First publ. in: Trends in genetics 20 (2004), 9, pp. 424-431

Novel evolutionary relationship among four fish model systems

Wei-Jen Chen[1,*], Guillermo Ortí[2] and Axel Meyer[1]

[1] Lehrstuhl für Zoologie und Evolutionsbiologie, University of Konstanz, 78457 Konstanz, Germany
[2] School of Biological Sciences, University of Nebraska, Lincoln, NE 68588, USA

Knowledge of the correct phylogenetic relationships among animals is crucial for the valid interpretation of evolutionary trends in biology. Zebrafish, medaka, pufferfish and cichilds are fish models for development, genomics and comparative genetics studies, although their phylogenetic relationships have not been tested rigorously. The results of phylogenomic analysis based on 20 nuclear protein-coding genes confirmed the basal placement of zebrafish in the fish phylogeny but revealed an unexpected relationship among the other three species, contrary to traditionally held systematic views based on morphology. Our analyses show that medaka (Beloniformes) and cichlids (Perciformes) appear to be more closely related to each other than either of them is to pufferfish (Tetraodontiformes), suggesting that a re-interpretation of some findings in comparative biology might be required. In addition, phylogenomic analyses show that fish typically have more copies of nuclear genes than land vertebrates, supporting the fish-specific genome duplication hypothesis.

Genomic information of vertebrates is increasing rapidly because of several complete or almost complete vertebrate genome projects (e.g. human, mouse, rat, *Fugu*, zebrafish and medaka). This information is invaluable to comparative biologists. Combined with developmental data gathered from model systems, such as *Xenopus,* mouse, chicken, medaka and zebrafish, and sophisticated post-genomic analyses, comparative genomic studies are starting to discover evolutionary mechanisms underlying the diversification of body forms and the increasing complexity of gene function of vertebrates [1,2]. The correct interpretation of comparative biological data requires an evolutionary framework (i.e. a well-supported phylogeny). Incorrect hypotheses of evolutionary relationships might result in misleading interpretations about, for example, taxon-specific genome characteristics and trends in gene family evolution, such as *Hox* gene cluster evolution [3–5]. The importance and applied examples of combining phylogenetic and comparative biological data to interpret early animal evolution [6] or evolution of the vertebrate limb [7] have been demonstrated [4,7,8]. Unfortunately, a sound and widely accepted phylogenetic hypothesis for major fish lineages is not yet available. Compared with other vertebrate groups, such as mammals, the relationships among the main branches of the evolutionary tree of ray-finned fish [Class ACTINOPTERYGII (see Glossary)] (Box 1) – comprising almost half of all vertebrate species – remain poorly defined. The composition of many taxonomic groups of fish at the ordinal and supra-ordinal level requires validation by rigorous molecular phylogenetic study.

Until now, knowledge of the relationships among the four major fish model systems [zebrafish, medaka, pufferfish and cichlids (pronounced sick-lids)] traditionally relied on a loosely formulated ‘phylogenetic syntheses’ [9] (Figure 1a),

Glossary

Actinopterygii: a clade of ‘bony fish’ in the animal phylum Chordata, comprising the ray-finned fish, which evolved during the end of the Silurian period, ~408 million years ago. They dominate the modern fauna and can be found in most aquatic habitats from the abyssal depths of the ocean, > 10 000 m, to high altitude freshwater streams and ponds. Both ‘fish’ and ‘bony fish’ (fish, other than lampreys, sharks and their relatives, with bony skeletons including ray-finned fish, lungfish and so on) are general terms in popular use but they are not derived from formal taxonomic classifications, in the sense of being derived from a single ancestral lineage (monophyletic group). In contrast, ray-finned fish are considered a monophyletic group. Traditionally three grades of Actinopterygii have been recognized: the Chondrostei (sturgeons, bichirs), Holostei (gars, bowfins) and Teleostei (the vast majority of extant ray-finned species). An example of how a ray-finned species are placed into a systematic hierarchical (cladistic) classification based on grouping taxa by their shared possession of derived characters is shown in Figure I.

Clade: a cluster of taxa derived from a single common ancestor.

Homoplasy: a similarity in character state (e.g. nucleotide in molecular phylogeny) resulting from chance or a selection constraint but not due to common history.

Long-branch attraction: a potential pitfall of phylogenetic reconstruction when the rate of sequence evolution among lineages is significantly different. That is, if a dataset contains some taxa that are placed at the tips of long branches (because of rapid evolution) and some other branches (both external and internal) that are short, then the long branches will attract each other and appear as sister taxa on the tree, even if they do not share recent common ancestry.

Orthologous genes: two homologous genes that evolved directly from their most recent common ancestor via speciation that did not undergo a gene duplication; otherwise they are termed paralogous, if they result from gene duplication.

Phylogenomic approach: an analytical approach in phylogenetics using large-scale or multiple gene loci sequence data (usually with less number of taxonomic representatives) obtained from intensive sequencing efforts and/or available genomic databases in order to reconstruct a well-supported phylogenetic tree of organisms or to understand the evolution of the individual gene trees (i.e. does a gene genealogy tree result from speciation, duplication or both of these two evolutionary processes?). Phylogenomics can be considered as an extension of molecular phylogenetics because it consists of the use of molecular data (not morphological characters) for evolutionary analysis. However, molecular phylogenetics is concerned with attempts to determine the rates and patterns of nucleotide and/or amino acid change occurring within a gene, whereas the focus of phylogenomics is on the evolutionary change among genes or the evolutionary trends of a group of genes such as a gene family. A sophisticated application of phylogenomics is to improve functional predictions for uncharacterized genes [44].

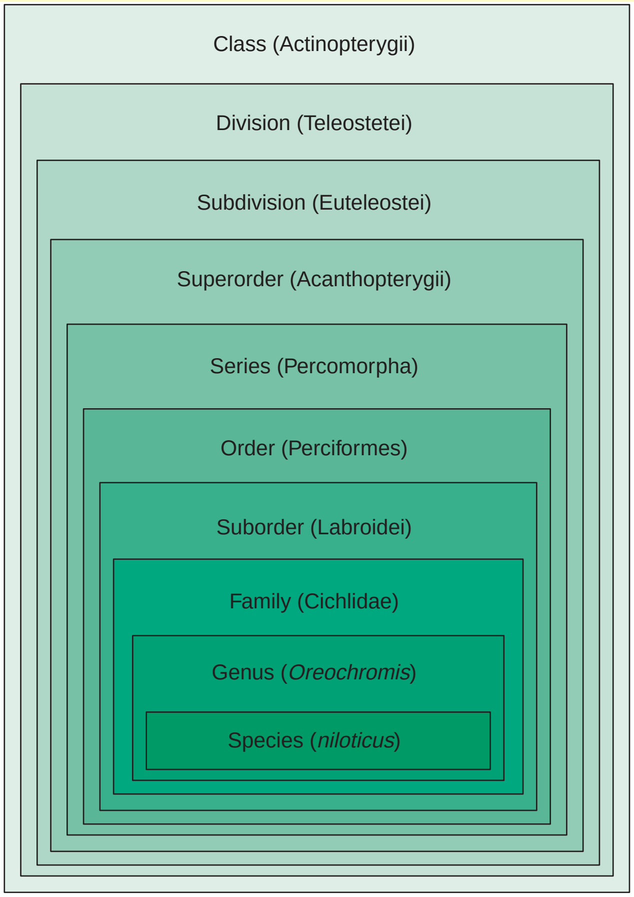
Corresponding author: Wei-Jen Chen (wjchen@bordeaux.inra.fr).

* Current address: INRA Bordeaux, Domaine de la Grande Ferrade, BP81, UMR SV (INRA-ENITAB), 33883 Villenave d'Ornon, France

Available online 21 July 2004

Box 1. Systematic hierarchical classification

An example of how the ray-finned species (*Oreochromis niloticus*) is placed into a systematic hierarchical (cladistic) classification based on grouping taxa by their shared possession of derived characters is shown in Figure I. In general, the hierarchy of categories includes species, genera, families, orders, classes, phyla and kingdoms. To provide a better presentation of the postulated hierarchical relationships, a relatively large number of categories were employed to ‘fish’ (e.g. suborder, series, superorder and so on), once a new monophyletic group was defined according to phylogenetic results. For more information on current recognized fish classification, see Ref. [9].



TRENDS in Genetics

Figure I. The systematic hierarchical categories of the classification of the Actinopterygii using the example of a cichlid species, *Oreochromis niloticus*.

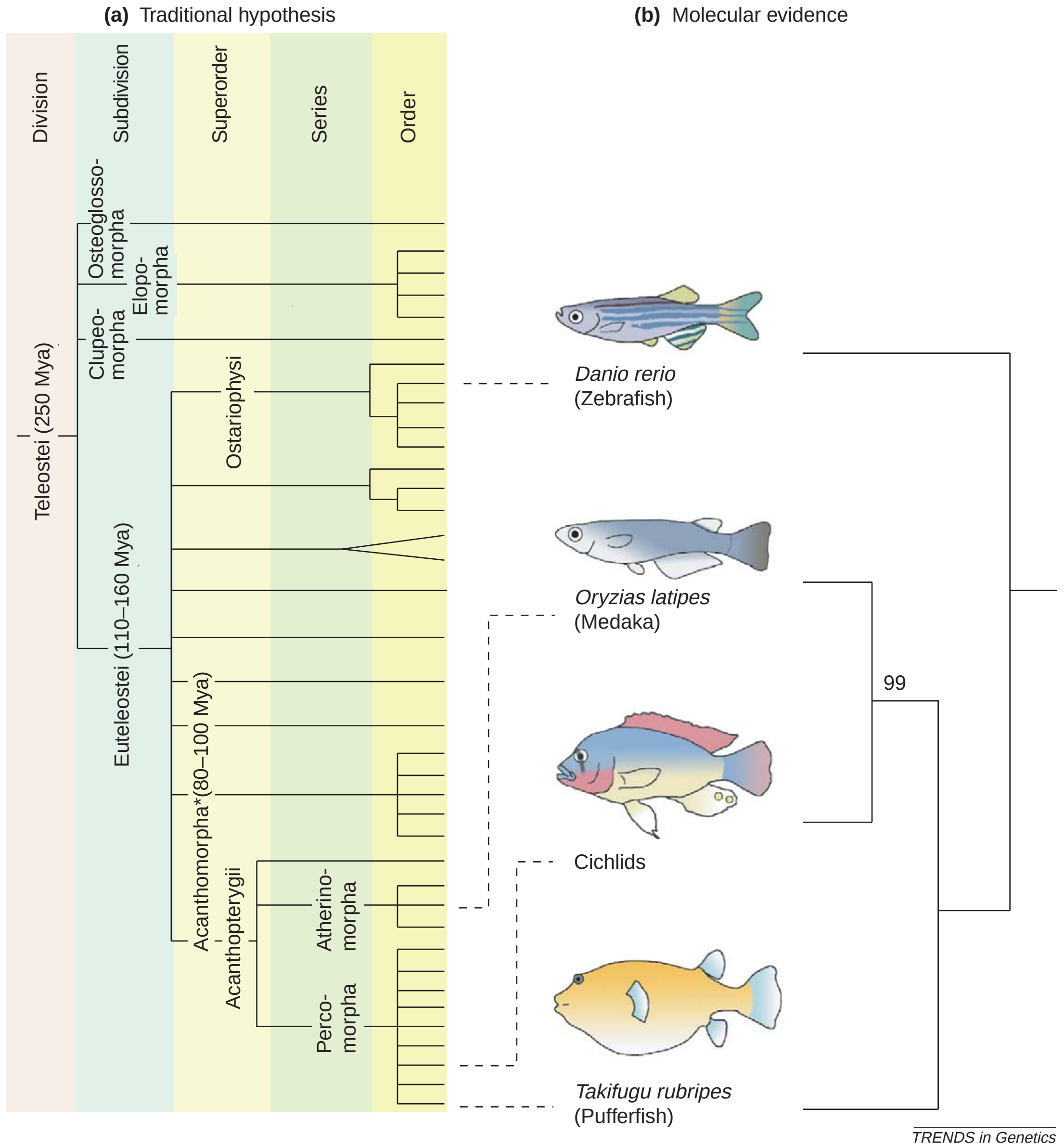
which was based on morphological studies of relatively small subsets of taxa. However, several efforts are now underway that rely on extensive taxonomic coverage of the diversity of ray-finned fish to define high-level systematic relationships (http://www.deepfin.org). These efforts are based on DNA sequences of either whole mitochondrial genomes [10–12] or multiple molecular markers (nuclear and mitochondrial [13]; W-J. Chen and G. Ortí, unpublished), and they provide a viable alternative to the difficult comparative morphological problem of establishing homologies among anatomical structures across widely divergent groups of fish. The emerging molecular evidence challenges some traditional hypotheses based on morphology, especially at the higher-taxonomic levels, and these need to be tested further by more intensive taxonomic sampling.

The goal of this study was to initiate a complementary, PHYLOGENOMIC APPROACH to fish phylogenetics based on sequence data that was assembled from genetic sequence (e.g. GenBank) and genomic (e.g. the complete *Fugu* genome) databases to assess phylogenetic relationships among the major four fish model organisms. We are particularly interested in the evolutionary affinities of cichlids, a newly established model system [14], whose genomic resources, such as ESTs, cDNAs, BAC libraries (A. Meyer *et al.*, unpublished) and microarrays are being compiled in several European, Japanese and USA laboratories (for more information, see the Cichlid Genome Consortium http://hcgs.unh.edu/cichlid/). The elucidation of phylogenetic relationships among fish would aid the study of comparative genomics and developmental evolutionary genetics of basal vertebrates.

What are the phylogenetic relationships among these four model systems?

Zebrafish (*Danio rerio*) belongs to the Family Cyprinidae (Order Cypriniformes). This order, together with other orders of primary freshwater fish that include, tetras, piranhas, catfish, electric knifefish and milkfish, forms a large monophyletic assemblage with ~6500 species known as the Ostariophysi. Recently, molecular and morphological studies [11,15,16] suggested that Clupeomorpha (anchovies and herrings) is the sister-group to Ostariophysi. These two lineages (otocephalans) are hypothesized to branch immediately after two basal teleost fish lineages, the Elopomorpha (eels) and the Osteglossomorpha (bony tongues), and constitute the sister-group of the Euteleostei. Among euteleosts, the most diverse group by far is the Acanthomorpha (true spiny fish), comprising >14 736 species, a result of putative rapid evolution that occurred during the Late Cretaceous period [17]. Pufferfish (*Takifugu rubripes*, Tetraodontidae, Order Tetraodontiformes), medaka (*Oryzias latipes*, Adrianichthyidae, Order Beloniformes) and cichlids (Cichlidae, Labroidei, Order Perciformes) all belong to the Acanthomorpha. Johnson and Patterson [18], based on their examination of morphological characters, assigned the perciformes (perch-like fish, which include cichlids), the Dactylopteriformes (flying gurnards), the Scorpaeniformes (Scorpionfish), the Pleuronectiformes (flatfish), and the Tetraodontiformes (the pufferfish and their relatives) to a single unnamed CLADE. This group, in turn, is the putative sister-group of the Smegmamorpha, which contains five lineages: Synbranchiformes (spiny and swamp eels), Mugiloidei (mullets), Elassomatidae (pigmy sunfish), Gasterosteiformes (pipefish and sticklebacks – another emerging fish model [19]) and Atherinomorpha (medaka and its allies, such as platyfish – another important fish model [20]).

However, recent molecular evidence began to challenge the long-held morphological hypothesis [21] of higher-level systematics of the Acanthomorpha [10,13]. One of the unexpected results arising from the molecular data is that cichlids, pomacentrids (another member family of the suborder Labroidei), a few perciforms and



TRENDS in Genetics

Figure 1. Phylogenetic trees depicting the evolutionary relationships among four fish models. **(a)** The morphology-based (traditional) hypothesis [9], notably, cichlids (Perciformes) and pufferfish (Tetraodontiformes) are more closely related to each other than either of them is to the medaka (Beloniformes, Atherinomorpha). **(b)** The molecular-based hypothesis, contrary to traditionally held systematic views. Acanthomorpha (spiny fish, marked with an asterisk) is currently recognized as a monophyletic group based on morphological evidence and includes Lampridiformes, Polymixiiformes, Paracanthopterygii (cods, goosefish and their relatives) and Acanthopterygii as depicted in the figure. Clupeomorpha should be included in the Euteleostei according to recent morphological and molecular results (see the main text for more details). Abbreviation: Mya, million years ago.

atherinomorphs appear to be more closely related to each other than to the other two labroid member families, labrids (wrasses) and scarids (parrotfish), which are grouped with other perciforms and the Tetraodontiformes (pufferfish). These results are from mitogenomic analyses [10–12] and from a combined dataset of four nuclear and mitochondrial genes [rhodopsin, recombination-activating gene 1 (*RAG-1*), 12S and 16S rDNA, together comprising ~3000 bp] with >65 acanthomorph representatives (W-J. Chen and G. Ortí, unpublished). In Figure 1, a simplified version of the traditional (morphological) hypothesis for these four fish models is compared with the new hypothesis arising from molecular studies.

Reliability of phylogenetic hypotheses

The issue of how phylogenetic hypotheses should be assessed for reliability is crucial, especially when new and somewhat unexpected hypotheses arise. In general, two alternative approaches have been advocated to enhance accuracy in phylogenetic inference. The first approach promotes the use of dense taxonomic sampling (i.e. inclusion of as many taxa as possible) to obtain taxon-rich datasets that will minimize the effect of systematic biases such as long-branch attraction [22,23]. This approach is being pursued with mitogenomic datasets [10] and *RAG-1* gene sequences (W-J. Chen and G. Ortí, unpublished), for several hundred representative fish taxa. The second phylogenomic approach used in this study is based on using as many genes as possible in a few representative lineages of particular interest. This kind of analysis has been useful in testing hypotheses of genome evolution and duplication [24,25] and is also a powerful approach for resolving difficult phylogenetic issues [26]. The benefits of sampling several independent gene genealogies to infer an organismal phylogeny with confidence are well established [13,27,28] because, ultimately, a better representation of the whole genome is highly desirable. Obviously, both approaches are not mutually exclusive, and given enough resources they should be combined to produce taxon-rich and gene-rich datasets for future phylogenetic inference.

In statistics, stochastic errors in the data can lead to incorrect inference when the sample size is small but these errors will disappear with infinite sample size. Thus, it is commonly expected that modern molecular phylogenetic analyses will be based on sequences from multiple gene sources, and will be performed using a 'total evidence' approach [29] or using simultaneous analysis. By combining all available data in a single matrix for analysis, the aspiration is to maximize the congruence of all relevant characters globally to obtain the preferred hypothesis. However, increasing the reliability, as more characters are analyzed, can be obtained only if the following basic assumptions are met: characters should be independent of each other and the distribution of HOMOPLASY (non-historical signal) should be random across data partitions. Unfortunately, examples are accumulating in molecular systematics, which show that homoplasy might accumulate in genes or genomes in ways that are not completely random [30]. The non-random aspects of molecular homoplasy can be understood by analyzing functional constraints and can be detected without phylogenetic tools, for example, by identifying mutational and/or base compositional biases within some positions or regions (i.e. at the third codon position of protein-coding genes). These molecular processes can originate and accumulate non-random homoplasy within a gene and can potentially mislead phylogenetic reconstruction. Another long-recognized potential problem in molecular phylogenetics involves the direct inference of organismal phylogeny from a single gene genealogy. If genes that are used for the analysis consist of a mixture of orthologous and paralogous copies, this inference will probably be incorrect. This particular problem is a major concern for inference involving ray-finned fish, because several putative single copy genes in vertebrates might have been duplicated during the evolution of ray-finned fish [25,31].

Therefore, we chose to perform phylogenetic analyses based on simultaneous (combined) analysis and separate analysis of a large number of genes. By doing so, we are able to consider the repeatability of CLADES as the prime criterion to establish reliability [13]; that is, based on the proportion of concordant phylogenetic results that were obtained from all genes (data partitions) independently. Considering that molecular homoplasy can have different effects on tree reconstruction from one gene to another, obtaining the same clade from the separate analysis of several genes increases the reliability of the results. The same phylogenetic result is expected to emerge from independent data partitions as an indication of a common structure in the data that must originate from common evolutionary (organismal) history. Thus, we focus on the presence of repeated clades among results obtained from sets of trees from different genes. It is unlikely that a false grouping resulting from mistaken orthology would be obtained repeatedly and consistently in the majority of the independently derived gene trees. This does not imply that every instance of mistaken orthology (or any other systematic bias) will necessarily be detected by this approach; rather, the implication is that the 'repeated clades' approach might be more robust in detecting examples of undetected paralogy in individual data partitions.

Towards resolving the evolutionary relationships between four model fish

During the past decade, the advances in molecular biology led to the accumulation of a considerable amount of sequence data of ray-finned fish in GenBank. In addition, several whole fish genomes have been either completely (*Fugu*) or almost completely (*Danio, Tetraodon*) sequenced. The medaka genome is almost finished and is expected to be available to the public this year. Such complete genome information can help to explore all possible homologous copies of particular genes to draw a complete picture of gene genealogies, and to gain a better understanding of genomic and organismal evolution.

To date, with all available sources, we were able to compile and analyze 20 datasets of nuclear protein-coding genes that are shared by the core taxa (the four fish model species) by mining GenBank and the *Fugu* and *Danio* genomic databases. ORTHOLOGOUS GENES and paralogous genes were identified using BLAST searches and phylogenetic reconstruction. First, we identified a common set of nuclear genes shared by cichlids and medaka (and/or by platyfish). Starting with these sequences, we conducted a broad search in the genomic databases for orthologous sequences for the other two fish models (zebrafish and pufferfish), we also included other teleost sequences as appropriate, to expand the taxonomic coverage. There are three important reasons to include additional teleost sequences: (i) they are likely to provide potentially useful information ('phylogenetic signal') to assess the relative phylogenetic position of the four fish models among a larger sample of fish taxa; (ii) a consistent increase in accuracy of phylogenetic inference is expected to be obtained when using more species [32]; and (iii) only by including orthologous and paralogous sequences from all of the available species will potential gene duplication events be discovered, which is essential to understand whether the inferred gene genealogy actually represents the organismal phylogeny. Therefore, each dataset compiled for analysis included all available orthologous and/or paralogous sequences for each gene, from four fish model systems as core taxa (zebrafish, pufferfish, medaka and cichlids), and a large array of diverse teleosts. Basal ray-finned fish (preferably) or sarcopterygian species such as mouse and human were chosen as outgroups to root the trees.

Sequences were aligned with Clustal X [33] based on the inferred amino acid translation. Regions where the amount of variation is high and the resulting alignment

Table 1. Summary of tree topologies depicting phylogenetic relationships among four fish models based on separate analyses of each gene sequence[a,b]

Gene	Size[c]	Nucleotide			Amino acid	
		ME	MP	Bayesian	MP	Bayesian
BMP4	1251	CF[d]	CF	CF	CF	X
cGnRH-II	261	X	X	X	X	X
DMO	672	CF	X	CF	X	CM
DMT	570	CF	X	X	CM	X
GH	633	CM[e]	CM	CM	CM	CM
GnRH	138	X[f]	X	X	X	X
GnRH-recpII	984	CF	CM	CM	CM	CM
HSP70	1860	CM	CM	CM	CM	X
EF1-alpha	1374	X	X	X	X	X
LWS	1062	X	X	CF	CM	CM
Otx2	546	CM	CM	CM	X	X
RH2	972	X	X	X	X	X
CYP1A	1530	CM	CM	CM	CM	CM
CYP19	1380	CM	CM	CM	X	CM
RPL18	393	CM	CF	X	X	X
sGnRH	219	CM	CM	CM	X	X
SWS	1029	CM	CM	CM	CM	CM
TMO4C4	510	X	X	X	X	X
TRF	549	X	X	X	X	X
VIT	2811	CF	CF	CM	CF	CM

[a] Abbreviations: *BMP4*, bone morphogenetic protein 4; *cGnRH-II*, chicken-II-type gonadotropin-releasing hormone; *DMO*, sex-determining protein; *DMT*, sex-determining protein; *GH*, growth hormone; *GnRH*, gonadotropin-releasing hormone; GnRH-recpII, gonadotropin-releasing hormone receptor; *HSP70*, heat shock protein 70; *EF1-α*, elongation factor 1-alph; *LWS*, long wavelength-sensitive cone opsin; *Otx2*, homeodomain protein; *RH2*, green sensitive cone opsin; *CYP1A*, cytochrome P450 1A; *CYP19*, cytochrome P450 aromatase; *RPL18*, ribosomal protein L18; *sGnRH*, salmon-type gonadotropin-releasing hormone; *SWS*, blue sensitive cone opsin; *TMO4C4*, 'titin-like protein'; *TRF*, transferrin; *VIT*, vitellogenin.

[b] Separate analyses with extensive taxa were performed by using maximum parsimony (MP), distance method with optimal criterion of minimum evolution (ME), and the maximum likelihood criterion with posterior probability approach in Bayesian analyses (Bayesian), as implemented in Phylogenetic Analysis Using Parsimony (PAUP*) (Version 4.0b10) [36] and MrBayes (version 2.01) [45], respectively. Taxa involved in these analyses are the same as shown in Figure S1 (see supplementary material online), which consist of the results from MP method at nucleotide level. The likelihood ratio tests, as implemented in model test 3.06 [46], were employed to choose models for ME and Bayesian analyses. Because MrBayes can only be applied with limited number of models and particular substitution types (the number of substitution types can only be one, two or six), approximate models were used instead. In addition, LogDet distance [47] can perform well in ME analysis when the base frequencies are heterogeneous across taxa (detected by using Chi-square test as implemented in PAUP*), ME LogDet+I analyses were performed for such datasets (HSP70, LWS, RH2, CYP1A, CYP19, SWS, TMO-4C4 and VIT). All Bayesian analyses at the amino acid sequence level applied the same model (Dayhoff model with corrected among-site rate variation).

[c] Sequence length (in bp) used in phylogenetic analyses for each gene.

[d] CF indicates that pufferfish was found as a sister group to cichlids in the phylogenetic analysis.

[e] CM indicates that medaka was found as a sister group to cichlids.

[f] X indicates that neither pufferfish nor medaka were found as a sister group to cichlids.

would be likely to contain invalid assertions of homology (e.g. large insertion and/or deletions segments showing high dissimilarity in sequence length) were discarded from the phylogenetic analyses. We performed a wide array of analyses using different phylogenetic methods (e.g. maximum parsimony and Bayesian analyses) for both nucleotide and amino acid sequences to gauge the robustness of the resulting hypothesis. The gene genealogies obtained from each of the separate analyses are used to identify repeated clades across datasets (Table 1 and the supplementary material online). For the combined analysis (total evidence), we identified the putative orthologous sequences for all fish models from all 20 genes but excluded potentially paralogous copies according to the topologies that were obtained for each gene tree (the sequences that were included are marked with an asterisk in the Figure S1 in the supplementary material online). The resulting phylogeny from the total evidence analysis is shown in Figure 1b. The measures of robustness (bootstrap proportions and posterior probabilities) and statistical tests for alternative tree topologies are shown in Table 2.

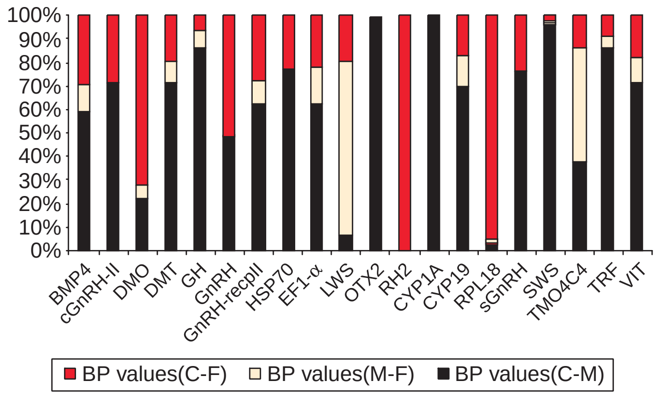
For all phylogenetic analyses based on individual genes, zebrafish (*Danio*, Cypriniformes) was placed basally to the other three model fish. In most cases, the zebrafish branch was placed as a sister group to a clade grouping the Protacanthopterygii (salmons and pikes), stomiiformes and the more derived teleosts such as true spiny-fish or Acanthomorpha. The results from various methods and data partitions (Table 1; Figure S1 in

Table 2. Support for alternative trees assessed by various tests from simultaneous-four taxa based analyses[a]

Tree[b]	Nucleotide						Amino acid		
	BP_MP	BP_ME	BP_ML	PP_ML	TP_MP	SH_ML	BP_MP	PP_ML	TP_MP
1	0.99	0.99	0.97	1	Best	Best	0.99	1	Best
2	<0.01	<0.01	<0.03	0	0.0245	0.0319	<0.01	0	0.0081
3	<0.01	<0.01	<0.03	0	<0.0001	0.0007	<0.01	0	<0.001

[a] Tests were followed by bootstrap proportion (BP), posteriori probability using Bayesian approach (PP), Templeton's nonparametric test (TP), and Shimodaira-Hasegawa's test (SH) under optimal criterion of tree reconstruction of maximum parsimony (MP) and Maximum likelihood (ML).

[b] Three alternative topologies are: Tree 1, zebrafish (*Fugu*(cichlids, medaka)); Tree 2, zebrafish (medaka (cichlids, *Fugu*)); Tree 3, zebrafish, (cichlids (*Fugu*, medaka)).



TRENDS in Genetics

Figure 2. The analyses of bootstrap components for each data partition (gene). The colors in the histograms represent the bootstrap proportions of clades in three alternative hypotheses based on four taxa analysis (zebrafish was chosen as an outgroup). These taxa were taken from each of the listings of bipartitions that were obtained from separate bootstrap analyses conducted by the maximum parsimony method (on nucleotide sequences). The three groupings are cichlids with pufferfish (C–F) in red, medaka with pufferfish (M–F) in yellow, and cichlids with medaka (C–M) in black. Bootstrap resamplings were performed using 500 replicates. Abbreviations: BP, bootstrap proportion; BMP4, bone morphogenetic protein 4; cGnRH-II, chicken-II-type gonadotropin-releasing hormone; DMO, sex-determining protein; DMT, sex-determining protein; GH, growth hormone; GnRH, gonadotropin-releasing hormone; GnRH-recpII, gonadotropin-releasing hormone receptor; HSP70, heat shock protein 70; EF1-α, elongation factor 1-alph; LWS, long wavelength-sensitive cone opsin; OTX2, homeodomain protein orthodenticle homolog 2; RH2, green sensitive cone opsin; CYP1A, cytochrome P450 1A; CYP19, cytochrome P450 aromatase; RPL18, ribosomal protein L18; sGnRH, salmon-type gonadotropin-releasing hormone; SWS, blue sensitive cone opsin; TMO4C4, 'titin-like protein'; TRF, transferrin; VIT, vitellogenin.

supplementary material online) show that a close relationship of cichlids and medaka is recovered repeatedly (number of instances is 7–9 of 20), whereas the cichlid and pufferfish grouping (the traditional hypothesis) is only recovered in a minority of cases (0–5 of 20). The cichlids and medaka grouping was supported by many genes with different functional constraints, indicating that this is a reliable result.

Figure 1b shows the tree of the simultaneous analysis from the concatenated sequences of 20 genes (18 708 bp) based on the maximum parsimony method. The other methods converge on the same topology. The sister group relationship between cichlids and medaka was supported by all methods with a high degree of confidence (97–100% bootstrap values and posterior probability). In addition, this tree was supported significantly by the tests of tree comparison (Table 2).

To understand the variation in potential support and conflict contained in each gene, we performed bootstrap analyses separately for each data partition. Bootstrap components for each data partition are shown in the form of a histogram for each possible clade in Figure 2. There are three possible topologies for a rooted tree with three lineages. If bootstrap values can be regarded as a measure of hierarchical signal to support each topology [34], this histogram represents the relative contribution from each individual data partition to such a signal. The bootstrap signal favors the grouping of cichlids and medaka to the exclusion of *Fugu* (the black bars in Figure 2) in the majority of the data partitions (14 of 20), whereas data partitions of sex-determining protein (DMO), long wavelength-sensitive cone opsin (LWS), green sensitive cone opsin (RH2) and ribosomal protein L18 (RPL18) present a strong signal against the grouping of cichlid and medaka (the yellow or red bars in Figure 2). To investigate the potential source and significance of incongruence between data partitions, we performed a partition-homogeneity test [35] as implemented in Phylogenetic Analysis Using Parsimony (PAUP*) [36]. A test that included all datasets showed significant heterogeneity among data partitions (*P* value = 0.006), whereas a test without these four apparently conflicting genes resulted in no significant incongruence among datasets (*P* value = 99.2). These results indicate that the striking incongruence detected among DMO, LWS, RH2, RPL18 and the other 16 gene genealogies might actually represent evidence for unrecognized paralogy or another systematic bias. A misleading signal could result in a systematic bias, for example, either because of the convergence in base composition (Table 1) or because of long-branch attraction in any of these four conflicting datasets. Fortunately, it is less likely that a wrong grouping would result recurrently from separate analyses for independent genes using a variety of phylogenetic reconstruction methods. Therefore, we conclude that the new hypothesis shown in Figure 1b is reliable.

Conclusions and perspectives

The approach used in this contribution provides a well-supported, yet novel, hypothesis of the evolutionary relationship among the major four fish models. The Atherinomorpha (including the medaka and platy) have been historically placed in an intermediate position among the other branches of the acanthomorph tree (true spiny fish) (Figure 1a) because they share several putative 'primitive' morphological features with more basal teleosts. However, the molecular data support a close relationship between atherinomorphs and putatively more derived perciform fish such the cichlids (Figure 1b). The ancestral features that are found among atherinomorphs include: (i) a protrusible upper jaw that lacks a ball-and-socket joint between the palatine and the maxilla (a feature that prevents the premaxillaries from being locked in the protruded position); and (ii) other generalized features, such as flexible spines on dorsal fines and pelvic fins in abdominal or subabdominal position [9]. According to the phylogeny results presented in this article, these features could be the result of a secondary loss that occurred during the evolution of spiny fish because the atherinomorphs are nested among perch-like fish. The same argument could also apply to the evolution of genomic features and developmental or regulatory mechanisms. For instance, the compact genome that is a characteristic of *Fugu* [37] is most likely the result of independent and unique evolution along the *Fugu* lineage [38], given that the other three fish models do not share this particular genomic feature.

In addition to resolving evolutionary relationships among the major four fish models systems, many examples of duplicated genes in teleost fish were shown from our phylogenomic analyses (Figure S1 in the supplementary material online). Several paralogous copies can be detected that are, presumably, the product of lineage-specific gene duplication events [39] or most likely arose from an ancestral polyploid genome. Among cypriniformes (e.g. *Danio*) and salmonids, the occurrence of polyploid genomes has been documented extensively [40,41]. Nonetheless, some 'duplicates' might be artifacts that were generated by sequencing errors so that two 'paralogous' sequences were deposited in GenBank. In such cases, the similarity between these sequences is usually suspiciously high. Perhaps other duplicates resulted from large-scale gene duplication events, such as the postulated fish-specific genome duplication that happened ~320 million years ago, before the divergence of most teleost species [25,42,43]. For example, the cytochrome P450 aromatase (*CYP19*) gene genealogy (Figure S1 in the supplementary material online) clearly shows two paralogous clusters (*CYP19a* and *CYP19b*) that branch off soon after the separation of sarcopterygians (e.g. tetrapods) and ray-finned fish. Interestingly, the other two cases [blue sensitive cone opsin (SWS) and vitellogenin (VIT)] suggest that these individual gene duplication events happened after the separation of the Ostariophysi (*Danio*) and the higher euteleosts. The rapid progress of genomic resources and developmental information for an increasing number of species – and especially model species – emphasized the imminent importance of a reliable phylogenetic framework in which to interpret comparative results correctly. Simple pairwise comparisons have only limited explanatory power because they lack knowledge of the evolutionary framework.

Acknowledgements

Part of this work was supported by grants from the Deutsche Forschungsgemeinschaft to A.M., and the National Science Foundation (DEB 9985045) to G.O. We thank E. Moriyama and the bioinformatics laboratory at the University of Nebraska for providing computational support. We also thank W. Salzburger for providing the fish pictures, and two anonymous reviewers and the Editor for their comments on the manuscript.

Supplementary data

Supplementary data associated with this article can be found at doi:10.1016/j.tig.2004.07.005

References

1 Holland, P.W.H. (1999) The future of evolutionary developmental biology. *Nature* 402, C41–C44

2 Shimeld, S.M. and Holland, P.W.H. (2000) Vertebrate innovations. *Proc. Natl. Acad. Sci. U. S. A.* 97, 4449–4452

3 Holland, L.Z. and Gibson-Brown, J.J. (2003) The *Ciona intestinalis* genome: when the constraints are off. *Bioessays* 25, 529–532

4 Holland, P.W.H. (1996) Hox genes and chordate evolution. *Dev. Biol.* 173, 382–395

5 Holland, P.W.H. (2003) More genes in vertebrates? *J. Struct. Funct. Genomics* 3, 75–84

6 Knoll, A.H. and Carroll, S.B. (1999) Early animal evolution: emerging views from comparative biology and geology. *Science* 284, 2129–2137

7 Mabee, P.M. (2000) Developmental data and phylogenetic systematics: evolution of the vertebrate limb. *Amer. Zool.* 40, 789–800

8 Tabin, C.J. *et al.* (1999) Out on a limb: Parallels in vertebrate and invertebrate limb patterning and the origin of appendages. *Amer. Zool.* 39, 650–663

9 Nelson, J.S. (1994) *Fishes of the World*, John Wiley and Sons

10 Miya, M. *et al.* (2003) Major patterns of higher teleostean phylogenies: a new perspective based on 100 complete mitochondrial DNA sequences. *Mol. Phylogenet. Evol.* 26, 121–138

11 Inoue, J.G. *et al.* (2003) Basal actinopterygian relationships: a mitogenomic perspective on the phylogeny of the "ancient fish". *Mol. Phylogenet. Evol.* 26, 110–120

12 Ishiguro, N.B. *et al.* (2003) Basal euteleostean relationships: a mitogenomic perspective on the phylogenetic reality of the "Prota-Protacanthopterygii". *Mol. Phylogenet. Evol.* 27, 476–488

13 Chen, W-J. *et al.* (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

14 Kocher, T.D. (2004) Adaptive evolution and explosive speciation: the cichlid fish model. *Nat. Rev. Genet.* 5, 288–298

15 Zaragüeta-Bagils, R. *et al.* (2002) Assessment of otocephalan and protacanthopterygian concepts in the light of multiple molecular phylogenies. *C. R. Biol.* 325, 1191–1207

16 Saitoh, K. *et al.* (2003) Mitochondrial genomics of ostariophysan fishes: perspectives on phylogeny and biogeography. *J. Mol. Evol.* 56, 464–472

17 Patterson, C. (1993) An overview of the early fossil record of acanthomorphs. *Bull. Mar. Sci.* 52, 29–59

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

19 Peichel, C.L. *et al.* (2001) The genetic architecture of divergence between threespine stickleback species. *Nature* 414, 901–905

20 Wittbrodt, J. *et al.* (2002) Medaka–a model organism from the far East. *Nat. Rev. Genet.* 3, 53–64

21 Greenwood, P.H. *et al.* (1966) Phyletic studies of teleostean fishes, with a provisional classification of living forms. *Bull. Am. Mus. Nat. Hist.* 131, 339–455

22 Pollock, D.D. *et al.* (2002) Increased taxon sampling is advantageous for phylogenetic inference. *Syst. Biol.* 51, 664–671

23 Hillis, D.M. *et al.* (2003) Is sparse taxon sampling a problem for phylogenetic inference? *Syst. Biol.* 52, 124–126

24 Durand, D. (2003) Vertebrate evolution: doubling and shuffling with a full deck. *Trends Genet.* 19, 2–5

25 Taylor, J.S. *et al.* (2003) Genome duplication, a trait shared by 22000 species of ray-finned fish. *Genome Res.* 13, 382–390

26 Rosenberg, M.S. and Kumar, S. (2003) Taxon sampling, bioinformatics, and phylogenomics. *Syst. Biol.* 52, 119–124

27 Zardoya, R. and Meyer, A. (1996) Phylogenetic performance of mitochondrial protein-coding genes in resolving relationships among vertebrates. *Mol. Biol. Evol.* 13, 933–942

28 Cummings, M.P. *et al.* (1995) Sampling properties of DNA sequence data in phylogenetic analysis. *Mol. Biol. Evol.* 12, 814–822

29 Kluge, A.G. (1989) A concern for evidence and a phylogenetic hypothesis of relationships among Epicrates (Boidae Serpentes). *Syst. Zool.* 38, 7–25

30 Naylor, G.J. and Brown, W.M. (1998) Amphioxus mitochondrial DNA, chordate phylogeny, and the limits of inference based on comparisons of sequences. *Syst. Biol.* 47, 61–76

31 Taylor, J.S. *et al.* (2001) Comparative genomics provides evidence for an ancient genome duplication event in fish. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 356, 1661–1679

32 Zwickl, D.J. and Hillis, D.M. (2002) Increased taxon sampling greatly reduces phylogenetic error. *Syst. Biol.* 51, 588–598

33 Thompson, J.D. *et al.* (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* 25, 4876–4882

34 Hillis, D.M. and Bull, J.J. (1993) An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Syst. Biol.* 42, 182–192

35 Farris, J.S. *et al.* (1995) Testing significance of incongruence. *Cladistics* 10, 315–319

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

37 Aparicio, S. *et al.* (2002) Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. *Science* 297, 1301–1310

38 Neafsey, D.E. and Palumbi, S.R. (2003) Genome size evolution in pufferfish: a comparative analysis of diodontid and tetraodontid pufferfish genomes. *Genome Res.* 13, 821–830

39 Robinson-Rechavi, M. *et al.* (2001) An ancestral whole-genome duplication may not have been responsible for the abundance of duplicated fish genes. *Curr. Biol.* 11, R458–R459

40 Allendorf, F.W. and Thorgaard, G. (1984) Polyploidy and the evolution of salmonid fishes. In *The Evolutionary Genetics of Fishes* (Turner, B.J. ed), pp. 1–53, Plenum Press

41 Larhammar, D. and Risinger, C. (1994) Molecular genetic aspects of tetraploidy in the common carp *Cyprinus carpio*. *Mol. Phylogenet. Evol.* 3, 59–68

42 Hoegg, S. *et al*. (2004) Phylogenetic timing of the fish-specific genome duplication correlates with phenotypic and taxonomic diversification in fishes. *J. Mol. Evol.* 59, 190–203

43 Vandepoele, K. *et al*. (2004) Major events in the genome evolution of vertebrates: paranome age and size differ considerably between ray-finned fishes and land vertebrates. *Proc. Natl. Acad. Sci. U. S. A.* 101, 1638–1643

44 Eisen, J.A. (1998) Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Res.* 8, 163–167

45 Huelsenbeck, J.P. and Ronquist, F. (2001) MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* 17, 754–755

46 Posada, D. and Crandall, K.A. (1998) MODEL TEST: testing the model of DNA substitution. *Bioinformatics* 14, 817–818

47 Lockhart, P.J. *et al*. (1994) Recovering evolutionary trees under a more realistic model of sequence evolution. *Mol. Biol. Evol.* 11, 605–612