In the phylogeny based on all ten loci, dussimieriids were well supported as the basal clupeiods and engraulids were grouped with pristigasterids, *chirocentrus* and some clupeids although with low support (Fig 5.2 left). The results suggest that including more nuclear protein-coding genes may improve the resolutions even for those short branches buried deep in time.

Table 5.1. Clupeiforms and outgroups sequenced for the ten loci.

Genus	Species	12s	16s	Rag1	Rag2	A	A5	L	M	U	Y
Hiodon	*alosoides*	NC_005145	NC_005145	Y	Y	Y	Y	Y	-	Y	Y
Danio	*rerio*	NC_002333	NC_002333	U71093	U71094	Y	Y	Y	Y	Y	Y
Ictalurus	*punctatus*	NC_003489	NC_003489	DQ492511	DQ492398	Y	Y	Y	Y	Y	Y
Chanos	*chanos*	NC_004693	NC_004693	Y	-	Y	-	Y	-	Y	Y
Pygocentrus	*nattereri*	Y	Y	Y	Y	-	Y	Y	-	Y	Y
Apteronotus	*albifrons*	NC_004692	NC_004692	-	-	Y	-	Y	-	Y	Y
Dorosoma	*cepedianum*	Y	Y	Y	Y	Y	Y	-	Y	-	Y
Denticeps	*clupeoides*	Y	Y	Y	Y	-	Y	-	Y	-	Y
Pellona	*flavipinnis*	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Pellona	*castelnaeana*	Y	Y	Y	Y	Y	-	-	Y	-	-
Sardinella	*aurita*	Y	Y	Y	Y	Y	-	-	-	-	Y
Brevoortia	*patronus*	Y	Y	Y	Y	Y	Y	-	Y	-	Y
Brevoortia	*tyrannus*	Y	Y	Y	Y	Y	Y	Y	Y	-	Y
Jenkinsia	*lamprotaenia*	Y	Y	Y	Y	Y	Y	-	-	-	Y
Anchoa	*delicatissima*	Y	Y	Y	Y	Y	Y	-	-	-	-
Engraulis	*mordax*	Y	Y	Y	Y	Y	-	-	-	Y	-
Etrumeus	*teres*	Y	Y	Y	Y	Y	Y	-	Y	-	Y
Opisthonema	*oglinum*	Y	Y	Y	Y	Y	Y	Y	Y	-	Y
Anchoa	*mitchilli*	Y	Y	Y	Y	Y	Y	-	-	-	-
Clupea	*harengus*	Y	Y	Y	Y	Y	Y	-	Y	Y	Y

Table 5.1. Clupeiforms and outgroups sequenced for 10 loci. (cont.).

Genus	**Species**	**12s**	**16s**	**Rag1**	**Rag2**	**A**	**A5**	**L**	**M**	**U**	**Y**
Alosa	*pseudoharengus*	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Alosa	*chrysochloris*	Y	Y	Y	Y	Y	-	Y	Y	Y	Y
Clupea	*pallasii*	Y	Y	Y	Y	Y	-	-	Y	Y	-
Engraulis	*eurystole*	Y	Y	Y	Y	Y	Y	-	-	Y	-
Harengula	*jaguana*	Y	Y	Y	Y	Y	Y	Y	-	Y	Y
Coilia	*nasus*	Y	Y	Y	Y	Y	Y	-	-	Y	-
Sardina	*pilchardus*	Y	Y	N	Y	Y	Y	Y	Y	-	Y
Coilia	*brachygnathus*	Y	Y	Y	Y	Y	-	-	Y	-	-
Ilisha	*elongata*	Y	Y	N	Y	Y	-	-	-	-	Y
Setipinna	*taty*	Y	Y	Y	Y	Y	Y	-	-	-	-
Chirocentrus	*dorab*	AP006229	AP006229	Y	Y	Y	Y	Y	Y	Y	Y
Sundasalanx	*mekongensis*	AP006232	AP006232	N	N	Y	Y	Y	-	Y	Y
Sprattus	*sprattus*	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Sardinella	*maderensis*	Y	Y	Y	-	-	-	-	Y	-	Y
Ethmalosa	*fimbriata*	Y	Y	Y	-	Y	Y	Y	Y	Y	Y
Spratelloides	*delicatulus*	Y	Y	Y	Y	Y	Y	Y	-	-	Y
Spratelloides	*gracilis*	Y	Y	Y	Y	Y	-	Y	-	-	Y
Dorosoma	*petenense*	Y	Y	Y	-	Y	-	-	Y	-	Y
Anodontostoma	*chacunda*	Y	Y	Y	-	Y	Y	Y	Y	-	Y
Nematalosa	*japonica*	Y	Y	Y	-	Y	Y	Y	-	Y	Y

Table 5.1. Clupeiforms and outgroups sequenced for 10 loci. (cont.).

Genus	**Species**	**12s**	**16s**	**Rag1**	**Rag2**	**A**	**A5**	**L**	**M**	**U**	**Y**
Pellonula	*leonensis*	Y	Y	Y	Y	Y	Y	-	Y	-	Y
Odaxothrissa	*vittata*	Y	Y	Y	Y	-	Y	Y	-	-	Y
Ilisha	*africana*	Y	Y	-	-	Y	Y	-	-	-	-
Clupeonella	*cultriventris*	Y	Y	Y	Y	Y	-	-	-	-	-

Y, indicates the locus has been sequenced in the present study; - indicates the locus cannot be amplified; accession number indicates the sequence was retrieved from GenBank.

Table 5.2. Characteristics of the ten loci amplified in clupeiforms.

Genes*	No. of bp	No. of var. sites	No. of PI sites	Average p-distance	Relative rate	CI-MP	No. of species sequenced
12S	600	312	244	0.167	1.456	0.307	44
16S	1388	821	704	0.192	1.627	0.330	44
RAG1	1545	781	644	0.162	1.088	0.389	39
RAG2	1269	773	619	0.200	1.082	0.438	35
zic1	891	318	261	0.104	0.415	0.418	40
RYR3	825	414	349	0.171	0.877	0.400	31
ENC1	801	373	234	0.136	0.665	0.519	22
Gylt	864	456	357	0.177	1.104	0.478	24
plagl2	792	313	212	0.137	0.661	0.571	20
sreb2	987	371	275	0.101	0.681	0.431	33

Table 5.3. GC content (G + C, %) of the ten loci in different taxanomic groups.

Taxa	12S16S*	RAGs*	zic1	RYR3	ENC1	Gylt	plagl2	Sreb2
Hidon	0.46	0.61	0.50	0.51	0.60	-	0.61	0.58
ostariophysans	0.46	0.51	0.56	0.48	0.54	0.48	0.56	0.58
clupeiforms	0.49	0.63	0.55	0.51	0.66	0.53	0.66	0.62
Denticeps	0.42	0.55	-	0.45	-	0.59	-	0.61
Jenkinsia	0.52	0.59	0.53	0.49	-	0.51	-	0.63
S. delicatulus	0.53	0.55	0.55	0.48	0.61	-	-	0.57
S. gracilis	0.51	0.66	0.55	-	0.67	-	-	0.63

*12S and 16S were combined together for analyses, because they have similar properties. RAG1 and RAG2 also were combined, see Chapter four.

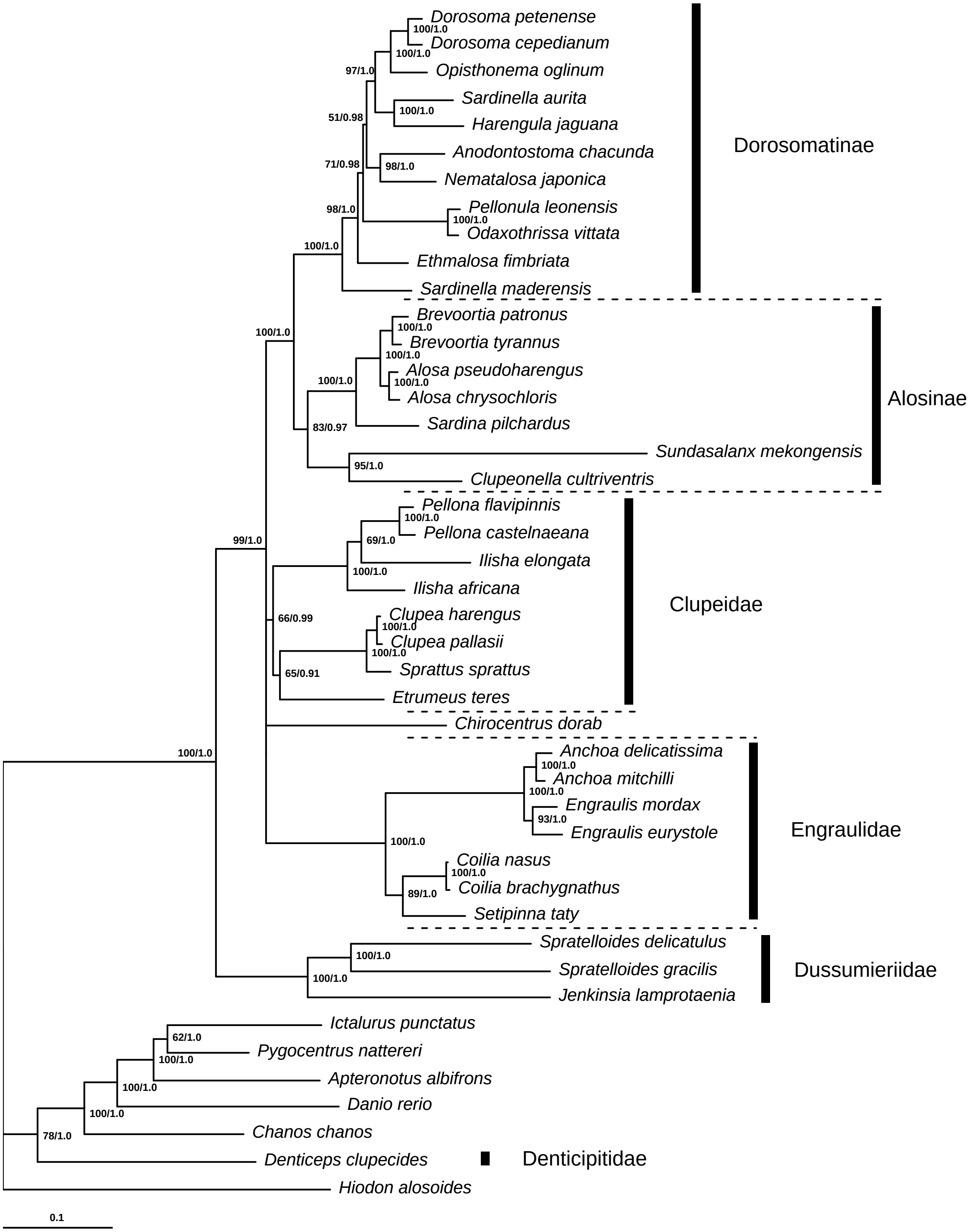

Fig. 5.1. ML phylogram of clupeiforms based on ten loci. The number on braches are ML bootstrap values and Bayesian posterior probabilities.

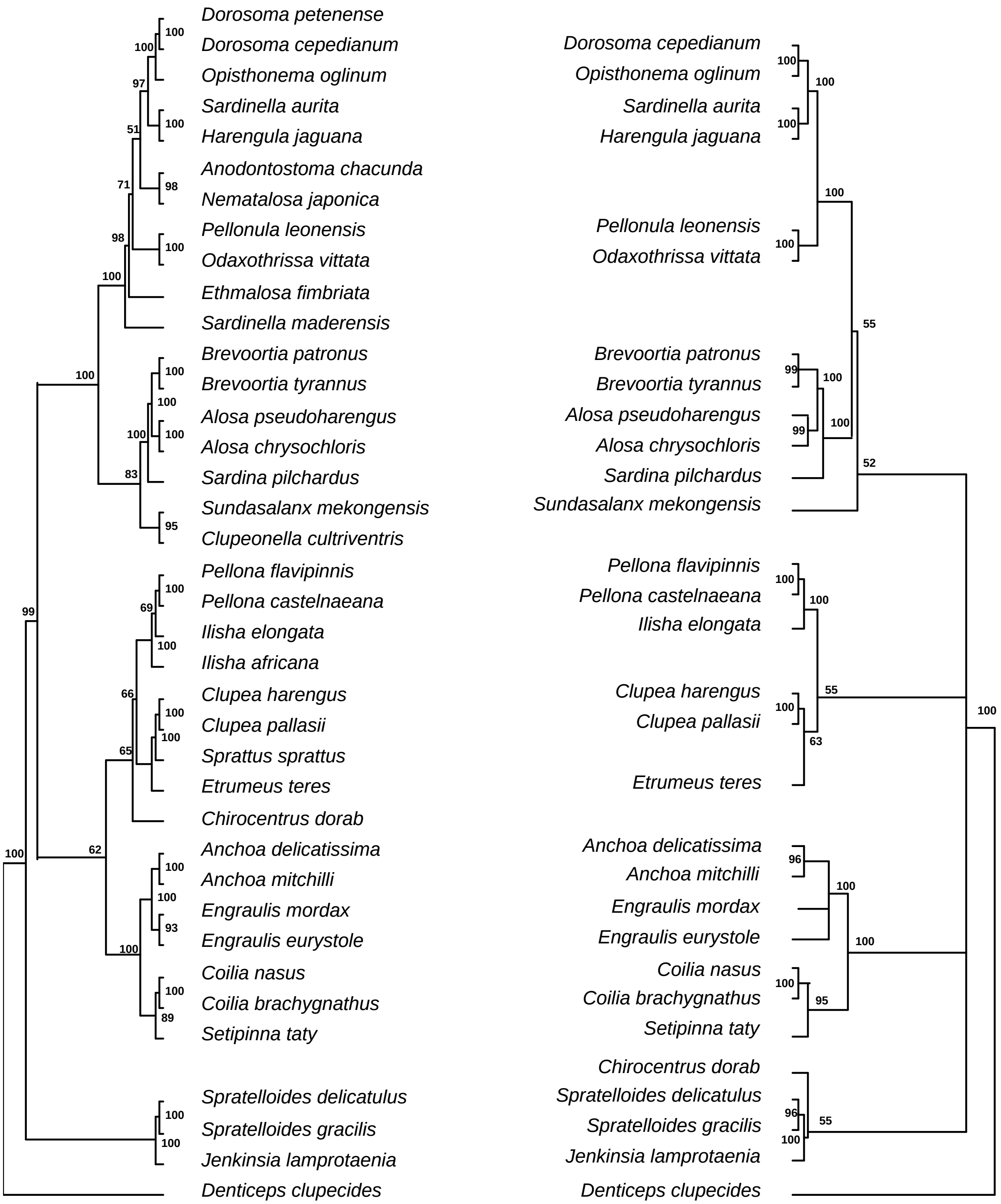

Fig. 5.2. Comparison between the ML phylogeny of clupeiforms inferred from ten loci (left side) and the phylogeny based on four loci (right side) as in Chapter four. Numbers on branches are bootstrap values.

References

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

Amores, A., Force, A., Yan, Y.L., Joly, L., Amemiya, C., Fritz, A., Ho, R.K., Langeland, J., Prince, V., Wang, Y.L., Westerfield, M., Ekker, M., Postlethwait, J.H., 1998. Zebrafish hox clusters and vertebrate genome evolution. Science 282, 1711-1714.

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

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.

Bapteste, E., Brinkmann, H., Lee, J.A., Moore, D.V., Sensen, C.W., Gordon, P., Durufle, L., Gaasterland, T., Lopez, P., Muller, M., Philippe, H., 2002. The analysis of 100 genes supports the grouping of three highly divergent amoebae: Dictyostelium, Entamoeba, and Mastigamoeba. Proc. Natl. Acad. Sci. USA 99, 1414-1419.

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

Blanquart, S., Lartillot, N., 2006. A Bayesian Compound Stochastic Process for Modeling Nonstationary and Nonhomogeneous Sequence Evolution. Mol. Biol. Evol. 23, 2058-2071.

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

Britz, R., Kottelat, M., 1999. Sundasalanx mekongensis, a new species of clupeiform fish from the Mekong basin (Teleostei: Sundasalangidae). Ichthyol. Explor. Freshwaters 10, 337-344.

Bryant, D., Galtier, N., Poursat, M., 2005. Likelihood calculation in molecular phylogenetics. In: Gascuel, O. (Ed.), Mathematics of evolution and phylogeny. Oxford University Press, New York, pp. 33-62.

Buckley, T.R., Simon, C., Chambers, G.K., 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., Huelsenbeck, J.P., Cunningham, C.W., Swofford, D.L., Waddell, P.J., 1993. Partitioning and combining data in phylogenetic analysis. Syst. Biol. 42, 384-397.

Burrow, C.J., Turner, S., 2000. Silurian vertebrates from Australia. In: Blieck, A., Turner, S. (Eds.), IGCP 328 Final Report. Courier Forschungsinstitut Senckenberg 223, pp. 169-174.

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

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

Chen, W.J., Bonillo, C., Lecointre, G., 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., Ortí, G., Meyer, A., 2004. Novel evolutionary relationship among four fish model systems. Trends Genet. 20, 424-431.

Churchill, G.A., von Haeseler, A., Navidi, W.C., 1992. Sample size for a phylogenetic inference. Mol. Biol. Evol. 9, 753-769.

Ciccarelli, F.D., von Mering, C., Suyama, M., Harrington, E.D., Izaurralde, E., Bork, P., 2005. Complex genomic rearrangements lead to novel primate gene function. Genome Res. 15, 343-351.

Clausen, H.S., 1959. Denticipitidae, a new family of primitive isospondylous teleosts from West African fresh-water. Vidensk. Medd. Dansk. Naturhist. Foren. 121, 141-151.

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

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

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

Crollius, H.R., Weissenbach, J., 2005. Fish genomics and biology. Genome Res. 15, 1675-1682.

Crow, K.D., Wagner, G.P., 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.

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

de Pinna, M.C.C., 1996. Teleostean monophyly. In: Stiassny, M.L.J., Parenti, L.R., Johnson, G.D. (Eds.), Interrelationships of fishes. Academic Press, San Diego, pp. 193-207.

Delsuc, F., Brinkmann, H., Philippe, H., 2005. Phylogenomics and the reconstruction of the tree of life. Nat. Rev. Genet. 6, 361-375.

Di Dario, F., 2002. Evidence supporting a sister-group relationship between Clupeoidea and Engrauloidea (Clupeomorpha). Copeia 102, 496-503.

Di Dario, F., 2004. Homology between the recessus lateralis and cephalic sensory canal, with the proposition of additional synapomorphies for the Clupeiformes and the Clupeoidei. Zool. J. Linnean Soc. 141, 257-270.

Di Dario, F., 2005. Relações filogenéticas entre os grandes grupos de Clupeomorpha e suas possíveis relações com Ostariophysi (Actinopterygii, Teleostei). Universidade de São Paulo, p. 629 pp.

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

Driskell, A.C., Ane, C., Burleigh, J.G., McMahon, M.M., O'Meara B, C., Sanderson, M.J., 2004. Prospects for building the tree of life from large sequence databases. Science 306, 1172-1174.

Drummond, A.J., Ho, S.Y., Phillips, M.J., Rambaut, A., 2006. Relaxed phylogenetics and dating with confidence. PLoS Biol. 4, e88.

Eisen, J.A., Fraser, C.M., 2003. Phylogenomics: intersection of evolution and genomics. Science 300, 1706-1707.

Farris, J.S., 1983. The logical basis of phylogenetic analysis. In: Platnick, N.I., Funk, V.A. (Eds.), Advances in cladistics, volume 2 : proceedings of the second meeting of the Willi Hennig Society. Columbia University Press, New York, pp. 7-36.

Farris, J.S., Kallersjo, M., Kluge, A.G., Bult, C., 1995a. Constructing a significance test for incongruence. Syst. Biol. 44, 570-572.

Farris, J.S., Kallersjo, M., Kluge, A.G., Bult, C., 1995b. Testing significance of incongruence. Cladistics 10, 315-319.

Felsenstein, J., 1978. Case in which parsimony or compatibility methods will be positively misleading. Syst. Biol. 27, 401-410.

Felsenstein, J., 2004. Inferring phylogenies. Sinauer Associates, Sunderland, Mass.

Felsenstein, J., 2005. PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author., Seattle.

Felsenstein, J., Sober, E., 1986. Parsimony and likelihood: an exchange. Syst. Zool. 35, 617-626.

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

Fink, S.V., Fink, W.L., 1996. Interrelationships of ostariophysan fishes (Teleostei). In: Stiassny, M.L.J., Parenti, L.R., Johnson, G.D. (Eds.), Interrelationships of fishes. Academic Press, San Diego, pp. 209-249.

Fink, W.L., 1984a. Basal euteleosts: relationships. In: Moser, H.G., Richards, W.J., Cohen, D.M., Fahay, M.P., Kendall, A.W., Richardson, S.L. (Eds.), Ontogeny and systematics of fishes. American Society of Ichthyologists and Herpetologists, Special Publication No. 1. Allen Press, Lawrence, Kansas, pp. 202-206.

Fink, W.L., 1984b. Salmoniforms: introduction. In: Moser, H.G., Richards, W.J., Cohen, D.M., Fahay, M.P., Kendall, A.W., Richardson, S.L. (Eds.), Ontogeny and systematics of fishes. American Society of Ichthyologists and Herpetologists, Special Publication No. 1. Allen Press, Lawrence, Kansas, pp. 1-139.

Fitch, W.M., 1970. Distinguishing homologous from analogous proteins. Syst. Zool. 19, 99-113.

Fitch, W.M., 1971. Rate of change of concomitantly variable codons. J. Mol. Evol. 1, 84-96.

Foster, P.G., 2004. Modeling compositional heterogeneity. Syst. Biol. 53, 485-495.

Foster, P.G., Hickey, D.A., 1999. Compositional bias may affect both DNA-based and protein-based phylogenetic reconstructions. J. Mol. Evol. 48, 284-290.

Friedlander, T.P., Regier, J.C., Mitter, C., 1992. Nuclear gene sequences for higher level phylogenetic analysis: 14 promising candidates. Syst. Biol. 41, 483-490.

Froese, R., Pauly, D., 2004. FishBase. World Wide Web electronic publication. www.fishbase.org, version (11/2005).

Funk, D.J., Omland, K.E., 2003. SPECIES-LEVEL PARAPHYLY AND POLYPHYLY: Frequency, Causes, and Consequences, with Insights from Animal Mitochondrial DNA. Annu. Rev. Ecol. Evol. Syst. 34, 397-423.

Galtier, N., 2001. Maximum-likelihood phylogenetic analysis under a covarion-like model. Mol. Biol. Evol. 18, 866-873.

Galtier, N., Gouy, M., 1998. Inferring pattern and process: maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis. Mol. Biol. Evol. 15, 871-879.

Galtier, N., Jean-Marie, A., 2004. Markov-modulated Markov chains and the covarion process of molecular evolution. J. Comput. Biol. 11, 727-733.

Gardiner, B.G., 1993. Basal actinopterygians. In: Benton, M.J. (Ed.), The fossil record. Chapman & Hall, London, pp. 148-152.

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

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

Gillespie, J.H., 1994. The causes of molecular evolution. Oxford University Press, USA.

Goldman, N., 1990. Maximum likelihood inference of phylogenetic trees, with special reference to a poisson process model of DNA substitution and to parsimony analyses. Syst. Zool. 39, 345-361.

Gowri-Shankar, V., Rattray, M., 2006. On the correlation between composition and site-specific evolutionary rate: implications for phylogenetic inference. Mol. Biol. Evol. 23, 352-364.

Grande, L., 1982. A revision of the fossil genus Knightia, with a description of a new genus from the Green River Formmation (Teleostei, Clupeidae). Am. Mus. Novitates 2731, 1-22.

Grande, L., 1985. Recent and fossil Clupeomorph fishes with materials for revision of the subgroups of clupeoids. Bull. Am. Mus. Nat. Hist. 181, 231-372.

Grande, L., Bemis, W.E., 1996. Interrelationships of Acipenseriformes, with comments on "Chondrostei". In: Stiassny, M.L.J., Parenti, L.R., Johnson, G.D. (Eds.), Interrelationships of fishes. Academic Press, San Diego, pp. 85-115.

Greenwood, P.H., 1968. The osteology and relationships of the Denticipitidae a family of Clupemorph fishes denticeps-clupeoides palaeodenticeps-tanganikae. Bull. Br. Mus. (Nat.-Hist.) Zool. 16, 15-273.

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

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

Groth, J.G., Barrowclough, G.F., 1999. Basal divergences in birds and the phylogenetic utility of the nuclear RAG-1 gene. Mol. Phylogenet. Evol. 12, 115-123.

Guindon, S., Gascuel, O., 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. 52, 696-704.

Hansen, J.D., Kaattari, S.L., 1996. The recombination activating gene 2 (RAG2) of the rainbow trout *Oncorhynchus mykiss*. Immunogenetics 44, 203-211.

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

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

Hillis, D.M., Pollock, D.D., McGuire, J.A., Zwickl, D.J., 2003. Is sparse taxon sampling a problem for phylogenetic inference? Syst. Biol. 52, 124-126.

Huber, K.T., Langton, M., Penny, D., Moulton, V., Hendy, M., 2002. Spectronet: a package for computing spectra and median networks. Appl. Bioinform. 1, 159-161.

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

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

Ishiguro, N.B., Miya, M., Inoue, J.G., Nishida, M., 2005. Sundasalanx (Sundasalangidae) is a progenetic clupeiform, not a closely-related group of salangids (Osmeriformes): mitogenomic evidence. J. Fish Biol. 67, 561-569.

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

Jaillon, O., Aury, J.M., Brunet, F., Petit, J.L., Stange-Thomann, N., Mauceli, E., Bouneau, L., Fischer, C., Ozouf-Costaz, C., Bernot, A., Nicaud, S., Jaffe, D., Fisher, S., Lutfalla, G., Dossat, C., Segurens, B., Dasilva, C., Salanoubat, M., Levy, M., Boudet, N., Castellano, S., Anthouard, V., Jubin, C., Castelli, V., Katinka, M., Vacherie, B., Biemont, C., Skalli, Z., Cattolico, L., Poulain, J., De Berardinis, V., Cruaud, C., Duprat, S., Brottier, P., Coutanceau, J.P., Gouzy, J., Parra, G., Lardier, G., Chapple, C., McKernan, K.J., McEwan, P., Bosak, S., Kellis, M., Volff, J.N., Guigo, R., Zody, M.C., Mesirov, J., Lindblad-Toh, K., Birren, B., Nusbaum, C., Kahn, D., Robinson-Rechavi, M., Laudet, V., Schachter, V., Quetier, F., Saurin, W., Scarpelli, C., Wincker, P., Lander, E.S., Weissenbach, J., Roest Crollius, H., 2004. Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. Nature 431, 946-957.

Jenner, R.A., 2004. Accepting partnership by submission? Morphological phylogenetics in a molecular millennium. Syst. Biol. 53, 333-342.

Jessen, H., 1972. Schultergürtel und Pectoralflosse bei Actinopterygiern. Fossils and Strata 1, 1–101.

Jobb, G., 2006. TREEFINDER. Distributed by the author at www.treefinder.de, Munich, Germany.

Jobb, G., von Haeseler, A., Strimmer, K., 2004. TREEFINDER: a powerful graphical analysis environment for molecular phylogenetics. BMC Evol Biol 4, 18.

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

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

Johnson, G.D., Patterson, C., 1996. Relationships of lower euteleostean fishes. In: Stiassny, M.L.J., Parenti, L.R., Johnson, G.D. (Eds.), Interrelationships of fishes. Academic Press, San Diego, pp. 251-332.

Jones, D.T., Taylor, W.R., Thornton, J.M., 1992. The rapid generation of mutation data matrices from protein sequences. Comput. Appl. Biosci. 8, 275-282.

Kelchner, S.A., Thomas, M.A., 2006. Model use in phylogenetics: nine key questions. Trends Ecol. Evol. 22, 87-94.

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

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

Kolaczkowski, B., Thornton, J.W., 2004. Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. Nature 431, 980-984.

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

Kumar, S., Tamura, K., Nei, M., 2004. MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. Brief Bioinform. 5, 150-163.

Kurland, C.G., Canback, B., Berg, O.G., 2003. Horizontal gene transfer: a critical view. Proc. Natl. Acad. Sci. USA 100, 9658-9662.

Lartillot, N., Philippe, H., 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol. Biol. Evol. 21, 1095-1109.

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

Lê, H.L., Lecointre, G., Perasso, R., 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.

Lecointre, G., Nelson, G., 1996. Clupeomorpha, sister group of Ostrariophysi. In: Stiassny, M.L.J., Parenti, L.R., Johnson, G.D. (Eds.), Interrelationships of fishes. Academic Press, San Diego, pp. 193-207.

Lewis, P.O., 2001. A likelihood approach to estimating phylogeny from discrete morphological character data. Syst. Biol. 50, 913-925.

Lockhart, P.J., Steel, M.A., Penny, D., Hendy, M.D., 1994. Recovering evolutionary trees under a more realistic model of sequence evolution. Mol. Biol. Evol. 11, 605-612.

Lopez, A.J., Chen, W.J., Ortí, G., 2004. Esociform phylogeny. Copeia 2004, 449-464.

Lopez, P., Casane, D., Philippe, H., 2002. Heterotachy, an important process of protein evolution. Mol. Biol. Evol. 19, 1-7.

Lovejoy, N.R., Collette, B.B., 2001. Phylogenetic relaionships of new world needlefishes (Teleostei: Belonidae) and the biogeography of transitions between marine and freshwater habitats. Copeia 2001, 324-338.

Lyons-Weiler, J., Hoelzer, G.A., Tausch, R.J., 1996. Relative apparent synapomorphy analysis (RASA). I: The statistical measurement of phylogenetic signal. Mol. Biol. Evol. 13, 749-757.

Maddison, W.P., 1997. Gene trees in species trees. Syst. Biol. 46, 523-536.

Maddison, W.P., Knowles, L.L., 2006. Inferring phylogeny despite incomplete lineage sorting. Syst. Biol. 55, 21-30.

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

Meyer, A., Van de Peer, Y., 2005. From 2R to 3R: evidence for a fish-specific genome duplication (FSGD). Bioessays 27, 937-945.

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

Miya, M., Kawaguchi, A., Nishida, M., 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., Nishida, M., 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., Satoh, T.P., Nishida, M., 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. Linnean Sco. Lond. 85, 289-306.

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

Mohammad-Ali, K., Eladari, M.E., Galibert, F., 1995. Gorilla and orangutan c-myc nucleotide sequences: inference on hominoid phylogeny. J. Mol. Evol. 41, 262-276.

Müller, K., 2005. SeqState: primer design and sequence statistics for phylogenetic DNA datasets. Appl. Bioinform. 4, 65-69.

Müller, K., 2006. Incorporating information from length-mutational events into phylogenetic analysis. Mol. Phylogenet. Evol. 38, 667-676.

Murphy, W.J., Eizirik, E., Johnson, W.E., Zhang, Y.P., Ryder, O.A., O'Brien, S.J., 2001. Molecular phylogenetics and the origins of placental mammals. Nature 409, 614-618.

Nelson, G.J., 1967. Gill arches of teleostean fishes of the family Clupeidae. Copeia, 389-399.

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, G.J., 1970. The hyobranchial apparatus of teleostean fishes of the families Engrauidae and Chirocentridae. Am. Mus. Novitates 2410, 1-30.

Nelson, J.S., 1976. Fishes of the world. Wiley, New York.

Nelson, J.S., 1984. Fishes of the world. Wiley, New York.

Nelson, J.S., 1994. Fishes of the world. J. Wiley, New York.

Nelson, J.S., 2006. Fishes of the world. John Wiley and Sons, Inc., New York.

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

Oettinger, M.A., 1992. Activation of V(D)J recombination by RAG1 and RAG2. Trends Genet. 8, 413-416.

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. Paleonto. 4, 481-499.

Ortí, G., Meyer, A., 1996. Molecular Evolution of Ependymin and the Phylogenetic Resolution of Early Divergences Among Euteleost Fishes. Mol. Biol. Evol. 13, 556-573.

Ortí, G., Meyer, A., 1997. The radiation of characiform fishes and the limits of resolution of mitochondrial ribosomal DNA sequences. Syst. Biol. 46, 75-100.

Page, R.D., 1996. TreeView: an application to display phylogenetic trees on personal computers. Comput. Appl. Biosci. 12, 357-358.

Page, R.D., Cotton, J.A., 2002. Vertebrate phylogenomics: reconciled trees and gene duplications. Pac. Symp. Biocomput., 536-547.

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

Pagel, M., Meade, A., 2005. Mixture models in phylogenetic inference. In: Gascuel, O. (Ed.), Mathematics of evolution and phylogeny. Oxford University Press, New York, pp. 63-90.

Pamilo, P., Nei, M., 1988. Relationships Between Gene Trees and Species Trees. Mol. Biol. Evol. 5, 568-583.

Patterson, C., 1973. Interrelationships of holosteans. In: Greenwood, P.H., Miles, R.S., Patterson, C. (Eds.), Interrelationships of fishes. Academic Press, London, pp. 207-226.

Patterson, C., Johnson, G.D., 1995. The intermuscular bones and ligaments of teleostean fishes. Smiths. Contrib. Zool. 559, 1-85.

Patterson, C., Rosen, D.E., 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., Rosen, D.E., 1989. The Paracanthopterygii revisited: order and disorder. In: Cohen, D.M. (Ed.), Papers on the systematics of gadiform fishes. . Natural History Museum of Los Angeles County, Los Angeles, California, pp. 5-36.

Peixoto, B.R., Mikawa, Y., Brenner, S., 2000. Characterization of the recombinase activating gene-1 and 2 locus in the Japanese pufferfish, *Fugu rubripes*. Gene 246, 275-283.

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

Philippe, H., Snell, E.A., Bapteste, E., Lopez, P., Holland, P.W., Casane, D., 2004. Phylogenomics of eukaryotes: impact of missing data on large alignments. Mol. Biol. Evol. 21, 1740-1752.

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

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

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

Poe, S., Swofford, D.L., 1999. Taxon sampling revisited. Nature 398, 299-300.

Posada, D., Buckley, T.R., 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., Crandall, K.A., 1998. MODELTEST: testing the model of DNA substitution. Bioinformatics 14, 817-818.

Poux, C., Madsen, O., Marquard, E., Vieites, D.R., de Jong, W.W., Vences, M., 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., Huchon, D., Cao, Y., Okada, N., Hasegawa, M., 2002. Combining multiple data sets in a likelihood analysis: which models are the best? Mol. Biol. Evol. 19, 2294-2307.

Pybus, O.G., 2006. Model selection and the molecular clock. PLoS Biol. 4, e151.

Reed, R.D., Sperling, F.A., 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., 445-461.

Roberts, T.R., 1981. Sundasalangidae, a new family of minute freshwater salmoniform fishes from southeast Asia. Proc. Calif. Acad. Sci. 42, 295-302.

Roberts, T.R., 1984. Skeletal anatomy and classification of the neotenic Asian salmoniform superfamily Salangoidea (icefishes or noodlefishes). Proc. Calif. Acad. Sci. 43, 179-220.

Rokas, A., Carroll, S.B., 2006. Bushes in the tree of life. PLoS Biol. 4, e352.

Rokas, A., Holland, P.W., 2000. Rare genomic changes as a tool for phylogenetics. Trends Ecol. Evol. 15, 454-459.

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

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

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

Ronquist, F., Huelsenbeck, J.P., 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19, 1572-1574.

Rosen, D.E., Greenwood, P.H., 1970. Origin of the Weberian apparatus and the relationships of the ostariophysan and gonorynchiform fishes. Am. Mus. Novitates 2428, 1-25.

Rozen, S., Skaletsky, H., 2000. Primer3 on the WWW for general users and for biologist programmers. Methods Mol. Biol. 132, 365-386.

Saint, K.M., Austin, C.C., Donnellan, S.C., Hutchinson, M.N., 1998. C-mos, a nuclear marker useful for squamate phylogenetic analysis. Mol. Phylogenet. Evol. 10, 259-263.

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

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

Sanderson, M.J., 1997. A nonparametric approach to estimating divergence times in the absence of rate constancy. Mol. Biol. Evol. 14, 1218-1231.

Sanderson, M.J., Kim, J., 2000. Parametric phylogenetics? Syst. Biol. 49, 817-829.

Schaeffer, B., 1973. Interrelationships of chondrosteans. In: Greenwood, P.H., Miles, R.S., Patterson, C. (Eds.), Interrelationships of fishes. Academic Press, London, pp. 207-226.

Shen, M., 1996. Fossil "osteoglossomorphs" in East Asia and their implications in teleostean phylogeny. In: Arratia, G., Viohl, G. (Eds.), MESOZOIC FISHES: SYSTEMATICS AND PALEOECOLOGY. Verlag Dr. F. Pfeil, München, Germany, pp. 261-272.

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

Siebert, D.J., 1997. Notes on the anatomy and relationships of Sundasalanx Roberts (Teleostei, Clupeidae), with descriptions of four new species from Borneo. Bull. Nat. Hist. Mus. Zool. Ser. 63, 13-26.

Simmons, M.P., Ochoterena, H., 2000. Gaps as characters in sequence-based phylogenetic analyses. Syst. Biol. 49, 369-381.

Small, R., Cronn, R., Wendel, J., 2004. L. A. S. Johnson Review No. 2. Use of nuclear genes for phylogeny reconstruction in plants. Aust. Syst. Botany 17, 145-170.

Sober, E., 2004. The contest between parsimony and likelihood. Syst. Biol. 53, 644-653.

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

Springer, V.G., Johnson, G.D., 2004. Study of the Dorsal Gill-arch Musculature of Teleostome Fishes, with Special Reference to the Actinopterygii. Bull. Biol. Soc. Wash. 11, 260.

Steel, M.A., 2005. Should phylogenetic models be trying to "fit an elephant"? Trends Genet. 21, 307-309.

Steel, M.A., Lockhart, P.J., Penny, D., 1993. Confidence in evolutionary trees from biological sequence data. Nature 364, 440-442.

Steel, M.A., Penny, D., 2000. Parsimony, likelihood, and the role of models in molecular phylogenetics. Mol. Biol. Evol. 17, 839-850.

Steinke, D., Salzburger, W., Meyer, A., 2006. Novel Relationships Among Ten Fish Model Species Revealed Based on a Phylogenomic Analysis Using ESTs. J. Mol. Evol. 62, 772-784.

Stiassny, M.L.J., Parenti, L.R., Johnson, G.D. (Eds.), 1996a. Interrelationships of fishes. Academic Press, San Diego.

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

Stiassny, M.L.J., Wiley, E.O., Johnson, G.D., de Carvalho, M.R., 2004. Gnathostome fishes. In: Cracraft, J., Donoghue, M.J. (Eds.), Assembling The Tree of Life. Oxford University Press, New York, pp. 410-429.

Swofford, D.L., 2003. PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4 Sinauer Associates, Sunderland, Massachusetts.

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

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

Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F., Higgins, D.G., 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. 25, 4876-4882.

Thompson, J.D., Higgins, D.G., Gibson, T.J., 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.

Tuffley, C., Steel, M., 1998. Modeling the covarion hypothesis of nucleotide substitution. Math Biosci. 147, 63-91.

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

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

Weisburg, W.G., Giovannoni, S.J., Woese, C.R., 1989. The Deinococcus-Thermus phylum and the effect of rRNA composition on phylogenetic tree construction. Syst. Appl. Microbiol. 11, 128-134.

Weisrock, D.W., Harmon, L.J., Larson, A., 2005. Resolving deep phylogenetic relationships in salamanders: analyses of mitochondrial and nuclear genomic data. Syst. Biol. 54, 758-777.

Whitehead, P.J.P., 1972. A synopsis of the clupeoid fishes of India. J. Mar. Biol. Assn. India 14, 160-256.

Whitehead, P.J.P., 1985. FAO species catalogue. Vol. 7. Clupeoid fishes of the world (suborder Clupeoidei). An annotated and illustrated catalogue of the herrings, pilchards, sprats, shads, anchovies and wold-herrings. Part 1 - Chirocentridae, Clupeidae and pristigasteridae. FAO-Fisheries-Synopsis: 303, No. 125 (7 Part 1).

Whittall, J.B., Medina-Marino, A., Zimmer, E.A., Hodges, S.A., 2006. Generating single-copy nuclear gene data for a recent adaptive radiation. Mol. Phylogenet. Evol. 39, 124-134.

Wiens, J.J., 2004. The role of morphological data in phylogeny reconstruction. Syst. Biol. 53, 653-661.

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

Willett, C.E., Cherry, J.J., Steiner, L.A., 1997. Characterization and expression of the recombination activating genes (rag1 and rag2) of zebrafish. Immunogenetics 45, 394-404.

Williams, R.R.G., 1987. The phylogenetic relationships of the salmoniform fishes based on the suspensorim and its muscles. . Dept. of Zoology. University of Alberta, Edmonton.

Woese, C.R., Achenbach, L., Rouviere, P., Mandelco, L., 1991. Archaeal phylogeny: reexamination of the phylogenetic position of Archaeoglobus fulgidus in light of certain composition-induced artifacts. Syst. Appl. Microbiol. 14, 364-371.

Wongratana, T., 1987. Four new species of clupeoid fishes Clupeidae and Engraulidae from Australian waters. Proc. Biol. Soc. Wash. 100, 104-111.

Woods, I.G., Wilson, C., Friedlander, B., Chang, P., Reyes, D.K., Nix, R., Kelly, P.D., Chu, F., Postlethwait, J.H., Talbot, W.S., 2005. The zebrafish gene map defines ancestral vertebrate chromosomes. Genome Res. 15, 1307-1314.

Wortley, A.H., Scotland, R.W., 2006. The effect of combining molecular and morphological data in published phylogenetic analyses. Syst. Biol. 55, 677-685.

Yang, Z., 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J. Mol. Evol. 39, 306-314.

Yang, Z., 2005. Bayesian inference in molecular phylogenetics. In: Gascuel, O. (Ed.), Mathematics of evolution and phylogeny. Oxford University Press, New York, pp. 63-90.

Yang, Z., Roberts, D., 1995. On the use of nucleic acid sequences to infer early branchings in the tree of life. Mol. Biol. Evol. 12, 451-458.

Zwickl, D.J., 2006. Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion. The University of Texas at Austin, Austin.

Appendix A. Exon ID, length, GC content of predicted single nuclear gene markers in zebrafish and torafugu, as well the blast result between orthologous genes.

	Zebrafish			Torafugu			Torafugu vs Zebra[‡]	
No. of markers	Exon ID	Exon length	GC content	Exon ID	Exon length	GC content	E-value	Identity (%)
1	ENSDARE00000015655[*]	968	0.55	SINFRUE00000662228	970	0.57	0	83
2	ENSDARE00000145053[*]	1664	0.49	SINFRUE00000786790	1227	0.48	3E-122	83.99
3	ENSDARE00000117872[*]	1402	0.49	SINFRUE00000719108	1429	0.57	2E-104	80.06
4	ENSDARE00000136964[*]	2605	0.46	SINFRUE00000561510	1483	0.58	3E-101	84.02
5	ENSDARE00000367269[*]	1482	0.53	SINFRUE00000681690	1770	0.58	1E-121	79.01
6	ENSDARE00000465292[*]	1307	0.48	SINFRUE00000577106	1408	0.53	3E-115	83.13
7	ENSDARE00000025410[*]	5811	0.47	SINFRUE00000644156	5799	0.48	2E-91	79.34
8	ENSDARE00000029022[*]	2894	0.47	SINFRUE00000628754	1116	0.57	0	86.29
9	ENSDARE00000039808[*]	1596	0.49	SINFRUE00000611615	1773	0.57	4E-53	87.26
10	ENSDARE00000055502[*]	1745	0.47	SINFRUE00000673034	844	0.58	2E-68	83.47
11	ENSDARE00000092751†	1636	0.50	SINFRUE00000725450	1636	0.58	5E-49	85.41
12	ENSDARE00000473520†	946	0.55	SINFRUE00000766736	856	0.56	3E-94	78
13	ENSDARE00000023056†	948	0.60	SINFRUE00000575639	969	0.61	0	81
14	ENSDARE00000053911[†]	534	0.54	SINFRUE00000649188	543	0.56	4E-165	85
15	ENSDARE00000281285[†]	640	0.55	SINFRUE00000774212	703	0.57	4E-133	81
16	ENSDARE00000008379	886	0.48	SINFRUE00000641978	1920	0.53	1E-22	82.32
17	ENSDARE00000014605	927	0.54	SINFRUE00000776709	1041	0.61	3E-68	88.16
18	ENSDARE00000021371	2073	0.45	SINFRUE00000577163	1086	0.68	3E-34	83.65
19	ENSDARE00000025341	971	0.56	SINFRUE00000609687	1935	0.62	5E-28	85.21
20	ENSDARE00000038832	1100	0.47	SINFRUE00000735032	1056	0.54	9E-93	80.7

Appendix A. (cont.).

	Zebrafish			Torafugu			Torafugu vs Zebra‡	
No. of markers	Exon ID	Exon length	GC content	Exon ID	Exon length	GC content	E-value	Identity (%)
21	ENSDARE00000039062	1188	0.53	SINFRUE00000776320	1203	0.60	2E-38	81.61
22	ENSDARE00000050276	1216	0.48	SINFRUE00000789399	1089	0.55	6E-23	84.4
23	ENSDARE00000051716	1867	0.46	SINFRUE00000690882	953	0.55	9E-28	80.49
24	ENSDARE00000057069	1394	0.48	SINFRUE00000732606	1638	0.50	2E-63	80.68
25	ENSDARE00000060643	1247	0.49	SINFRUE00000723019	1319	0.49	6E-36	81.36
26	ENSDARE00000072303	883	0.50	SINFRUE00000618086	844	0.53	2E-53	87.04
27	ENSDARE00000072794	1940	0.45	SINFRUE00000634400	1263	0.51	2E-60	84.67
28	ENSDARE00000075160	1203	0.49	SINFRUE00000722545	1206	0.54	2E-26	85.93
29	ENSDARE00000075532	1002	0.52	SINFRUE00000733528	996	0.59	3E-120	83.43
30	ENSDARE00000080271	1654	0.46	SINFRUE00000768063	826	0.58	3E-27	82.5
31	ENSDARE00000083490	846	0.54	SINFRUE00000688050	894	0.53	5E-26	83.53
32	ENSDARE00000094312	915	0.56	SINFRUE00000626899	948	0.62	4E-30	79.18
33	ENSDARE00000101104	2013	0.54	SINFRUE00000703687	2533	0.57	2E-25	83.33
34	ENSDARE00000105670	2202	0.43	SINFRUE00000588080	914	0.62	5E-60	84.24
35	ENSDARE00000108088	1319	0.49	SINFRUE00000564119	925	0.50	2E-35	80.75
36	ENSDARE00000111350	948	0.51	SINFRUE00000800129	1022	0.55	1E-30	83.01
37	ENSDARE00000113193	1994	0.47	SINFRUE00000706470	2496	0.45	4E-29	80.88
38	ENSDARE00000113527	1290	0.47	SINFRUE00000607191	2304	0.50	4E-32	82.81
39	ENSDARE00000114437	2899	0.49	SINFRUE00000634453	2307	0.57	1E-25	81.78
40	ENSDARE00000118208	1545	0.49	SINFRUE00000673962	2670	0.48	7E-34	79.89
41	ENSDARE00000121572	1446	0.53	SINFRUE00000575529	1126	0.52	5E-33	86.99

Appendix A. (cont.).

	Zebrafish			Torafugu			Torafugu vs Zebra[‡]	
No. of		Exon	GC		Exon	GC		
markers	Exon ID	length	content	Exon ID	length	content	E-value	Identity (%)
42	ENSDARE00000121853	863	0.57	SINFRUE00000618756	857	0.62	2E-102	86.45
43	ENSDARE00000127244	1204	0.51	SINFRUE00000699178	1095	0.54	1E-36	84.69
44	ENSDARE00000135137	1995	0.51	SINFRUE00000646724	2007	0.63	9E-36	79.52
45	ENSDARE00000140117	888	0.55	SINFRUE00000650663	816	0.56	5E-23	83.03
46	ENSDARE00000146317	825	0.50	SINFRUE00000623975	1968	0.54	5E-28	81.1
47	ENSDARE00000149196	1678	0.54	SINFRUE00000648016	924	0.70	2E-22	83.55
48	ENSDARE00000156722	1054	0.55	SINFRUE00000642853	1655	0.56	3E-32	87.59
49	ENSDARE00000156742	1647	0.60	SINFRUE00000582617	1570	0.63	4E-28	85.62
50	ENSDARE00000158301	982	0.54	SINFRUE00000581861	964	0.58	1E-30	81.93
51	ENSDARE00000158601	1459	0.51	SINFRUE00000663337	904	0.52	7E-44	81.88
52	ENSDARE00000160152	819	0.55	SINFRUE00000673736	825	0.60	8E-28	81.03
53	ENSDARE00000164315	840	0.48	SINFRUE00000699740	840	0.51	1E-29	83.51
54	ENSDARE00000172488	3750	0.43	SINFRUE00000662708	3288	0.57	8E-25	81.69
55	ENSDARE00000180133	1101	0.52	SINFRUE00000723234	1143	0.53	1E-21	83.23
56	ENSDARE00000180576	2040	0.40	SINFRUE00000668186	929	0.57	4E-27	84.38
57	ENSDARE00000182877	2180	0.46	SINFRUE00000652910	1174	0.58	1E-21	84.44
58	ENSDARE00000189313	891	0.57	SINFRUE00000680694	918	0.55	3E-46	84.75
59	ENSDARE00000189500	1407	0.43	SINFRUE00000684238	2022	0.46	5E-28	85.62
60	ENSDARE00000197458	2251	0.51	SINFRUE00000684419	1555	0.54	1E-24	85.94
61	ENSDARE00000204844	1147	0.51	SINFRUE00000680436	985	0.53	5E-45	82.61
62	ENSDARE00000206420	1075	0.50	SINFRUE00000572111	1123	0.58	8E-41	85.02

Appendix A. (cont.).

	Zebrafish			Torafugu			Torafugu vs Zebra[‡]	
No. of markers	Exon ID	Exon length	GC content	Exon ID	Exon length	GC content	E-value	Identity (%)
63	ENSDARE00000206479	1196	0.53	SINFRUE00000580687	1214	0.58	8E-29	85.16
64	ENSDARE00000219160	1085	0.50	SINFRUE00000607190	1935	0.54	5E-37	82.26
65	ENSDARE00000219263	1742	0.42	SINFRUE00000666050	1064	0.56	4E-24	82.2
66	ENSDARE00000229740	1349	0.57	SINFRUE00000690755	1406	0.51	1E-37	80.98
67	ENSDARE00000254677	2832	0.45	SINFRUE00000575897	1491	0.58	3E-47	82.37
68	ENSDARE00000264881	954	0.53	SINFRUE00000812202	951	0.55	2E-53	82.3
69	ENSDARE00000272936	992	0.49	SINFRUE00000699845	1026	0.49	2E-35	81.13
70	ENSDARE00000281441	2586	0.52	SINFRUE00000610710	2502	0.55	4E-26	86.61
71	ENSDARE00000281522	802	0.56	SINFRUE00000694569	836	0.56	8E-25	81.73
72	ENSDARE00000282174	1036	0.52	SINFRUE00000685586	1087	0.64	4E-30	83.78
73	ENSDARE00000282665	1555	0.47	SINFRUE00000650606	812	0.55	1E-48	83.7
74	ENSDARE00000285110	1232	0.51	SINFRUE00000627739	1290	0.53	7E-60	83.01
75	ENSDARE00000285860	2245	0.48	SINFRUE00000623301	882	0.56	2E-34	85.98
76	ENSDARE00000293219	3252	0.47	SINFRUE00000749920	1509	0.59	1E-27	84.47
77	ENSDARE00000306073	1548	0.54	SINFRUE00000745372	1551	0.55	1E-21	83.23
78	ENSDARE00000308452	891	0.55	SINFRUE00000635758	891	0.53	4E-30	80.78
79	ENSDARE00000311138	1419	0.49	SINFRUE00000599257	1314	0.51	7E-51	80
80	ENSDARE00000311461	1489	0.55	SINFRUE00000610969	964	0.58	4E-30	83.78
81	ENSDARE00000323279	1033	0.53	SINFRUE00000601349	1051	0.49	4E-55	81.32
82	ENSDARE00000332176	1670	0.44	SINFRUE00000602884	1131	0.62	1E-42	80.7
83	ENSDARE00000335381	829	0.52	SINFRUE00000615205	1020	0.61	2E-99	80.46

Appendix A. (cont.).

	Zebrafish			Torafugu			Torafugu vs Zebra[‡]	
No. of		Exon	GC		Exon	GC		
markers	Exon ID	length	content	Exon ID	length	content	E-value	Identity (%)
84	ENSDARE00000342020	936	0.61	SINFRUE00000632131	837	0.61	9E-74	82.74
85	ENSDARE00000344553	854	0.53	SINFRUE00000565494	860	0.58	6E-69	80.89
86	ENSDARE00000347062	843	0.58	SINFRUE00000601793	1386	0.64	1E-43	81.63
87	ENSDARE00000358071	833	0.53	SINFRUE00000591640	857	0.62	2E-115	82.93
88	ENSDARE00000358117	1401	0.49	SINFRUE00000650742	1482	0.51	3E-22	80.51
89	ENSDARE00000359173	1065	0.49	SINFRUE00000608787	1272	0.56	6E-48	78.6
90	ENSDARE00000360719	1062	0.58	SINFRUE00000611346	848	0.63	2E-84	81.5
91	ENSDARE00000360787	1543	0.47	SINFRUE00000667348	992	0.63	1E-33	88.15
92	ENSDARE00000370814	1439	0.53	SINFRUE00000690230	1430	0.61	5E-55	86.12
93	ENSDARE00000377477	2762	0.40	SINFRUE00000802706	1055	0.63	3E-46	80.23
94	ENSDARE00000381363	870	0.60	SINFRUE00000757942	845	0.55	2E-34	91.07
95	ENSDARE00000386979	2706	0.43	SINFRUE00000592475	1072	0.48	1E-27	83.82
96	ENSDARE00000389841	868	0.50	SINFRUE00000695948	862	0.60	3E-43	79.43
97	ENSDARE00000389876	940	0.48	SINFRUE00000695933	1769	0.51	7E-30	85.09
98	ENSDARE00000391626	1110	0.46	SINFRUE00000695204	1089	0.54	1E-21	83.23
99	ENSDARE00000392437	818	0.47	SINFRUE00000619713	818	0.57	3E-27	81.7
100	ENSDARE00000396273	889	0.52	SINFRUE00000656325	1152	0.57	2E-38	82.4
101	ENSDARE00000397971	887	0.53	SINFRUE00000687744	950	0.59	2E-22	86.21
102	ENSDARE00000402487	1593	0.47	SINFRUE00000602810	2076	0.43	1E-22	78.12
103	ENSDARE00000402673	1533	0.53	SINFRUE00000718128	1645	0.49	2E-24	80.56
104	ENSDARE00000403799	970	0.43	SINFRUE00000597153	854	0.58	2E-25	80.3

Appendix A. (cont.).

	Zebrafish			Torafugu			Torafugu vs Zebra[‡]	
No. of markers	Exon ID	Exon length	GC content	Exon ID	Exon length	GC content	E-value	Identity (%)
105	ENSDARE00000404770	1797	0.49	SINFRUE00000667654	1947	0.49	8E-30	84.39
106	ENSDARE00000407314	1174	0.53	SINFRUE00000721499	1122	0.50	1E-98	81.18
107	ENSDARE00000409838	818	0.53	SINFRUE00000709146	1020	0.49	1E-23	82.42
108	ENSDARE00000410488	2042	0.51	SINFRUE00000691278	2082	0.54	5E-105	81.26
109	ENSDARE00000418749	823	0.52	SINFRUE00000730367	919	0.61	4E-33	82.23
110	ENSDARE00000418930	1156	0.51	SINFRUE00000720787	1175	0.50	3E-34	86.11
111	ENSDARE00000420489	1653	0.54	SINFRUE00000561462	1590	0.64	1E-40	84.3
112	ENSDARE00000421998	1027	0.52	SINFRUE00000590718	1030	0.61	2E-72	80.7
113	ENSDARE00000424213	931	0.54	SINFRUE00000771338	938	0.66	2E-31	83.25
114	ENSDARE00000429938	831	0.48	SINFRUE00000805544	840	0.50	1E-23	93.24
115	ENSDARE00000435042	1030	0.55	SINFRUE00000597578	1033	0.65	9E-56	86.84
116	ENSDARE00000435786	1092	0.53	SINFRUE00000606878	1056	0.55	6E-97	88.79
117	ENSDARE00000435942	874	0.50	SINFRUE00000717374	874	0.49	8E-93	79.86
118	ENSDARE00000440228	1767	0.41	SINFRUE00000802590	935	0.49	9E-28	81.9
119	ENSDARE00000440514	17148	0.41	SINFRUE00000777929	894	0.44	3E-24	83.23
120	ENSDARE00000441380	924	0.58	SINFRUE00000582889	930	0.58	2E-38	86.29
121	ENSDARE00000442073	2121	0.45	SINFRUE00000772689	1452	0.48	6E-33	78.6
122	ENSDARE00000442814	2167	0.44	SINFRUE00000577022	821	0.58	5E-66	82.22
123	ENSDARE00000452862	822	0.58	SINFRUE00000618557	951	0.66	1E-36	81.09
124	ENSDARE00000461814	1201	0.52	SINFRUE00000585763	846	0.63	2E-22	84.03
125	ENSDARE00000463567	2094	0.49	SINFRUE00000669034	1725	0.61	3E-23	84.51

Appendix A. (cont.).

	Zebrafish			Torafugu			Torafugu vs Zebra[‡]	
No. of markers	Exon ID	Exon length	GC content	Exon ID	Exon length	GC content	E-value	Identity (%)
126	ENSDARE00000468050	2222	0.45	SINFRUE00000565243	1089	0.58	2E-84	81.96
127	ENSDARE00000472455	862	0.52	SINFRUE00000808658	975	0.58	1E-23	82.94
128	ENSDARE00000472797	1303	0.43	SINFRUE00000642469	828	0.59	8E-31	84.94
129	ENSDARE00000479861	927	0.58	SINFRUE00000569048	960	0.60	3E-37	81.31
130	ENSDARE00000485260	1035	0.56	SINFRUE00000657696	1068	0.52	4E-30	84.85
131	ENSDARE00000490915	3547	0.47	SINFRUE00000717980	1590	0.53	2E-48	78.73
132	ENSDARE00000495706	2848	0.47	SINFRUE00000605519	2866	0.51	3E-24	83.54
133	ENSDARE00000502459	1784	0.45	SINFRUE00000620332	1585	0.55	1E-62	81
134	ENSDARE00000506413	1323	0.52	SINFRUE00000589030	1032	0.56	6E-23	88.12
135	ENSDARE00000509406	1212	0.49	SINFRUE00000691757	1104	0.51	6E-57	79.96
136	ENSDARE00000510312	2289	0.49	SINFRUE00000624350	3542	0.59	6E-32	77.34
137	ENSDARE00000513536	818	0.61	SINFRUE00000649321	807	0.58	1E-88	83.58
138	ENSDARE00000513917	3058	0.53	SINFRUE00000784235	3540	0.58	6E-66	89.19

[†]markers successfully passed through the *in silico* as well experimental tests; [‡]markers passed through the *in silico* but failed in the experimental tests; [§]result of blasting zebrafish sequences with torafugu sequences.

Appendix B. Taxon sampling and AC numbers (accession numbers of sequences determinded in this study are EU001863-EU002148).

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*	*deomestica*	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*	this study		this study	this study		this study			this study	this study
Albuliformes	Albulidae	*Albula*	*vulpes*	this study	this study				this study				
Amiiformes	Amiidae	*Amia*	*calva*	EF032909	EF032922	EF032935	EF032948	EF032961		EF032987	EF033000	EF033013	EF033026
Anguilliformes	Anguillidae	*Anguilla*	*rostrata*	this study	this study		this study		this study		this study		
Argentiniformes	Argentinidae	*Argentina*	*sialis*	this study	this study	this study	this study	this study	this study	this study	this study		this study
Atheriniformes	Atherinopsidae	*Labidesthes*	*sicculus*	this study	this study	this study	this study	this study	this study		this study		this study
Aulopiformes	Synodontidae	*Synodus*	*foetens*	this study	this study	this study	this study	this study	this study	this study	this study		this study
Batrachoidiformes	Batrachoididae	*Porichthys*	*plectrodon*	this study	this study		this study	this study	this study	this study		this study	this study
Beloniformes	Adrianichthyidae	*Oryzias*	*latipes*	EF032914	EF032927	EF032940	EF032953	EF032966	EF032979	EF032992	EF033005	EF033018	EF033031
Beryciformes	Holocentridae	*Myripristis*	*violacea*	this study	this study	this study	this study	this study	this study	this study	this study	this study	this study
Characiformes	Characidae	*Pygocentrus*	*nattereri*		this study	this study	this study		this study		this study	this study	this study
Clupeiformes	Chirocentridae	*Chirocentrus*	*dorab*	this study	this study	this study			this study	this study		this study	this study
Clupeiformes	Clupeidae	*Dorosoma*	*cepedianum*	this study	this study	this study				this study			this study
Clupeiformes	Pristigasteridae	*Pellona*	*flavipinnis*	this study	this study	this study	this study		this study	this study		this study	this study
Cypriniformes	Cyprinidae	*Danio*	*rerio*	EF032910	EF032923	EF032936	EF032949	EF032962	EF032975	EF032988	EF033001	EF033014	EF033027
Cypriniformes	Cyprinidae	*Notemigonus*	*crysoleucas*	this study	this study	this study	this study	this study	this study	this study		this study	this study
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*	this study	this study		this study	this study	this study	this study	this study	this study	this study

Appendix B. (cont.).

Orders	**Families**	**Genus**	**Species**	**zic1**	**myh6**	**RYR3**	**ptr**	**tbr1**	**ENC1**	**Glyt**	**SH3PX3**	**plagl2**	**sreb2**
Elopiformes	Elopidae	*Elops*	*saurus*	this study	this study	this study	this study	this study	this study		this study	this study	this study
Esociformes	Esocidae	*Esox*	*lucius*	this study	this study	this study	this study		this study	this study	this study	this study	this study
Gadiformes	Gadidae	*Gadus*	*morhua*	this study	this study		this study		this study	this study	this study	this study	this study
Gadiformes	Macrouridae	*Coryphaenoides*	*rupestris*		this study		this study		this study	this study	this study		
Gasterosteiformes	Gasterosteidae	*Gasterosteus*	*aculeatus*	EF032912	EF032925	EF032938	EF032951	EF032964	EF032977	EF032990	EF033003	EF033016	EF033029
Gonorynchiformes	Chanidae	*Chanos*	*chanos*	this study	this study		this study	this study	this study			this study	this study
Gymnotiformes	Apteronotidae	*Apteronotus*	*albifrons*	this study			this study	this study	this study		this study	this study	this study
Lampriformes	Regalecidae	*Regalecus*	*glesne*	this study	this study		this study			this study	this study	this study	this study
Lepisosteiformes	Lepisosteidae	*Lepisosteus*	*osseus*	this study	this study	this study						this study	this study
Lophiiformes	Lophiidae	*Lophius*	*gastrophysus*	this study	this study		this study	this study	this study		this study	this study	this study
Mugiliformes	Mugilidae	*Mugil*	*curema*	this study	this study	this study	this study	this study	this study	this study	this study	this study	this study
Myctophiformes	Neoscopelidae	*Neoscopelus*	*macrolepidotus*	this study	this study		this study	this study	this study	this study	this study		this study
Ophidiiformes	Ophidiidae	*Brotula*	*multibarbata*	EF032920	EF032933	EF032946	EF032959	EF032972	EF032985	EF032998	EF033011	EF033024	EF033037
Osmeriformes	Osmeridae	*Thaleichthys*	*pacificus*	this study	this study		this study	this study	this study	this study	this study		this study
Osteoglossiformes	Hiodontidae	*Hiodon*	*alosoides*	this study	this study	this study	this study	this study	this study		this study	this study	this study
Osteoglossiformes	Osteoglossidae	*Osteoglossum*	*bicirrhosum*	this study				this study	this study			this study	this study
Perciformes	Cichlidae	*Cichlasoma*	*cyanoguttatum*	this study	this study		this study	this study	this study	this study	this study	this study	this study
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*	this study	this study	this study	this study	this study	this study	this study	this study		this study
Pleuronectiformes	Pleuronectidae	*Pleuronectes*	*platessa*	this study	this study	this study	this study	this study		this study	this study	this study	this study
Polymixiiformes	Polymixiidae	*Polymixia*	*japonica*	this study	this study	this study	this study		this study	this study	this study		
Polypteriformes	Polypteridae	*Polypterus*	*senegalus*	this study		this study	this study	this study		this study		this study	this study
Salmoniformes	Salmonidae	*Oncorhynchus*	*mykiss*	EF032911	EF032924	EF032937	EF032950	EF032963	EF032976	EF032989	EF033002	EF033015	EF033028

Appendix B. (cont.).

Orders	**Families**	**Genus**	**Species**	**zic1**	**myh6**	**RYR3**	**ptr**	**tbr1**	**ENC1**	**Glyt**	**SH3PX3**	**plagl2**	**sreb2**
Scorpaeniformes	Sebastidae	*Sebastes*	*ruberrimus*	this study	this study	this study	this study	this study	this study	this study	this study	this study	
Siluriformes	Ictaluridae	*Ictalurus*	*punctatus*	EF032916	EF032929	EF032942	EF032955	EF032968	EF032981	EF032994	EF033007	EF033020	EF033033
Stomiiformes	Stomiidae	*Stomias*	*boa*	this study	this study		this study	this study	this study	this study	this study		this study
Synbranchiformes	Synbranchidae	*Monopterus*	*albus*	this study	this study	this study	this study	this study	this study	this study	this study	this study	this study
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*	this study	this study	this study	this study		this study	this study	this study	this study	this study