Molecular Biology and Evolution

Different Phylogenomic Approaches to Resolve the Evolutionary Relationships among Model Fish Species

Enrico Negrisolo,*[1] Heiner Kuhl,[2] Claudio Forcato,[3] Nicola Vitulo,[4] Richard Reinhardt,[2] Tomaso Patarnello,[1] and Luca Bargelloni[1]

[1]Department of Public Health, Comparative Pathology and Veterinary Hygiene, University of Padova, Agripolis, Legnaro, Italy

[2]Max-Planck-Institute for Molecular Genetics, Berlin, Germany

[3]Department of Environmental Agronomy and Crop Science, University of Padova, Agripolis, Legnaro, Italy

[4]Department of Biology, Centro Ricerche Interdepartimentale Biotecnologie Innovative Biotechnology Centre, University of Padova, Padova, Italy

***Corresponding author:** E-mail: enrico.negrisolo@unipd.it.

Associate editor: Hervé Philippe

Research article

Abstract

Comparative genomics holds the promise to magnify the information obtained from individual genome sequencing projects, revealing common features conserved across genomes and identifying lineage-specific characteristics. To implement such a comparative approach, a robust phylogenetic framework is required to accurately reconstruct evolution at the genome level. Among vertebrate taxa, teleosts represent the second best characterized group, with high-quality draft genome sequences for five model species (*Danio rerio*, *Gasterosteus aculeatus*, *Oryzias latipes*, *Takifugu rubripes*, and *Tetraodon nigroviridis*), and several others are in the finishing lane. However, the relationships among the acanthomorph teleost model fishes remain an unresolved taxonomic issue. Here, a genomic region spanning over 1.2 million base pairs was sequenced in the teleost fish *Dicentrarchus labrax*. Together with genomic data available for the above fish models, the new sequence was used to identify unique orthologous genomic regions shared across all target taxa. Different strategies were applied to produce robust multiple gene and genomic alignments spanning from 11,802 to 186,474 amino acid/nucleotide positions. Ten data sets were analyzed according to Bayesian inference, maximum likelihood, maximum parsimony, and neighbor joining methods. Extensive analyses were performed to explore the influence of several factors (e.g., alignment methodology, substitution model, data set partitions, and long-branch attraction) on the tree topology. Although a general consensus was observed for a closer relationship between *G. aculeatus* (Gasterosteidae) and *Di. labrax* (Moronidae) with the atherinomorph *O. latipes* (Beloniformes) sister taxon of this clade, with the tetraodontiform group *Ta. rubripes* and *Te. nigroviridis* (Tetraodontiformes) representing a more distantly related taxon among acanthomorph model fish species, conflicting results were obtained between data sets and methods, especially with respect to the choice of alignment methodology applied to noncoding parts of the genomic region under study. This may limit the use of intergenic/noncoding sequences in phylogenomics until more robust alignment algorithms are developed.

Key words: teleosts, phylogenomics, model fish species, electronic chromosome painting.

Introduction

At the dawn of molecular systematics, phylogenies were based on single genes or even gene fragments (e.g., Field et al. 1988). Since then, the increasing throughput capacity of DNA sequencing technology has made available an ever-growing amount of sequence information, mainly in the form of large collections of expressed sequence tags (ESTs) or genome sequences (e.g., Clark et al. 2003; Jaillon et al. 2004). To take advantage of such a wealth of data, phylogenomics was proposed to replace molecular phylogenetics (Rokas et al. 2003; Philippe, Delsuc, et al. 2005), through the use of genomic rather than single gene information. The rationale for a phylogenomic approach is that data sets that are one or several orders of magnitude larger than traditional ones should make possible more robust phylogenetic reconstructions, allowing the resolution of difficult taxonomic problems; however, large data sets, similar to traditional ones, might be subject to several sources of error (e.g., compositional biases, model misspecification, and saturation of the phylogenetic signal), which, because of data set size, could lead to a much stronger bias and produce highly supported, yet incorrect, phylogenetic trees (Philippe, Delsuc, et al. 2005; Jeffroy et al. 2006).

In most cases, and especially in vertebrate taxonomy, phylogenomics actually means multiple gene phylogenies, produced through the simple addition of single gene or EST information (Chen et al. 2004; Philippe and Telford 2006; Steinke et al. 2006). A major drawback of this approach is that it multiplies the risk of comparing orthologous with paralogous gene copies as a consequence of lineage-specific evolution of ancestrally duplicated genes, which is especially important for molecular phylogenies of vertebrates that have experienced more than one round of whole-genome duplication (WGD; Taylor et al. 2003; Jaillon et al. 2004). In particular, and highly relevant to the present study, the teleost genome underwent an additional WGD

during its evolutionary history, followed by chromosomal rearrangements mainly in the form of fusions and translocations (Taylor et al. 2003; Jaillon et al. 2004; Kasahara et al. 2007; Nakatani et al. 2007). Such major genomic changes represent a further challenge to establish the relationships of orthology versus paralogy. The use of large, contiguous genomic sequences could provide a more reliable set of data for phylogenomic analysis because the issue of gene orthology and paralogy can be evaluated in the context of chromosome evolution (e.g., Kassahn et al. 2009). To this end, the identification of orthologous genomic regions derived from the same proto-chromosome might help prevent potential artifacts due to the presence of ohnologous chromosomal regions (i.e., paralogous chromosomal regions produced by a WGD) in the data set.

In the present study, a chromosome electronic painting strategy (e.g., Kasahara et al. 2007; Nakatani et al. 2007) was implemented to address this issue, which is particularly relevant for the ingroup species.

However, the analysis of large genomic regions, which include mostly noncoding sequences (introns, promoters, intergenic regions), poses new challenges, or more appropriately, old challenges in a new form. Reliable bioinformatic tools that are able to appropriately handle genomic sequences (Brudno et al. 2003; Kurtz et al. 2004; Margulies et al. 2006) are the most important requirement for phylogenomic studies based on genomic alignments. The crucial point, as outlined in recent articles, is the quality of multiple alignment for large genomic sequences (e.g., Wong et al. 2008). In the present study, the most recent tools for phylogenomic analysis are evaluated, using as a case study the relationships among the teleost model species that are included in the Acanthomorpha (true spiny fish; Rosen 1973).

The traditional classification places *Danio rerio* (zebra fish) within Cypriniformes (family Cyprinidae), *Oryzias latipes* (medaka) within Beloniformes (family Adrianichthyidae), *Dicentrarchus labrax* (European sea bass) within Perciformes (family Moronidae), *Gasterosteus aculeatus* (three-spined stickleback) within Gasterosteiformes (family Gasterosteidae), and *Takifugu rubripes* (Japanese puffer fish) and *Tetraodon nigroviridis* (spotted green puffer fish) within Tetraodontiformes (family Tetraodontidae). The latter five taxa belong to the Acanthomorpha, whereas *Da. rerio* can be used as an outgroup reference taxon to address the phylogenomic relationships among Acanthomorpha. Within acanthomorphs, *O. latipes* is included in the series Atherinomorpha, whereas *Di. labrax*, *G. aculeatus*, *Te. nigroviridis*, and *Ta. rubripes* are classified in the Percomorpha series (Nelson 2006). The Gasterosteiformes and Perciformes orders are polyphyletic groups, as proven by previous molecular analyses (Miya et al. 2003, 2005; Mabuchi et al. 2007; Kawahara et al. 2008; Setiamarga et al. 2008; Li et al. 2009). Different phylogenetic reconstructions based on either morphological characters or molecular data (nuclear genes as well as complete mitochondrial genomes) suggest a substantial reconsideration of the Percomorpha as currently accepted (Nelson 2006), with the inclusion of Beloniformes within this series (Johnson and Patterson 1993; Miya et al. 2003, 2005; Smith and Wheeler 2004; Smith and Craig 2007), among various other changes.

Resolving the relationships among teleost fish lineages belonging to the Acanthomorpha is a major task in vertebrate taxonomy due to the extremely large size of the Acanthomorpha clade (Li et al. 2009). The complexity of this task is further increased as a consequence of the rapid diversification of acanthomorphs, which started approximately 200 Ma and gave rise to more than one-third of all extant vertebrate species, including several model organisms (Nelson 2006). It is well known that radiation-like events leave little time to accumulate lineage-specific sequence divergence, and therefore, the phylogenetic signal is often quite limited. If a radiation occurred in the distant past, multiple independent changes have likely accumulated on the same sites, further reducing information content. Molecular phylogenies dealing with the evolutionary relationships among acanthomorphs have identified several monophyletic taxa departing from the traditional taxonomic arrangement, but many issues remain to be solved (Miya et al. 2001, 2003, 2005; Chen et al. 2003; Smith and Wheeler 2004; Dettai and Lecointre 2005; Smith and Craig 2007; Yamanoue et al. 2007; Kawahara et al. 2008; Li et al. 2009).

Despite their relevance in diverse fields such as developmental biology, genetics, and comparative genomics, just to mention a few, a limited number of studies have investigated the relationships among model fish species (Chen et al. 2004; Steinke et al. 2006; Yamanoue et al. 2006). Furthermore, in these studies, taxon sampling was not constant and did not fully overlap with the species considered in the present study. Conversely, the teleost fishes analyzed here have been included in different incomplete combinations in previous molecular phylogenies (Miya et al. 2001, 2003, 2005; Chen et al. 2003; Smith and Wheeler 2004; Dettai and Lecointre 2005; Smith and Craig 2007; Yamanoue et al. 2007; Kawahara et al. 2008). A very recent article (Li et al. 2009) included all the species considered in the present study, but it could not completely resolve their phylogenetic relationships.

Over 1.2 million base pairs were sequenced from contiguous genomic clones of *Di. labrax* that were then assembled in a single scaffold. The newly determined genomic sequence was used in combination with the existing data for other teleost fishes to implement different methodological phylogenomic approaches, some of which have been applied for the first time in the present study, to resolve the evolutionary relationships within the acanthomorph model species.

Materials and Methods

Genomic Sequencing of 10 *Dicentrarchus labrax* Bacterial Artificial Chromosome Clones

Clones from the *Di. labrax* bacterial artificial chromosome (BAC) library (Whitaker et al. 2006) were comparatively mapped by end sequencing and BlastN alignment (Altschul et al. 1990) to the *G. aculeatus* genomic sequence (Kuhl

et al. 2010). BAC DNA from 10 clones covering approximately 1.3 Mb in *Di. labrax* was isolated by alkaline lysis. Subsequently, the remaining *Escherichia coli* DNA was removed by cesium chloride density gradient centrifugation or ATP-dependent exonuclease digestion. Purified BAC DNA was sheared by ultrasonic sound and size selected to 1- to 4-kb fragments. Fragments were end polished with T4 DNA polymerase and DNA polymerase I (Klenow) and afterward ligated with T4 –DNA ligase into a puC19 sequencing vector. Competent *E. coli* DH10B cells were transformed by electroporation. For each BAC, a library with approximately 10× coverage was constructed, and plasmid DNA was purified for sequencing using ABI BigDye v3.1 Terminator chemistry on ABI 3730xl capillary sequencers. Raw sequences were processed by Phred, and BACs were assembled using Phrap (http://www.phrap.org). The *Di. labrax* whole scaffold encompassing 10 BACs is available in GenBank under the accession number FP017272.

Annotation of the *Dicentrarchus labrax* Genomic Region Sequenced

Gene prediction in *Di. labrax* scaffold FP017272 was performed using a custom bioinformatic platform developed at the CRIBI bioinformatics laboratory of Padua University, which combines multiple and heterogeneous sources of information to predict gene locations. The platform is formed by three modules: 1) ab initio predictions, 2) alignment of ESTs, and 3) alignment of proteins. The ab initio predictors are probabilistic models generally based on hidden Markov models gain prediction ability through a training data set that is composed of annotated known genes. The input is the raw nucleotide sequence, and the output is the gene structure. For gene prediction in the *Di. labrax* scaffold, the ab initio gene finders implemented in the GenScan (Burge and Karlin 1997) and GeneID (Parra et al. 2000) programs were used. In this analysis, the parameter "appropriate organism" was set to vertebrate for GenScan and to *Te. nigroviridis* for GeneID. A collection of fish proteins downloaded from the Ensembl database (http://www.ensembl.org/) was aligned against the *Di. labrax* scaffold using a custom pipeline based on the BLAT (Kent 2002) and GeneWise (Birney et al. 2004) programs. Furthermore, fish ESTs downloaded from the Unigene database (http://www.ncbi.nlm.nih.gov/unigene) were aligned using the est2genome algorithm (Mott 1997).

Different types of prediction tools may produce conflicting results. JigSaw software (Allen and Salzberg 2005), a tool developed to combine evidence coming from heterogeneous data, was used to resolve such conflicts. JigSaw creates a sort of consensus determining a clear and an unambiguous gene structure. The JigSaw outputs were used for the final structure prediction of various genes.

Identification of Orthologous Genomic Regions

The genomes of *Da. rerio*, *G. aculeatus*, *O. latipes*, *Ta. rubripes*, and *Te. nigroviridis* were searched to identify orthologous counterparts of the newly determined genomic portion of *Di. labrax*. The last genome release was used to achieve this task. Data were downloaded from the Ensembl Web site (Ensembl 53 release, March 2009; http://www.ensembl.org/index.html). The identification of orthologous genomic regions was straightforward for *G. aculeatus*, *O. latipes*, *Ta. rubripes*, and *Te. nigroviridis*. Initially, multiple basic local alignment search tool (BLAST) searches (Altschul et al. 1990) were run against each genome using as queries a 2-kb sequence every 50 kb along the genomic sequence of *Di. labrax*. The BLAST results provided evidence to select the contig/chromosome including them. Finally, the 5′ and 3′ ends of the *Di. labrax* genomic region were used to better define the boundaries of the selected syntenic genomic regions. The orthology among selected genomic regions was further assessed through electronic chromosome painting and pairwise and multiple alignments performed with the Multi-LAGAN and MUMmer 3.0 programs (Brudno et al. 2003; Kurtz et al. 2004; see below).

The identification of an unambiguous orthologous genomic region for *Da. rerio* required the more complex approach outlined below. Initially, the *Di. labrax* 1.2-Mb sequence was aligned against the whole genome of *Da. rerio* using the MUMmer 3.0 program, which is software specifically devoted to the pairwise rapid alignment of large genomes (Kurtz et al. 2004). The PROmer script of the MUMmer package was applied to perform this analysis. Two genomic alignments were performed using a masked and an unmasked version of the *Da. rerio* genome. Default parameters were relaxed to allow the comparison among divergent sequences. The length of maximal exact matches (−l option) and the minimum length of the cluster (−c option) were reduced. Conversely, the maximum allowed distance between matches within a cluster was increased (−g option). Alignments obtained through MUMmer (see Results) allowed the identification of a non-ambiguous orthologous region located on *Da. rerio* chromosome 18 that was added to the set of orthologous sequences used to build up the genomic multiple alignment (see below).

Identification of Orthologous/Paralogous Chromosomes

The MUMmer program was used to perform pairwise genomic alignments of whole set of *G. aculeatus* chromosomes against *O. latipes* chromosomes 3 and 6, and *Te. nigroviridis* chromosomes 5 and 13. The electronic chromosome painting approach described below was used to assess orthology/ohnology relationships among chromosomes of the fishes under study.

Tetraodon nigroviridis chromosome 5, *O. latipes* chromosome 3, and *Da. rerio* chromosome 7 are included in a cluster of orthologous chromosomes, whereas *Te. nigroviridis* chromosome 13, *O. latipes* chromosome 6, and *Da. rerio* chromosomes 18 and 25 belong to a second cluster of orthologous chromosomes (see Results for further details; Kasahara et al. 2007).

Genes contained in *Te. nigroviridis* chromosomes 5 and 13, *O. latipes* chromosomes 3 and 6, *G. aculeatus* chromosome 2, and *Da. rerio* chromosomes 7, 18, and 25 were downloaded from the Ensembl Web site. Five data sets were created, each including the genes of a single species. Pairwise comparisons among the five data sets were performed to identify the best bidirectional hit (BBH) for each gene in every data set (Hulsen et al. 2006). Comparisons were made using the BlastP algorithm (Altschul et al. 1990), and BLAST results were parsed with a cutoff value of $e = 10^{-5}$, 30% identity, 70% similarity, and 70% coverage. Identification of BBHs allowed us to assign the orthologous genes to their respective chromosomes. This strategy represents a crude estimation of orthology relationships in some cases (Koski and Golding 2001; Kuzniar et al. 2008) because it does not allow us to identify the evolutionary mechanisms that distort the orthology, such as the birth-and-death process or gene conversion/recombination (Nei 2005). We tried to minimize this problem using the stringent parameters listed above. Furthermore, despite the limits mentioned, the BBH strategy allowed the straightforward comparison of our results with those produced in previous analyses of fish chromosomes that had been performed all according to a BBH approach (Jaillon et al. 2004; Kasahara et al. 2007; Nakatani et al. 2007; Kassahn et al. 2009). Finally, orthology among chromosome regions inferred through BBH analysis was further corroborated by genomic pairwise alignments performed with the MUMmer program (see above).

Figures depicting chromosomes based on electronic painting were produced with a custom software program. This software is a homemade Perl script that uses the GD graphic library (http://www.libgd.org). The program uses as an input file the parsed BLAST output file containing the list of orthologous genes, chromosome gene positions, and chromosome lengths. The output of the program is a chromosome painting picture in png format.

Gene/Genomic Multiple Alignments

Three strategies were applied to align gene/genomic sequences. First, each group of orthologous single gene sequences, which were identified through BBH comparison and further assessed through visual inspection of BLAST results, was used to produce a multiple sequence alignment (MSA). Initially, the encoded polypeptides were aligned using the version 6 of the MAFFT program (Katoh et al. 2002, 2005), which has been shown to be one of the most accurate programs to obtain MSAs (Nuin et al. 2006; Wilm et al. 2006; Carroll et al. 2007). MAFFT was used according to the default settings available at its Web site (http://align.bmr.kyushu-u.ac.jp/mafft/online/server/). The amino acid MSAs were used as the backbone to align the corresponding nucleotide sequences with the DAMBE program (Xia and Xie 2001).

In a second approach, the non-ambiguous orthologous genomic regions were divided into five partitions, and each partition was aligned with the MAFFT program. Partitioning was necessary because the use of MAFFT proved computationally unfeasible when the alignment of the entire genomic region was attempted. The boundaries of each partition were selected to maximize the matching of orthologous segments identified through pairwise genomic alignments performed with MUMmer (see Supplementary fig. S1, Supplementary Material online).

Finally, a single multiple alignment of orthologous genomic regions was produced using the Multi-LAGAN program that has been especially developed for this task (Brudno et al. 2003). The MSA was performed by applying the default parameters implemented in the Multi-LAGAN version available at the VISTA web server (http://genome.lbl.gov/vista/lagan/submit.shtml; Frazer et al. 2004). The translated anchoring option was also used to improve the alignment quality among divergent genomic sequences.

Two options for masking coding regions were used prior to performing MSAs with MAFFT and Multi-LAGAN.

In the full-masking approach, all identified protein-coding regions were masked before running the alignment program, and thus, the MSA was restricted to non–protein-coding regions. In the partial-masking approach, the coding sequences that were not present and arranged in the same order in all analyzed taxa were masked. In the partial-masking approach, five polypeptides/coding sequences present in the unambiguous orthologous genomic regions identified for the six analyzed species (see Supplementary fig. S2 and table S1, Supplementary Material online) were left unmasked.

A third option was implemented in Multi-LAGAN, keeping all coding and non–protein-coding sequences unmasked, irrespective of the presence/absence of the corresponding orthologous counterparts in different species.

In addition to MAFFT and Multi-LAGAN, the two programs FSA (Bradley et al. 2009) and Mauve (Darling et al. 2004) were evaluated. However, in both cases, the total length of the alignment was drastically reduced (<15 kb) after implementing Gblocks (see below; Castresana 2000), and for this reason, the data sets obtained with FSA and Mauve were not further analyzed.

Genomic alignments, obtained through MAFFT and Multi-LAGAN, were further processed with the Gblocks program to prevent/minimize the violation of the positional homology principle (Castresana 2000). Blocks of conserved nucleotides were selected by applying the default settings (gaps not allowed).

Construction of Data Sets for Phylogenetic Analysis

Different data sets were produced for phylogenetic purposes following four alignment strategies:

Type 1 MSAs—multigenes phylogenomic data sets: Genes located in the unambiguous orthologous genomic regions of ingroup species were used as a starting point to create MSAs (see Supplementary fig. S2 and table S1 for a fine-scale annotation of these regions, Supplementary Material online).

Putative orthologs were identified through BBH comparison and subsequent visual inspection of BLAST results. At the end of this process, 20 genes were retained that

could be unambiguously aligned and were present in all analyzed fish species. Amino acid and nucleotide MSAs were produced for each set of orthologous genes following the procedure described in the previous subsection. Finally, the amino acid MSAs obtained with MAFFT were concatenated in a single data set hereafter called $MSA1_{PRmf}$ (11,802 amino acids). The corresponding nucleotide MSAs were concatenated in a single data set named MSA2p1-$p3_{CSmf}$ (35,406 bp). An additional MSA (MSA2p1-$p2_{CSmf}$, 23,604 bp) was created removing the third codon positions from MSA2p1-$p3_{CSmf}$.

Type 2 MSAs—genomic alignments with MAFFT: Five nonoverlapping partitions of the entire genomic region (fully masked and partially masked), aligned with MAFFT and processed with Gblocks, produced blocks of conserved nucleotides that were concatenated in the data set $MSA3_{NCmf}$ (151,068 bp) and the $MSA4_{CD+NCmf}$ alignment (163,775 bp).

Type 3 MSAs—genomic alignments with Multi-LAGAN: MSAs performed with Multi-LAGAN and processed with Gblocks produced the sets $MSA5_{CD+NCmla}$, $MSA6_{NCmla}$, and $MSA7_{CD+NCmla}$. The $MSA5_{CD+NCmla}$ (85,439 bp) was obtained from partially masked genomic sequences and $MSA6_{NCmla}$ (74,711 bp) from fully masked genomic sequences. $MSA7_{CD+NCmla}$ (88,036 bp) was produced through the alignment of unmasked genomic sequences.

Type 4 MSA—combination of MSAs obtained with type 1 and 2 strategies: A combined data set, $MSA8_{CSmfNCmf}$ (186,474 bp), was produced merging MSA2p1-$p3_{CSmf}$ with $MSA3_{NCmf}$.

Type 5 MSA—combination of MSAs obtained with type 1 and 3 strategies: A combined data set, $MSA9_{CSmfNCmla}$ (110,117 bp), was obtained merging MSA2p1-$p3_{CSmf}$ with $MSA6_{NCmla}$.

Testing for Mutational Saturation

The level of mutational saturation was estimated by plotting uncorrected *P*-distances (based on observed substitutions) against ML-estimated distances (i.e., general time reversible [GTR] + *I* + *G*) for each MSA. After fitting a regression line, its slope (*m*) was used as a measure of mutational saturation. If $m = 1$, no saturation is inferred, whereas for $m << 1$, the phylogenetic signal is largely saturated.

Phylogenetic Analysis

Phylogenetic trees were inferred using Bayesian inference (BI), maximum likelihood (ML), maximum parsimony (MP), and neighbor joining (NJ) methods (Felsenstein 2004). The ProtTest program was used to select the best-fitting evolutionary model for protein MSAs (Abascal et al 2005). Models that best fitted the nucleotide MSAs were identified using the Modeltest program according to the Akaike criterion (Posada and Crandall 1998). Analyses on nucleotide MSAs were based on codon, GTR + *I* + *G*, and CAT-GTR (Lartillot and Philippe 2004) models, or a combination of them, depending on the type of multiple alignment. Simpler evolutionary models were also used to evaluate the effect of model selection. Partitioning of multiple alignments was applied when appropriate in BI and ML analyses (Nishihara et al. 2007). The number of data set partitions ranged from 1 to 21.

BI trees were obtained with MrBayes 3.2 (Ronquist and Huelsenbeck 2003) and PhyloBayes 3.2d (Lartillot et al. 2009). In MrBayes, two simultaneous runs, each of four chains, were performed in all analyses. Each run consisted of 100,000–1,000,000 generations, and trees were sampled every 10–100 generations. Stationarity was considered to be reached when the average standard deviation of split frequencies was less than 0.001. Burn-in was also increased respectively to 50%, 70%, and 90% without any appreciable change in tree topology and posterior probability values. Analyses performed with PhyloBayes were carried out following the guidelines provided in the program manual. Stationarity was considered to be reached when maxdiff was less than 0.1 between two independent runs. Once stationarity was reached, a minimum of 1,000 trees was used to generate a majority-rule posterior consensus tree.

ML analyses were performed using PhyML 3 (Guindon and Gascuel 2003; Guindon et al. 2009), TREEFINDER (Jobb et al. 2004), PAUP* (Swofford 2002), and RaxML 7.3 (Stamatakis 2006).

An exhaustive search approach was applied with the steepest descent option activated in the PAUP* program (Swofford et al. 1996). In ML analyses done with RaxML, 10 independent runs using randomized MP starting trees were performed to assess the stability of the obtained tree topology.

MP analyses were done using algorithms implemented in the MEGA 4 (Tamura et al. 2007) and/or PAUP* program (Swofford 2002; Tamura et al. 2007).

NJ trees were computed for both amino acid and nucleotide MSAs by applying several evolutionary models available in MEGA 4 or in PAUP*.

Statistical Tests on Tree Topologies

Nonparametric bootstrap (BT) tests (Felsenstein 1985) were performed to assess the robustness of ML, MP, and NJ tree topologies (1,000 replicates in all cases). Posterior probabilities were calculated for each node of the BI trees.

The approximately unbiased (AU) and the weighted Shimodaira and Hasegawa (WSH) tests (Shimodaira 2002) were performed for all tree topologies to evaluate alternative phylogenetic hypotheses supported by the different data sets. First, all 105 topologies that can be obtained from a data set containing the six taxa were generated with the program PAUP*. Subsequently, AU and WSH values were calculated for all the topologies using TREEFINDER.

Compositional Heterogeneity—Mutational Saturation

To minimize the effects of compositional bias, three approaches were followed: 1) MSA recoding (Phillips et al. 2004), 2) selective removal of sites within MSAs (Ruiz-Trillo

et al. 1999), and 3) application of mixture models (Pagel and Meade 2004).

1. MSA recoding: Nucleotide MSAs were recoded using the simplified R(AG) Y(CT) scheme (Phillips et al. 2004). In $MSA1_{PRmf}$, amino acids were recoded into four categories [Dayhoff4 (A,G,P,S,T) (D,E,N,Q) (H,K,R) (F,Y,W,I,L,M,V) (C = ?)] that allow us to apply a GTR + *G* + *I* substitution model (Rodriguez-Ezpeleta et al. 2007). All recoded data sets were analyzed as described for the original MSAs. Recoded data sets are listed as RY-MSA_i or Dayhoff4-$MSA1_{PRmf}$.
2. Selective removal of sites within MSAs: Among-site heterogeneity was modeled using a gamma distribution approximated by eight rate categories. This procedure was repeated for every MSA using TREE1, TREE2 (see Results), or their strict consensus tree as the reference topology. Each position was assigned to one of these eight categories (from a low to a high level of heterogeneity). Among-site rate heterogeneity parameters were calculated with PUZZLE 5.2 (Schmidt et al. 2002). Sites were selectively removed from MSAs using a Perl script developed by one of the authors (N.V.), starting from positions in category 8 (C8), which includes the most variable sites, and ending with the exclusion of positions in category 4 (C4). Data sets obtained deleting C4 to C8 positions were analyzed according to an ML criterion using PhyML 3.
3. Application of mixture models: On selected MSAs, BI phylogenetic analyses were performed with BayesPhylogenies, a program that implements a mixture model that allows the user to fit more than one model of sequence evolution without partitioning the data (Pagel and Meade 2004). Different mixture models involving up to eight independent substitution matrices were applied, either alone or in combination with a four-categories gamma distribution.

Effect of the Alpha Parameter Variation on Tree Topology

The parameter alpha determines the shape of the gamma distribution (Felsenstein 2004). Estimation of alpha may be subject to stochastic error, potentially affecting the obtained tree topology. To test the stability of the ML topologies recovered from various MSAs, the shape parameter alpha was varied according to fixed values ranging from 0.05 (implying a strong rate heterogeneity among sites) to 5 (weak rate heterogeneity). ML phylogenetic trees obtained imposing different alpha values were compared using the best ML topology that was obtained using the estimated value of unconstrained alpha for every analyzed MSA.

Results

Identification of Orthologous Chromosomes/Genomic Regions

The newly determined region of the *Di. labrax* genome is a 1,296,077-bp scaffold derived from shotgun sequencing of 10 BAC clones. By applying an *in silico* annotation approach (see Supplementary fig. S2 and table S1, Supplementary Material online), at least 26 putative peptide-coding sequences could be identified. These putative genes are distributed along the entire sequenced region, belong to different gene families, and do not form any specific gene cluster.

Through BLAST searches, we identified a single genomic region spanning nearly 1 Mb in each acanthomorph species considered in the present work (table 1). Pairwise alignments of these genomic sequences performed with Multi-LAGAN proved to be straightforward and revealed a very high level of sequence conservation. The length of pairwise alignments ranged from 141,047 bp with sequence identity of 78.2% (*O. latipes* vs. *Ta. rubripes*) to 547,902 bp with a sequence identity of 79.4% (*Di. labrax* vs. *G. aculeatus*), with an average length of 265,347 ± 130,962 bp and mean nucleotide identity of 78.86 ± 0.64%. For three of the four acanthomorph species under study, it was possible to unambiguously assign the unique genomic sequence matching the *Di. labrax* genomic scaffold to a single chromosome, namely chromosome II of *G. aculeatus*, chromosome 3 of *O. latipes*, and chromosome 5 of *Te. nigroviridis* (table 1). For *Ta. rubripes*, three separate, nonoverlapping scaffolds were identified. Unfortunately, the *Ta. rubripes* genome (release 4.0) consists of 7,213 contigs, all of which have yet to be assigned to a specific chromosome.

A pairwise alignment between the 1.3-Mb genomic sequence of *Di. labrax* and the whole genome of *Da. rerio* was performed using the MUMmer package, which allowed us to align chromosomes that have experienced local inversions, a likely phenomenon in more distantly related species such as *Di. labrax* and *Da. rerio*. The first 500,000 bp of the *Di. labrax* sequence exhibited significant matches with chromosome 7 (24,093 positions, with 61.04% identity), chromosome 18 (32,499 positions, with 54.93% identity), and chromosome 25 (40,047 positions, with 57.96% identity) of *Da. rerio* (fig. 1). The remaining part of the *Di. labrax* genomic sequence could be aligned only with a region of *Da. rerio* chromosome 18.

As already mentioned, the ancestral teleost genome experienced a WGD followed by rearrangements of some chromosomes (Kasahara et al. 2007, fig. 2*A*). In some lineages, genomes were subject to further changes including single chromosome duplications, as in the case of *Da. rerio* (fig. 2*B*). Kasahara et al. (2007) reconstructed teleost genome evolution for three species (*O. latipes*, *Te. nigroviridis*, and *Da. rerio*). We also identified homology relationships for specific chromosomes in *G. aculeatus* (fig. 2C). Electronic chromosome painting based on BBHs confirmed that *O. latipes* chromosome 3, *Te. nigroviridis* chromosome 5, and *Da. rerio* chromosome 7 are orthologous (Kasahara et al. 2007). Additionally, our analysis revealed the corresponding ortholog in *G. aculeatus* (chromosome II) (fig. 2C). Likewise, ohnology (paralogy) to the above chromosomes could be inferred for chromosome 6 in *O. latipes*, for chromosome 13 in *Te. nigroviridis*, chromosomes 18 and 25 in *Da. rerio* (Kasahara et al. 2007), and chromosome XIX in *G. aculeatus* (fig. 2C). For *O. latipes*, *Te. nigroviridis*, and

Table 1. Genomic Regions Included in the Data Set Aligned with Multi-LAGAN.

Species	Syntenic Genome Portion			Starting Sequence in Ensembl	
	Start	End	Length (bp)	Genome Source	SEP
Danio rerio	26100000	28000000	1,900,000	Chromosome 18	26100000–28000000
Oryzias latipes	550000	1317476	767,476	Chromosome 3	20984098–22301573
Dicentrarchus labrax	500000	1296077	796,077	Single scaffold	1–1296077
Gasterosteus aculeatus	400000	1010000	610,000	Chromosome group II	13030204–11681984
Takifugu rubripes	260000	787168	527,168	Scaffolds 309, 2075, and 14	(309) 1–287026; (2075) 1–29866; (14) 2331662 –1864477
Tetraodon nigroviridis	400000	920000	520,000	Chromosome 5	4000000—4997803

NOTE.—SEP, start/end point in Ensembl sequences.

G. aculeatus (all acanthomorph species), a single match was found with a region located on orthologous chromosomes, whereas in the case of the outgroup species *Da. rerio*, the first 500,000 bp of *Di. labrax* exhibited matches both with

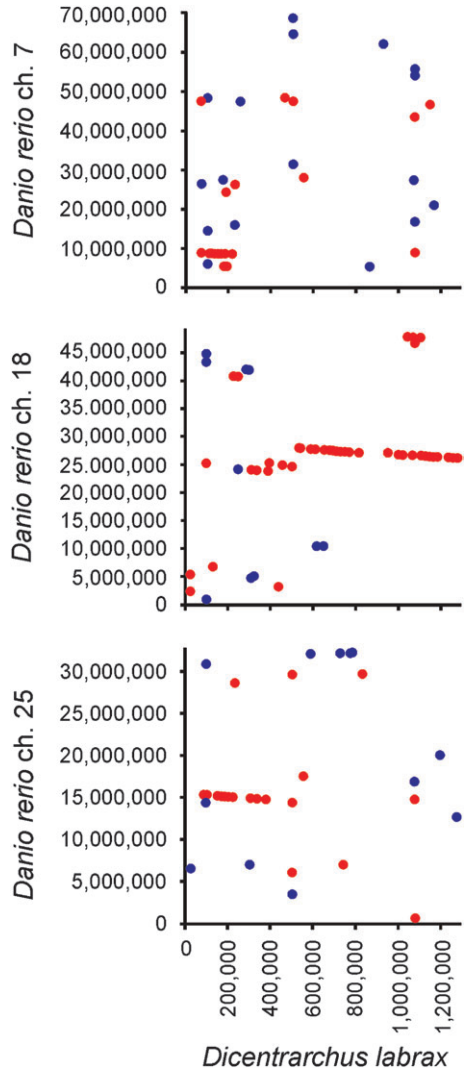


FIG. 1. Pairwise alignment between the 1.2-Mb genomic region of *Dicentrarchus labrax* and the whole genome of *Danio rerio*, performed with the MUMmer package. Matches on the same strand are presented as blue dots, whereas matches in the opposite strands are presented as red dots. Numbers on both axes refer to base positions in the genomic region/chromosomes.

the orthologous chromosome 7 and with its ohnologs, chromosomes 18 and 25, although with different degrees of sequence conservation (fig. 2C). Thus, orthologous and paralogous copies of this region seem to have been retained in the *Da. rerio* genome. The remaining part of the *Di. labrax* scaffold unambiguously matched only with *Da. rerio* chromosome 18. The simplest explanation for this finding would require two independent deletion events, one affecting chromosome 7 and the other on chromosome 25, which led to retention of only one paralogous copy on chromosome 18. However, the analysis of genomic sequences flanking the studied region suggests a more complex scenario. The upstream region on stickleback chromosome II contains three genes (ENSGACP00000021012, ENSGACP00000021022, and ENSGACP00000021024) that have orthologs on *Da. rerio* chromosome 7 (fig. 3). Likewise, three genes (ENSGACP00000021151, ENSGACP00000021157, and ENSGACP00000021159), which represent the downstream boundary on stickleback chromosome II, have their homologues on *Da. rerio* chromosome 7 (fig. 3). Gene sequences with the best reciprocal hits to the *G. aculeatus* chromosome II genes mentioned above are also present in *Da. rerio* chromosomes 18 and 25 (fig. 3). However, synteny is disrupted in both chromosomes 18 and 25 because not all counterparts are found on *Da. rerio* chromosomes 18 and 25 (fig. 3). The conservation of the upper and lower boundaries delimiting the matching regions on, respectively, *G. aculeatus* chromosome II and *Da. rerio* chromosome 7 indicates that at least eight independent genomic modifications (six deletions and two translocations) were necessary to produce the present genomic organization (fig. 3*A*). An alternative and plausible evolutionary scenario involving the matching fragment on chromosome 18 would also require eight genomic rearrangements in the *Da. rerio* genome (five deletions and three translocations) (fig. 3*B*). A translocation between *Da. rerio* chromosomes 7 and 18 had already been identified by Kasahara et al. (2007), although the precise limits of this event were not precisely reported. In the case of translocation (fig. 3*B*), the matching region on zebra fish chromosome 18 will represent the true ortholog of the sea bass scaffold and their corresponding sequences on the other acanthomorph species.

Because of its higher degree of sequence conservation and the consequent better alignment quality, for all subsequent analyses, only the region spanning 2 Mb (26–28

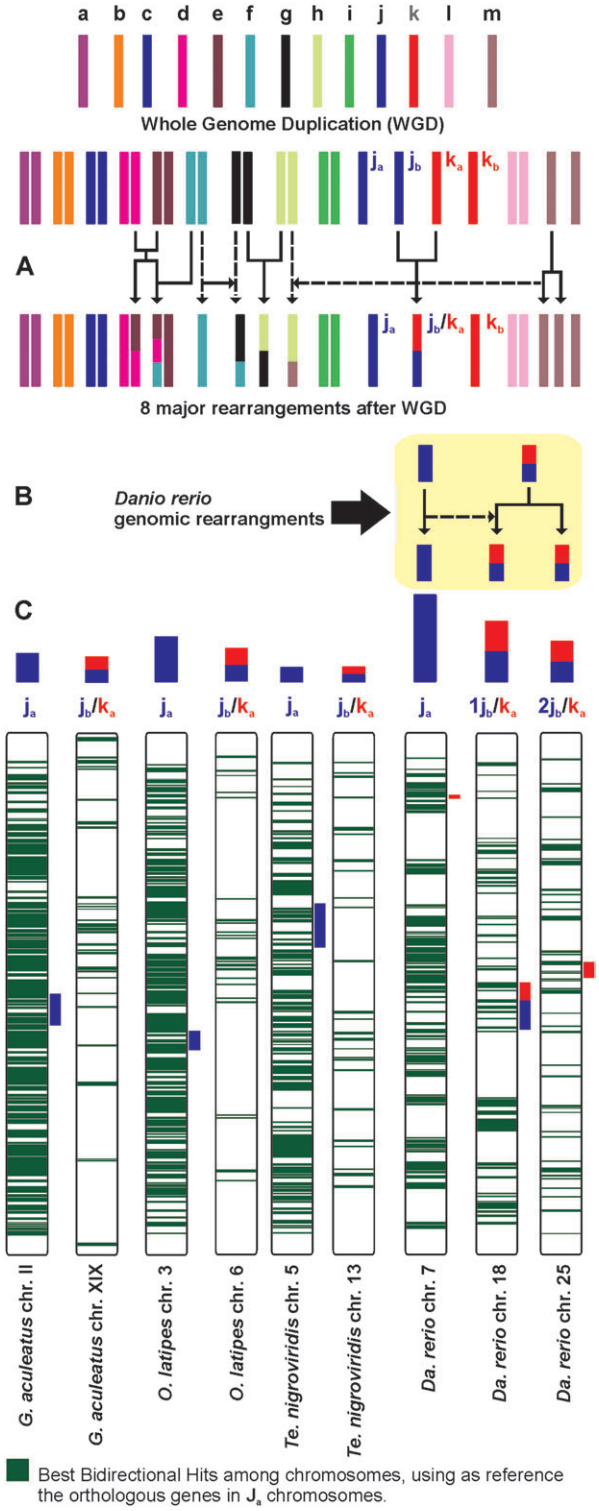


FIG. 2. (*A*) The ancestral teleost genome was subject to a WGD followed by at least eight major rearrangements (redrawn and modified from Kasahara et al. 2007). (*B*) Further single chromosome and/or regional duplications as well as chromosomal rearrangements characterized the evolution of *Danio rerio* genome. Here, only changes that occurred in zebra fish chromosomes 7, 18, and 25 are represented (redrawn and modified from Kasahara et al. 2007). (*C*) Electronic chromosome painting of selected chromosomes. Blue/red miniatures of orthologous/ohnolog chromosomes are drawn to scale. Electronic painted chromosomes are not to scale. Vertical dark-blue bars indicate unique genomic sequences in each species that could be aligned with the portion of the *Dicentrarchus labrax* genome that we sequenced. Vertical red bars indicate multiple homologous segments that are located on different zebra fish chromosomes and can be aligned with the same region of the *Di. labrax* genomic sequence.

Mb) of *Da. rerio* chromosome 18 was retained. In fact, even under the first scenario, with the *Da. rerio* chromosome 18 segment being a true paralogous copy, *Da. rerio* represents the outgroup species; therefore, rooting of the ingroup tree can be appropriately attained by applying a paralogous rooting strategy (e.g., Iwabe et al. 1989; Gribaldo and Cammarano 1998; Kollman and Doolittle 2000; Zhaxybayeva et al. 2005). The corresponding genomic regions identified in *Te. nigroviridis*, *Ta. rubripes*, *G. aculeatus*, *O. latipes*, and *Di. labrax* were trimmed by removing their 5′ segment that produced matches with different chromosomes of *Da. rerio*. Table 1 summarizes the final sequences retained for subsequent phylogenomic analyses (see also Supplementary fig. S1 and table S1, Supplementary Material online).

Assessing Mutational Saturation in MSAs

The mutational saturation present in the original 10 MSAs (table 2) and in their RY/Dayhoff4 recoded version MSAs was investigated by considering the *m* slope of the regression line passing through the origin of an *xy* scatter-plot where the *P*-distance values were listed on the *x* axis and the GTR + *I* + *G* values were listed on the *y* axis. The obtained results are summarized in figure 4. Among the original MSAs, the lowest mutational saturation was identified in $MSA1_{PRmf}$, $MSA2p1\text{-}p2_{CSmf}$, $MSA5_{CD+NCmla}$, $MSA6_{NCmla}$, and $MSA7_{CD+NCmla}$, whereas $MSA3_{NCmf}$, $MSA4_{CD+NCmf}$, and $MSA8_{CSmfNCmf}$ exhibited the highest saturation. Thus, the genomic MSAs produced with Multi-LAGAN had a lower saturation than those created with MAFFT. Recoded MSAs always showed a lower saturation than the original MSAs and exhibited a similar pattern concerning the saturation of different MSAs. The reduction in saturation was very marked on $RY\text{-}MSA2p1\text{-}p2_{CSmf}$, $RY\text{-}MSA6_{NCmla}$, $RY\text{-}MSA7_{CD+NCmla}$, and $RY\text{-}MSA9_{CSmfNCmla}$.

Phylogenetic Inference

Analyses performed on 10 data sets using BI, ML, MP, and NJ methods and applying different evolutionary models produced two alternative topologies (henceforth TREE1 and TREE2), which are presented in figure 5. The critical point was represented by the placement of *O. latipes*, which represented the sister taxon of all other acanthomorph fishes in TREE1 and the sister species of (*Di. labrax* + *G. aculeatus*) in TREE 2 (fig. 5).

All analyses performed on $MSA1_{PRmf}$ produced the same topology (TREE1). Nodes 1, 3, and 4 received very high statistical support (BI = 1.00 and BT $\geq$ 99%) irrespective of the phylogenetic method applied (BI, ML, MP, NJ) and of the applied evolutionary model. Node 2 in TREE1 received strong BI support (BI = 1.00), whereas BT values ranged

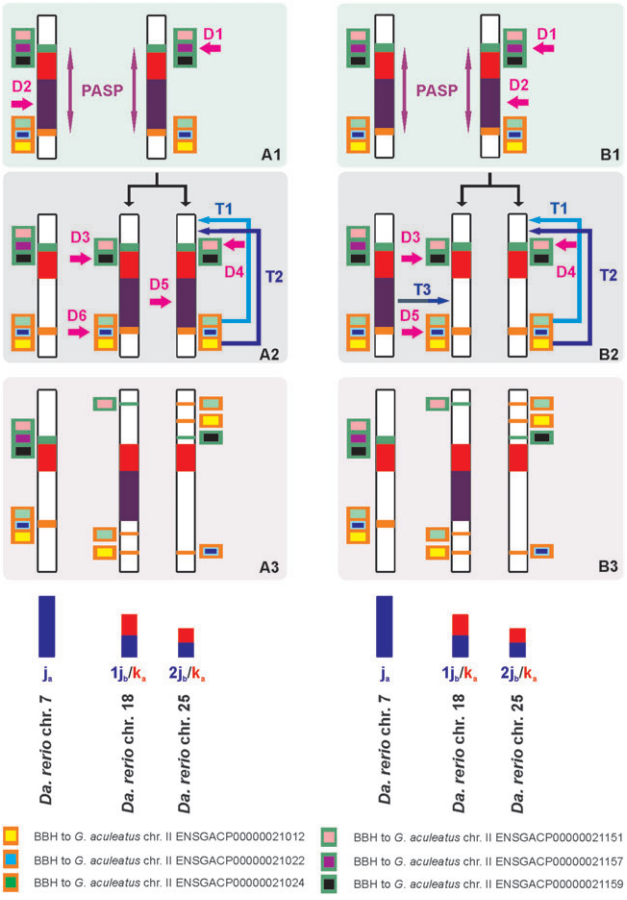


Fig. 3. Alternative genomic rearrangements in the evolution of *Danio rerio* chromosomes 7 18, and 25. Purple double arrow points to putative ancestral syntenic portions (PASP) among *Gasterosteus aculeatus* chromosome II and *Da. rerio* chromosome 7 and the ancestral chromosome for zebra fish chromosomes 18 and 25. 5′ boundary, orange bar; 3′ boundary, green bar; D, deletion; T, translocation.

from 58% to 100% depending on the different methods and evolutionary models.

Analyses performed on MSA2p1-p2$_{CSmf}$ and MSA2p1-p3$_{CSmf}$ produced conflicting results depending on the method/model used. MP and NJ always favored TREE1, with strong statistical support for all nodes (BT $\geq$ 99%). BI analyses based on the GTR + *I* + *G* model applied to a single partition identified TREE2 as the best topology. However, very strong statistical support for node 5 was obtained excluding the third-codon positions (MSA2p1-p2$_{CSmf}$; BI = 1.00), whereas no significant BI was observed when including them (MSA2p1-p3$_{CSmf}$; BI = 0.61). When MSA2p1-p2$_{CSmf}$ and MSA2p1-p3$_{CSmf}$ were divided into 20 partitions and independent GTR + *I* + *G* models were applied to each partition, the first alignment supported TREE2 (BI =1.00), whereas the second data set strongly favored TREE1 (BI =1.00). Finally, BI analysis of MSA2p1-p2$_{CSmf}$ under the CAT + GTR model favored TREE2 with strong statistical support for node 5 (BI = 0.96), whereas the analysis performed on MSA2p1-p3$_{CSmf}$ identified a different topology with a clade (*O. latipes* + [*Ta. rubripes* + *Te. nigroviridis*]) that did not received any support (BI = 0.73). All ML analyses performed on either MSA2p1-p3$_{CSmf}$ or MSA2p1-p2$_{CSmf}$ using single/multiple partition(s) recovered TREE2 with very strong BT (86–100%) support for node 5.

How does the inclusion of noncoding genomic regions, which represent the largest part of the data set, influence the outcome of phylogenetic reconstructions? All analyses performed on MSA3$_{NCmf}$, MSA4$_{CD+NCmf}$, and MSA8$_{CSmfNCmf}$ multiple alignments, that is, MSAs obtained when using MAFFT on noncoding genomic regions, with or without subsequent addition of coding genes aligned separately, always supported TREE1. All nodes received very high statistical support (BI = 1.00 and BT $\geq$ 99%) irrespective of the phylogenetic method applied (BI, ML, MP, NJ). The same was true irrespective of the applied evolutionary model. Models ranged from an uncorrected *P*-distance (NJ) across all sites to the implementation of 20 independent codon-based substitution models for coding genes and a GTR + *I* + *G* for noncoding sequences (MSA8$_{CSmfNCmf}$, 21 partitions).

A more complex scenario resulted from analyses based on Multi-LAGAN MSAs (MSA5$_{CD+NCmla}$, MSA6$_{NCmla}$, MSA7$_{CD+NCmla}$, and MSA9$_{CSmfNCmla}$). MP analyses always recovered TREE1, and each node received very high statistical support (BT $\geq$ 99%). NJ analyses favored TREE2 with very high BT support ($\geq$99%) at all nodes for most data sets and models (from Juke–Cantor to composite likelihood),

Table 2. Data Sets Description.

MSA Name	Data Type	Size[a]	ALNp
MSA1$_{PRmf}$ [g]	PR	11,802	MAFFT (mf)
MSA2p1-p2$_{CSmf}$ [g]	CS	23,604	MAFFT (mf)
MSA2p1-p3$_{CSmf}$ [g]	CS	35,406	MAFFT (mf)
MSA3$_{NCmf}$	NC	151,068	MAFFT (mf)
MSA4$_{CD+NCmf}$	CD + NC	163,775	MAFFT (mf)
MSA5$_{CD+NCmla}$	CD + NC	85,439	Multi-LAGAN (mla)
MSA6$_{NCmla}$	NC	74,711	Multi-LAGAN (mla)
MSA7$_{CD+NCmla}$	CD + NC	88,036	Multi-LAGAN (mla)
MSA8$_{CSmfNCmf}$ [g]	CS + NC	186,474	MAFFT (mf)
MSA9$_{CSmfNCmla}$ [g]	CS + NC	110,117	MAFFT (mf)/Multi-LAGAN (mla)

Note.—PR, protein; CS, codons; NC, noncoding DNA; CD, coding DNA; ALNp, alignment program. Superscript "g" refers to MSA including gaps derived through the alignment with MAFFT of amino acids/codons (see Materials and Methods for details).

[a] Number of positions in the alignment.

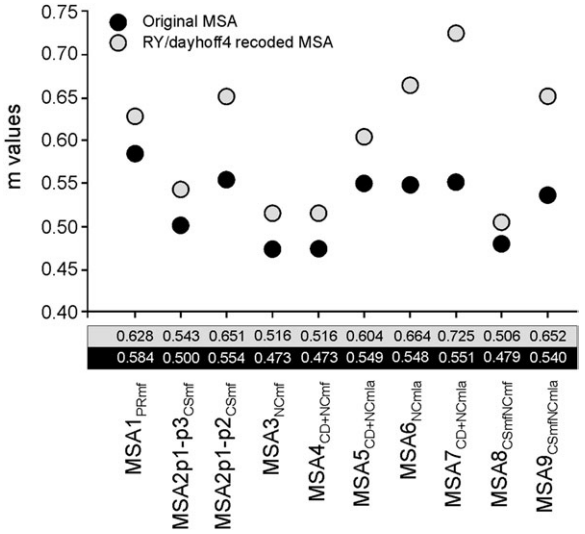


FIG. 4. Mutational saturation of MSAs. Scatter plot of *m* values. The *m* is the slope of the $y = mx$ regression line, where $x_i = P_i$-distance and $y_i = ML_i$-distance. Solid black, *m* values calculated for the original MSAs; gray background, *m* values computed for RY/Dayhoff4 recoded MSAs.

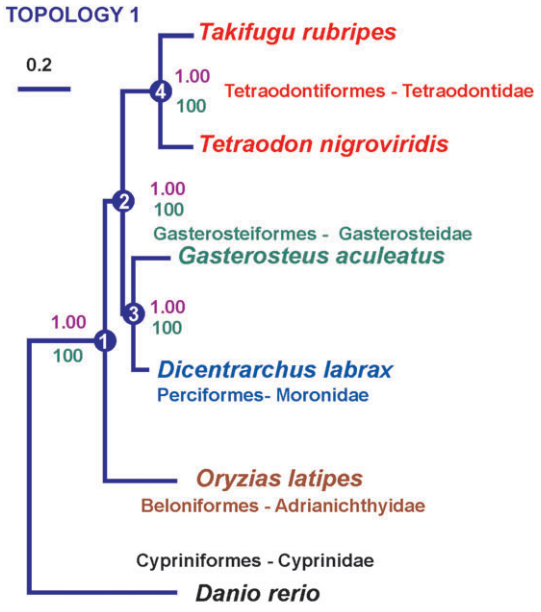
with the only exception of $MSA5_{CD+NCmla}$/$MSA7_{CD+NCmla}$, where NJ using uncorrected *P*-distances produced TREE1 with moderate/marginal statistical support (BT = 78/51%) for the group ([*Di. labrax* + *G. aculeatus*] + [*Te. nigroviridis* + *Ta. rubripes*]). All BI and ML analyses performed on Multi-LAGAN MSAs, irrespective of the complexity of the evolutionary model applied (single partition vs. 21 partitions; GTR vs. codon-based models, etc.) were concordant on a single topology (TREE2) with high statistical support for all nodes (BT $\geq$ 99%; BI = 1.00). Results obtained with the different methods applied to various MSAs are summarized in table 3.

Testing Alternative Phylogenetic Hypotheses

The observed discrepancies for the ML phylogenetic analysis were further evaluated using the AU and WSH tests. The results of these analyses performed on all possible tree topologies (105) using different data sets are provided in table 4. $MSA1_{PRmf}$ and $MSA2p1\text{-}p3_{CSmf}$ (MSAs consisting of only protein-coding sequences) were not conclusive in rejecting TREE2 and TREE1, respectively. The $MSA2p1\text{-}p2_{CSmf}$ rejected TREE1 in the AU test but not in the more conservative WSH test. $MSA3_{NCmf}$, $MSA4_{CD+NCmf}$, and $MSA8_{CSmfNCmf}$ (MSAs obtained with MAFFT) exclusively supported TREE1 with the rejection of all alternative trees. Conversely, $MSA5_{CD+NCmla}$, $MSA6_{NCmla}$, $MSA7_{CD+NCmla}$, and $MSA9_{CSmfNCmla}$ (MSAs obtained with Multi-LAGAN) exclusively supported TREE2.

Effects of Compositional Bias and Rate Heterogeneity across Sites on Phylogenetic Inference

Recoded MSAs provided phylogenetic results largely mirroring those obtained from the corresponding original



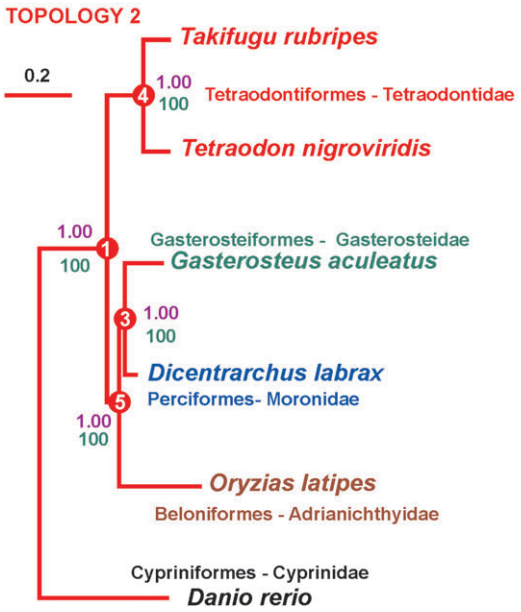


FIG. 5. TREE1. ML tree (−lnL = −983265.64859; HKY85 model) inferred from $MSA8_{CSmfNCmf}$. TREE2. ML tree (−lnL = 448535.89; GTR + *I* + *G* model) inferred from $MSA7_{CD+NCmla}$ alignment. Bar represents 0.2 substitutions per site. Green-colored numbers indicate BT expressed as percentage, whereas purple-colored numbers indicate clade posterior probabilities computed through BI analysis on the same data sets. Nodes representing topology discrepancies are numbered differently for the purpose of clarity in the main text.

MSAs with a few exceptions, which are described below. The BI and ML analyses performed on Dayhoff4-$MSA1_{PRmf}$ data set supported TREE2 instead of TREE1 but without support to node 5 (BI = 0.7; BT = 46%). The NJ trees obtained from RY-$MSA2p1\text{-}p2_{CSmf}$ provided dubious topologies because the group (*Ta. rubripes* + *Te. nigroviridis*) could not be recovered. The NJ tree derived from

Table 3. Tree Topologies Supported by Different MSAs Analyzed Using BI, ML, MP, and NJ Methods.

		BI1		BI2		ML		MP	NJ1		NJ2	
MSA Name	ALNp[c]	STT	EvMd	STT	EvMd	STT	EvMd	STT	STT	EvMd	STT	EvMd
$MSA1_{PRmf}$	MAFFT (mf)	Tp1	Dayhoff/WAG/JTT (+I) + G (+ F) (20 partitions)			Tp1	WAG/JTT + I + G	Tp1	Tp1	p-D to JTT + I + G		
$MSA2p1\text{-}p2_{CSmf}$	MAFFT (mf)	Tp2	GTR + *I* + *G* (single/20 partitions)			Tp2	HKY85 to GTR + *I* + *G*	Tp1	Tp1	p-D to MCL		
$MSA2p1\text{-}p3_{CSmf}$	MAFFT (mf)	Tp2	GTR + *I* + *G*	Tp1	GTR + *I* +*G* or CDSmP (20 partitions)	Tp2	HKY85 to (20 partitions)	Tp1	Tp1	p-D to MCL		
$MSA3_{NCmf}$	MAFFT (mf)	Tp1	GTR + *I* + *G*			Tp1	HKY85 to GTR + *I* + *G*	Tp1	Tp1	p-D to MCL		
$MSA4_{CD+NCmf}$	MAFFT (mf)	Tp1	GTR + *I* + *G*			Tp1	HKY85 to GTR + *I* + *G*	Tp1	Tp1	p-D to MCL		
$MSA5_{CD+NCmla}$	M-LAGAN (mla)	Tp2	GTR + *I* + *G*			Tp1	HKY85 to GTR + *I* + *G*	Tp1	Tp1	p-D	Tp2	JK69 to MCL
$MSA6_{NCmla}$	M-LAGAN (mla)	Tp2	GTR + *I* + *G*			Tp2	HKY85 to GTR + *I* + *G*	Tp1	Tp2	p-D to MCL		
$MSA7_{CD+NCmla}$	M-LAGAN (mla)	Tp2	GTR + *I* + *G*			Tp2	HKY85 to GTR + *I* + *G*	Tp1	Tp1	p-D	Tp2	JK69 to MCL
$MSA8_{CSmfNCmf}$	MAFFT (mf)	Tp1	CDSmP + GTR + *I* + *G* (21 partitions)			Tp1	HKY85 to GTR + *I* + *G*	Tp1	Tp1	p-D to MCL		
$MSA9_{CsmfNCmla}$	MAFFT (mf)/ M-LAGAN (mla)	Tp2	CDSmP + GTR + *I* + *G* (21 partitions)			Tp2	HKY85 to GTR + *I* + *G*	Tp1	Tp2	p-D to MCL		

NOTE.—STT, supported tree topology; Tp, topology; EvMd, evolutionary model; p-D, *P*-distance; MCL, maximum composite likelihood; CDSmP, codon model for each partition; Dayhoff, WAG (Whelan and Goldman), JTT (Jones, Taylor, and Thorton); HKY85, JK69, MCL, and GTR + *I* + *G* (Felsenstein 2004; Tamura et al. 2007).

Table 4. AU and WSH Values Calculated for All 105 Topologies.

	AU[a]	WSH[a]
ML topology 1 ($MSA3_{NCmf}$, $MSA4_{CD+NCmf}$, $MSA8_{CSmfNCmf}$)	1	1
104 alternative topologies to topology 1 ($MSA3_{NCmf}$, $MSA4_{CD+NCmf}$, $MSA8_{CSmfNCmf}$)	0	0
$MSA1_{PRmf}$, ML topology 1	0.94	1
$MSA1_{PRmf}$, ML topology 2	0.08	0.44
$MSA2p1\text{-}p3_{CSmf}$, ML topology 1	0.44	0.95
$MSA2p1\text{-}p3_{CSmf}$, ML topology 2	0.57	1
$MSA2p1\text{-}p2_{CSmf\text{-}P}$, ML topology 1	0.03	0.32
$MSA2p1\text{-}p2_{CSmf\text{-}P12}$, ML topology 2	1	0.96
ML topology 2 ($MSA5_{CD+NCmla}$, $MSA6_{NCmla}$, $MSA7_{CD+NCmla}$, $MSA9_{CSmfNCmla}$)	1	1
104 alternative topologies to topology 2 ($MSA5_{CD+NCmla}$, $MSA6_{NCmla}$, $MSA7_{CD+NCmla}$, $MSA9_{CSmfNCmla}$)	0	0

[a] *P* values.

RY-$MSA2p1\text{-}p3_{CSmf}$ using ML distance supported TREE2 instead of TREE1 but without support for node 5.

ML results from the selective removal of nucleotide positions based on gamma distribution categories are summarized in table 5. Despite the number of possible combinations, two general patterns seem to emerge. First, MSAs originally favoring TREE2 never support TREE1 after site removal, irrespective of the topology (TREE1, TREE2, SCT) used for estimating rate categories. When only the most variable categories (C8 or C7/C8) are excluded, TREE2 is always recovered. When less variable categories (C6–8, C5–8, C4–8) are also removed, topologies different from either TREE1 or TREE2 are obtained. These topologies are considered dubious because the well-established tetraodontiform group (*Ta. rubripes* + *Te. nigroviridis*) is often not recovered. The only exception was observed for the nucleotide data set containing only protein-coding genes and including all codon positions ($MSA2p1\text{-}p3_{CSmf}$, table 5). Second, progressive removal of the variable sites (C6–8) from MSAs originally supporting TREE1 recovered the alternative topology (TREE2). In addition, the tree used for estimating the rate categories appears to have a relevant effect in this case (table 5). Mixture models including up to eight substitution matrices, with or without site heterogeneity (gamma distribution with four categories), were applied to $MSA4_{CD+NCmf}$ and $MSA5_{CD+NCmla}$ in their original and RY-recoded versions as representatives of MAFFT- and Multi-LAGAN-based alignments, respectively. Phylogenetic analyses performed with BayesPhylogenies using mixture models produced the same topologies obtained in the original BI with maximum statistical support (BI = 1.00) for all nodes.

To test the effect of the shape parameter alpha on the ML tree reconstructions, different fixed alpha values were imposed. The results are summarized in table 6. MSAs including noncoding genomic sequences show low-moderate rate heterogeneity across sites (see estimated alpha values, table 6) and produce the same topology under a broad range of alpha values, although MAFFT-based alignments appear to be more sensitive to low alpha (0.25–0.5). However, ML trees that are based on concatenated protein-coding genes ($MSA1_{PRmf}$, $MSA2p1\text{-}p2_{CSmf}$, and $MSA2p1\text{-}p3_{CSmf}$) have an estimated alpha value (range 0.526–0.386) that suggests the presence of substantial rate heterogeneity. The amino acid MSA yields the same topology irrespective of the imposed alpha value. Conversely, artificially fixing the parameter alpha to a higher-than-estimated value (1.0–5.0) for protein-coding nucleotide MSAs produces a different topology (TREE1) compared with the best tree (TREE2) under the estimated alpha (table 6).

Discussion

The ultimate goal of this study was to assess the potential and limitations of phylogenomics, which can be summarized with two main questions. First, what are the critical issues and the benefits of identifying and sequencing a large contiguous genomic region for phylogenetic analysis of distantly related species? Second, phylogenomic studies generally rely on data sets of concatenated protein-coding genes; how does this approach compare to the analysis of a genomic region that contains protein-coding genes as well as a large fraction of noncoding sequences?

The first question concerns the potential problem of genomic rearrangements within the target region, which should be preliminarily evaluated, especially when a distant outgroup is used. However, if this problem can be solved, the use of contiguous genomic regions might provide information on sequence orthology and additional phylogenetic signals from noncoding regions. The limited number of available fish genomic sequences imposed the use of *Da. rerio* as outgroup. This species diverged approximately 280 Ma from the lineage Euteleostei, which includes all ingroup taxa (Azuma et al. 2008). During this long separation, the *Da. rerio* genome underwent several rearrangements, which did not allow a complete use of the target region. To analyze only the genomic regions that are unique to each species, it was necessary to discard genomic segments spanning approximately 700,000 bp (with reference to the *Da. labrax* sequenced region). As a consequence, the final MSAs were obtained starting from different segments totaling approximately 500,000 bp. After filtering nucleotide positions with Gblocks, the final size of the aligned noncoding regions ranged between 74,711 and 151,068 bp, plus 8,136–8,592 bp for five syntenic genes. Although the

Table 5. Topologies, Inferred with PhyML, Obtained from MSAs Where the More Heterogeneous Positions Were Removed from the Alignment.

		Topology Favored by MSA without Cx-Cy					
Data Set	PzTree	C4–8	C5–8	C6–8	C7–8	C8	bTree
$MSA1_{PRmf}$	TREE1	*	*	TREE2	TREE1	TREE1	TREE1
	TREE2	*	*	*	TREE2	TREE2	
	SCT	*	*	*	TREE2	TREE2	
$MSA2p1\text{-}p2_{CSmf}$	TREE1	*	*	*	*	TREE2	TREE2
	TREE2	*	*	*	*	TREE2	
	SCT	*	*	*	*	TREE2	
$MSA2p1\text{-}p3_{CSmf}$	TREE1	*	*	*	TREE1	TREE1	TREE2
	TREE2	*	*	TREE2	TREE2	TREE2	
	CT	*	*	*	TREE2	TREE1	
$MSA3_{NCmf}$	TREE1	TREE2	TREE2	TREE1	TREE1	TREE1	TREE1
	TREE2	TREE2	TREE2	TREE2	TREE2	TREE2	
	SCT	TREE2	TREE2	TREE2	TREE1	TREE1	
$MSA4_{CD+NCmf}$	TREE1	*	TREE2	TREE1	TREE1	TREE1	TREE1
	TREE2	TREE2	TREE2	TREE2	TREE2	TREE1	
	SCT	TREE2	*	TREE2	TREE1	TREE1	
$MSA5_{CD+NCmla}$	TREE1	*	*	*	TREE2	TREE2	TREE2
	TREE2	*	*	*	TREE2	TREE2	
	SCT	*	*	*	TREE2	TREE2	
$MSA6_{NCmla}$	TREE1	*	*	*	TREE2	TREE2	TREE2
	TREE2	*	*	*	TREE2	TREE2	
	SCT	*	*	TREE2	TREE2	TREE2	
$MSA7_{CD+NCmla}$	TREE1	*	*	*	TREE2	TREE2	TREE2
	TREE2	*	*	*	*	TREE2	
	SCT	*	*	*	TREE2	TREE2	
$MSA8_{CSmfNCmf}$	TREE1	*	TREE2	TREE1	TREE1	TREE1	TREE1
	TREE2	*	TREE2	TREE2	TREE2	TREE2	
	SCT	*	TREE2	TREE2	TREE1	TREE1	
$MSA9_{CSmfNCmla}$	TREE1	*	*	*	TREE2	TREE2	TREE2
	TREE2	*	*	*	TREE2	TREE2	
	SCT	*	*	TREE2	TREE2	TREE2	

NOTE.—PzTree, tree used as reference to calculate gamma C categories with TREE-PUZZLE program; Cx/y, gamma rate heterogeneity category calculated with TREE-PUZZLE; bTree, ML best tree supported by original MSA with no positions removed from the alignment; SCT, Strict Consensus Tree of TREE1 + TREE2; *, topology different from T1, T2 mostly favoring the group (*Oryzias latipes* + [*Takifugu rubripes* + *Tetraodon nigroviridis*]), or implying the disruption of the clade (*Ta. rubripes* + *Te. nigroviridis*). TREE1 and TREE2 (fig. 5).

Table 6. Variation of Shape Parameter Alpha and Its Effect on Tree Topology.

	Tree Topology Obtained from Fixed α value						ML Result	
Data Set	0.050	0.250	0.500	1.000	2.500	5.000	bTree	est-α
$MSA1_{PRmf}$	TREE1	TREE1	TREE1	TREE1	TREE1	TREE1	TREE1	0.513
$MSA2p1\text{-}p2_{CSmf}$	TREE2	TREE2	TREE2	TREE1	TREE1	TREE1	TREE2	0.386
$MSA2p1\text{-}p3_{CSmf}$	TREE1	TREE2	TREE2	TREE1	TREE1	TREE1	TREE2	0.526
$MSA3_{NCmf}$	TREE1 *		TREE2	TREE1	TREE1	TREE1	TREE1	2.138
$MSA4_{CD+NCmf}$	TREE1 *		TREE1	TREE1	TREE1	TREE1	TREE1	1.937
$MSA5_{CD+NCmla}$	TREE2	TREE2	TREE2	TREE2	TREE2	TREE2	TREE2	1.114
$MSA6_{NCmla}$	TREE1 *		TREE2	TREE2	TREE2	TREE2	TREE2	1.395
$MSA7_{CD+NCmla}$	TREE2	TREE2	TREE2	TREE2	TREE2	TREE2	TREE2	1.117
$MSA8_{CSmfNCmf}$	TREE1	TREE2	TREE2	TREE1	TREE1	TREE1	TREE1	1.558
$MSA9_{CSmfNCmla}$	TREE2	TREE2	TREE2	TREE2	TREE2	TREE2	TREE2	0.963

NOTE.—bTree, best tree in ML analysis (GTR + G evolutionary model); est-α, estimated value of shape parameter α in ML analysis; * topology different from TREE1, TREE2.

genomic region that could be usefully analyzed still represents one of the largest data sets examined so far, there was a dramatic reduction in size (10- to 15-fold). Although the *Da. rerio* genome experienced significant modifications involving the studied region, the same genome fragment seems to be largely conserved across ingroup species (Supplementary fig. S2, Supplementary Material online). Electronic chromosome painting suggests that the target region is orthologous among the acanthomorph taxa examined here. This hypothesis is further confirmed by the conservation of gene content/order (gene colinearity), indicating that the target segment has likely been retained without major gene loss or rearrangements in the species under study. Synteny analysis can provide strong evidence for orthology/paralogy, in parallel to sequence similarity (e.g., Kassahn et al. 2009). Therefore, although gene conversion/recombination events involving paralogous copies on different chromosomes cannot be completely excluded, a phylogenomic analysis of a contiguous genomic region allows accounting of other sources of unrecognized paralogy. As mentioned above, multiple nuclear gene phylogenies of vertebrates, and particularly of teleosts, might be affected by the incorrect use of paralogous rather than orthologous gene copies. In the teleost fish, three rounds of WGD have occurred. After each WGD, duplicated copies present in two different species may follow several evolutionary pathways. Both copies can be retained, for example, as a consequence of the evolution of novel functions or through subfunctionalization, or one copy of orthologous genes can be lost through pseudogenization or gene deletion in both species. However, it might be possible that one paralogous copy is lost in each species, an event termed reciprocal gene loss (RGL). In a recent study (Kassahn et al. 2009) using five fish model species (*Da. rerio*, *Te. nigroviridis*, *Ta. rubripes*, *O. latipes*, *G. aculeatus*) and focusing on 754 gene families, it was estimated that for 154 (20%) of these families, RGL has occurred in at least one evolutionary lineage. Even with this conservative estimate obtained from a small number of taxa, RGL appears to be not negligible and might represent the most important source of error when defining orthology/paralogy relationships for multiple gene phylogenies, where genes are sampled randomly across the genome (e.g., Blair and Hedges 2005; Philippe, Lartillot, et al. 2005; Rokas et al. 2005; Bourlat et al. 2006; Delsuc et al. 2006; Savard et al. 2006; Dunn et al. 2008; Ruiz-Trillo et al. 2008). At present, sequencing a large contiguous genomic region is technically more challenging; yet, novel targeted enrichment methods and next-generation sequencing technologies will likely reduce the cost and difficulty of sequencing large, contiguous genomic regions, making the approach described in the present study more feasible. An additional critical point might be the fact that teleost genomes exhibit high plasticity in terms of chromosomal duplications, rearrangements, and fusions (Kasahara et al. 2007). Sequencing and aligning a large genomic region yielded a relatively large data set (ranging from 74,711 bp for $MSA6_{NCmla}$ to 163,775 bp for $MSA4_{CD+NCmf}$) for five ingroup species, even

enforcing stringent alignment parameters and dealing with a "problematic" outgroup. Because whole-genome sequencing projects are ongoing for several other fish species is (e.g., Nile tilapia, Atlantic salmon, Atlantic cod), a phylogenomic approach based on large, contiguous genomic regions will soon be available to study one of the largest vertebrate groups, the teleost fish. The region of the genome analyzed here is syntenic over a broad taxonomic range and represents a good reference data set to provide a robust phylogenetic framework for such a comparative approach. However, it remains to be evaluated whether the current tools will be able to cope with a broader taxonomic sampling, which is required even for a limited representation of the different teleost lineages, and to determine how this will impact on the final size of the data set.

Regarding the second question, the present study provides the opportunity to test two different types of data, concatenated protein-coding genes and contiguous genomic sequences, in the same taxa and under a similar broad array of analytical methods. As a result, both sets of data clearly indicate a close relationship between *Di. labrax* and *G. aculeatus* (Moronidae + Gasterosteidae), whereas both yield conflicting evidence for deeper nodes in the tree. Only a limited number of studies (Dettai and Lecointre 2005; Smith and Craig 2007; Li et al. 2009) have investigated the phylogenetic position of Moronidae (here represented by *Di. labrax*) together with Tetraodontiformes (*Ta. rubripes* and *Te. nigroviridis* in our sample), and Gasterosteidae (*G. aculeatus* here), either with limited support for different topologies, [[Moronidae, Tetraodontiformes], Gasterosteidae] in Smith and Craig (2007) and [[Moronidae, Gasterosteidae], Tetraodontiformes] in Li et al. (2009), or with no resolution at all [Moronidae, Tetraodontiformes, Gasterosteidae] in Dettai and Lecointre (2005). Therefore, using a larger data set, either as concatenated protein-coding genes or as a contiguous genomic region, seems sufficient to unambiguously solve this phylogenetic issue. In this case, (data set) size does matter.

Whereas the position of *Di. labrax* appears unambiguous across data sets, evolutionary models, and phylogenetic methods, the placement of *O. latipes* relative to (*Ta. rubripes* + *Te. nigroviridis*) and (*G. aculeatus* + *Di. labrax*) shows irreconcilable evidence with strong support found for two alternative topologies (TREE1 and TREE2; table 4). Previous phylogenetic studies based on either complete mitochondrial sequences (Miya et al. 2001, 2003, 2005; Yamanoue et al. 2006; Azuma et al. 2008; Kawahara et al. 2008; Setiamarga et al. 2008) or nuclear genes (Chen et al. 2003; Smith and Wheeler 2004; Dettai and Lecointre 2005) always recovered a clade including *G. aculeatus* and *Ta. rubripes* and/or *Te. nigroviridis* (Tetraodontidae), with *O. latipes* being most distantly related, that is, reflecting TREE1 from this study. No study has so far reported a closer relationship between *G. aculeatus* and *O. latipes*, with the exclusion of *Ta. rubripes* and/or *Te. nigroviridis* (TREE2). Yet, TREE1 cannot be reliably considered the correct tree because supporting evidence is quite limited in previous reports. When good, or even complete, support is obtained for alternative topologies depending on the analyzed data set and/or the implemented phylogenetic method, tree reconstruction artifacts should be suspected. Several factors might favor systematic errors in tree inference, such as across-site rate variation, heterotachy, site-interdependent evolution, compositional heterogeneity, and site-heterogeneous nucleotide/amino acid replacement (Rodriguez-Ezpeleta et al. 2007 and references therein). If long-separated lineages and/or fast-evolving sequences are analyzed, these factors combine with the problem of multiple substitutions at the same site (mutational saturation) and long-branch attraction (LBA), increasing the level of "non-phylogenetic" signal (Rodriguez-Ezpeleta et al. 2007). When "true" phylogenetic signal is inherently low because of rapid lineage sorting, as is likely the case for acanthomorph taxa, all these combined factors might lead to artificial, yet highly supported, topologies.

Several different strategies have been proposed to partially correct for the systematic errors listed above. Not all these strategies could be applied in the present study. Modifying taxon sampling was not possible because of the small number of analyzed species, whereas discarding part of the protein-coding genes, fast-evolving sequences as in Nishihara et al. (2007), would have led to a concatenated data set of limited size. Other approaches, that is, RY, Dayhoff recoding, or exclusion of third-codon positions or highly variable sites, have been tested whenever possible. Additional exploratory analyses such as variation of parameter alpha, use of mixture models, and data partitioning have been implemented as well. As partially different methods have been used on either concatenated protein-coding genes or contiguous genomic sequences and because of the inherent difference between these two types of data sets, the corresponding results will be discussed separately. In the case of protein-coding genes, four major points can be made. First, distance-based and parsimony methods (NJ and MP) always agree on TREE1. Second, probabilistic methods (ML and BI) favor TREE1 when amino acid sequences are analyzed and TREE2 when nucleotides are considered, either as single positions or under a codon model. Third, neither TREE1 nor TREE2 can be significantly rejected in ML tests for alternative topologies (table 4), with the only exception of TREE2 when excluding third-codon positions (MSA2p1-p2$_{CSmf}$). Fourth, the implementation of different approaches to correct potential systematic errors suggests a substantial robustness of TREE2 compared with TREE1. Recoding amino acid positions (Dayhoff4) or nucleotide positions (RY) appears to reduce the mutational saturation (fig. 4 and after recoding, nucleotide data sets still support TREE2, whereas recoded amino acids favor TREE2, albeit without support). Data set partitioning into 20 independent data sets has little effect on tree topology, whereas selective removal of highly variable positions tends to shift tree topology from TREE1 to TREE2, especially if rate categories are estimated on alternative trees. A possible interpretation of these results is that TREE2 is the correct topology, though mutational saturation and compositional

heterogeneity at the amino acid level affect phylogenetic inference, especially on the protein data set. Conflicting evidence between protein- and nucleotide-based tree reconstructions has been reported to occur (e.g., Baldauf et al. 2000; Pollard et al. 2006). Recoding amino acid positions and selective removal of fast-evolving sites appears to reconcile the conflict, although with poor statistical support. An additional complication might be the possible occurrence of an LBA artifact in TREE1 between the long branches leading, respectively, to the outgroups *Da. rerio* and *O. latipes*. Indeed, NJ and MP, two methods that are well known to be highly sensitive to LBA, always favor TREE1. However, under any possible interpretation, it seems clear that a data set of 20 concatenated protein-coding genes, despite its relatively large size (>35,000 bp) does not allow us to chose between the two alternative topologies (TREE1 and TREE2) that depict of acanthomorph relationships. Contiguous genomic sequences provide MSAs that are twice to several times larger than the protein-coding type examined in the present study. Quite remarkably, such an increase in size makes the conflict between alternative topologies appear even sharper. The most interesting result here is that genomic MSAs produced with either MAFFT or Multi-LAGAN supported mutually exclusive tree topologies under ML, BI, and NJ. Furthermore, statistical support for either TREE1 or TREE2 is complete (e.g., table 4). As phylogenetic methods and taxa are the same, this means that MSAs produced with one of these two programs are largely affected by systematic errors and convey sufficient "non-phylogenetic signal" to completely obscure true phylogenetic information. It should be noted here that joining coding regions with noncoding regions under different combinations ($MSA4_{CD+NCmf}$, $MSA5_{CD+NCmla}$, $MSA7_{CD+NCmla}$—$MSA9_{CSmfNCmla}$) does not produce detectable differences compared with the corresponding noncoding-only MSAs ($MSA3_{NCmf}$, $MSA6_{NCmla}$). Although this is expected for $MSA6_{NCmla}$, which yields the same topology (TREE2) as the nucleotide concatenated coding sequences (MSA2p1-$p3_{CSmf}$ and MSA2p1-$p2_{CSmf}$), in the case of $MSA3_{NCmf}$ it is likely that the much larger size (>150,000 bp) of the noncoding region overwhelms the signal from the coding sequences. As observed for the concatenated protein-coding genes, TREE2 appears more robust upon selective removal of fast-evolving sites or variation of parameter alpha (tables 5 and 6). RY recoding has no effect on the best topology; yet, it substantially reduces the mutational saturation in MSAs supporting TREE2 ($MSA6_{NCmla}$, $MSA7_{CD+NCmla}$, $MSA9_{CSmfNCmla}$), which already have lower saturation compared with MSAs obtained using MAFFT (fig. 4). In the latter case, the effect of recoding is rather minimal. A possible hypothesis to explain the obtained results is that, similar to what was observed for protein-coding MSAs, systematic errors combined with LBA might lead to recovering TREE1, which is not the correct topology. An LBA artifact is suggested by the fact that MP analyses, which are particularly prone to LBA (Felsenstein 1978), always supported TREE1, where the two long branches leading, respectively, to *Da. rerio* and *O. latipes* depart from the base of the tree. Assuming this hypothesis is correct, it remains to be found which systematic error might affect MSAs obtained with MAFFT. Because the largest part of MSAs is represented by noncoding sequences, support for positional homology could not be obtained using information derived from gene/protein structure/consensus sequences. Therefore, violation of positional homology might be partially responsible for the observed results. However, precisely to avoid, or at least to minimize this problem, very stringent settings in Gblocks (e.g., no gaps allowed) were selected for automated filtering of all MSAs. Although it cannot be excluded, it seems unlikely that systematic violation of positional homology largely affects the MSAs analyzed here. It appears more plausible that, even without violating the principle of positional homology, MAFFT tends to include a large fraction of highly variable sites/regions in the alignment, whereas Multi-LAGAN might be more conservative, as it considers only regions above a certain threshold of sequence similarity (70% under default settings). In fact, the alignment produced with MAFFT on noncoding sequences is twice the size of the corresponding one obtained with Multi-LAGAN (table 2). If this hypothesis is correct, mutational saturation and possibly compositional bias, which has been found to be associated with highly saturated positions (e.g., Rodriguez-Ezpeleta et al. 2007) could affect MSAs obtained with MAFFT to a much greater extent. In turn, this would account for the higher mutational saturation observed in $MSA3_{NCmf}$, $MSA4_{CD+NCmf}$, and $MSA8_{CSmfNCmf}$, and the minimal improvement after RY recoding (fig. 4). This approach is based on the well-known substitution bias toward transitions, which causes transversions to accumulate more slowly and therefore to reach saturation lately. However, at fast-evolving sites, transversions might also reach saturation, making RY recoding less effective. Likewise, as noted above, either assuming high-rate heterogeneity across sites and imposing a low value of alpha (0.5; table 6), or excluding highly variable sites (table 5) tends to favor TREE2, even with MAFFT-MSAs.

The accuracy of multiple alignments is a long-standing, albeit somehow neglected, problem in phylogenetic analysis. Novel algorithms for sequence alignment are constantly developed; yet, several challenges remain, particularly in the production of robust MSAs for phylogenetic purposes (see Rosenberg 2009 and references therein). In particular, methods for multiple alignment of large orthologous genomic regions or whole chromosomes are still in their infancy (Kumar and Filipski 2007; Rosenberg 2009). Here, two rather different approaches were evaluated. MAFFT is one of the best performing programs in single gene multiple alignments, which ensures reasonable speed with good accuracy (Nuin et al. 2006; Wilm et al. 2006; Carroll et al. 2007). It uses a simple progressive method similar to ClustalW (Thompson et al. 1994), which builds a guide tree that is constructed based on a pairwise distance matrix and then uses the tree to iteratively improve the alignment, but with several technical modifications that allow increased speed and

accuracy. Multi-LAGAN, one of the best programs for genomic alignments available to date (Kumar and Filipski 2007), relies on short anchoring sequences to reduce the computational complexity of aligning large genomic sequences to smaller distinct fragments (Brudno et al. 2003). A guide tree is used, but it is has to be specified by the user. With respect to this point, however, alternative guide trees with Multi-LAGAN did not produce different alignments in terms of phylogenetic results (data not shown). It might be possible that, in addition to, or in combination with a less conservative approach compared with Multi-LAGAN, MAFFT is more sensitive to the choice of the guide tree.

In conclusion, size does matter because complete support for the best topology is achieved when increasing the number of aligned positions. The relative importance of taxon sampling versus gene sampling, particularly in higher phylogenetic relationships, is still an open issue (e.g., Hillis 1998; Heat et al. 2008). Theoretical and experimental studies suggest that the accuracy of phylogenetic inference is better ensured by an extensive sequence sampling rather than increasing the taxon sampling for a limited number of genes (Mitchell et al. 2000; Rosenberg and Kumar 2001, 2003; Rokas and Carroll 2005). However, as already pointed out in previous studies (Philippe, Delsuc, et al. 2005; Jeffroy et al. 2006; Nishihara et al. 2007; Rodriguez-Ezpeleta et al. 2007), tree reconstruction artifacts might find greater support when large data sets are analyzed and sparse taxon sampling may have a major impact on tree topology (e.g., Heat et al. 2008). Thus, the evolutionary relationships recovered in the present study require further corroboration based on a better taxonomic representativeness of acanthomorph fishes as well as outgroups. Indeed, the large size of the analyzed data sets and the forcedly limited taxon sampling could potentially have introduced systematic errors and unrecognized biases into our phylogenomic results. Therefore, following the Latin maxim *in medio stat virtus* (virtue lies in the center), increasing alignment size should be well balanced with greater attention to potential sources of systematic errors, such as model violations, mutational saturation, and LBA, because the effect of systematic biases likely increases with alignment size. In light of the results obtained here, current substitution models appear to still be inadequate, whereas the accuracy of methods for multiple alignment needs to be substantially improved to deal with large regions of noncoding sequences.

Supplementary Material

Supplementary figures S1 and S2 and table S1 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We express our most sincere thanks to three anonymous reviewers who provided constructive criticism on earlier versions of the manuscript. We are particularly indebted to the associate editor Hervé Philippe who provided very useful suggestions and comments that helped us to improve the final version of this work. The authors acknowledge funding by the European Commission of the European Union through the Network of Excellence Marine Genomics Europe (contract GOCE-CT-2004-505403).

References

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

Allen JE, Salzberg SL. 2005. JIGSAW: integration of multiple sources of evidence for gene prediction. *Bioinformatics* 21:3596–3603.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol.* 215:403–410.

Azuma Y, Kumazawa Y, Miya M, Mabuchi K, Nishida M. 2008. Mitogenomic evaluation of the historical biogeography of cichlids toward reliable dating of teleostean divergences. *BMC Evol Biol.* 8:215.

Baldauf SL, Roger AJ, Wenk-Siefert I, Doolittle WF. 2000. A kingdom-level phylogeny of Eukaryotes based on combined protein data. *Science* 290:972–977.

Birney E, Clamp M, Durbin R. 2004. GeneWise and Genomewise. *Genome Res.* 14:988–995.

Blair JE, Hedges SB. 2005. Molecular phylogeny and divergence times of deuterostome animals. *Mol Biol Evol.* 22:2275–2284.

Bourlat SJ, Juliusdottir T, Lowe CJ, et al. (14 co-authors). 2006. Deuterostome phylogeny reveals monophyletic chordates and the new phylum Xenoturbellida. *Nature* 444:85–88.

Bradley RK, Roberts A, Smoot M, Juvekar S, Do J, Dewey C, Holmes I, Pachter L. 2009. Fast statistical alignment. *PLoS Comput Biol.* 5:e1000392.

Brudno M, Do CB, Cooper GM, Kim MF, Davydov E, Green ED, Sidow A, Batzoglou S. 2003. LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res.* 13:721–731.

Burge C, Karlin S. 1997. Prediction of complete gene structures in human genomic DNA. *J Mol Biol.* 268:78–94.

Carroll H, Beckstead W, O'Connor T, Ebbert M, Clement M, Snell Q, McClellan D. 2007. DNA reference alignment benchmarks based on tertiary structure of encoded proteins. *Bioinformatics* 23: 2648–2649.

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

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

Clark MS, Edwards YJ, Peterson D, et al. (13 co-authors). 2003. *Fugu* ESTs: new resources for transcription analysis and genome annotation. *Genome Res.* 13:2747–2753.

Darling ACE, Mau B, Blatter FR, Perna NT. 2004. Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res.* 14:1394–1403.

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

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

Dunn CW, Hejnol A, Matus DQ, et al. (18 co-authors). 2008. Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature* 452:745–749.

Felsenstein J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. *Syst Zool.* 27:401–410.
Felsenstein J. 1985. Confidence limits on phylogenies: an approach using bootstrap. *Evolution* 39:783–791.
Felsenstein J. 2004. Inferring phylogenies. Sunderland (MA): Sinauer Associates.
Field KG, Olsen GJ, Lane DJ, Giovannoni SJ, Ghiselin MT, Raff EC, Pace NR, Raff RA. 1988. Molecular phylogeny of the animal kingdom. *Science* 239:748–753.
Frazer KA, Pachter L, Poliakov A, Rubin EM, Dubchak I. 2004. VISTA: computational tools for comparative genomics. *Nucleic Acids Res.* 32:W273–W279.
Gribaldo S, Cammarano P. 1998. The root of the universal tree of life inferred from anciently duplicated genes encoding components of the protein-targeting machinery. *J Mol Evol.* 47:508–516.
Guindon S, Delsuc F, Dufayard JF, Gascuel O. 2009. Estimating maximum likelihood phylogenies with PhyML. *Methods Mol Biol.* 537:113–137.
Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol.* 52:696–704.
Heat TC, Zwickl DJ, Kim J, Hillis DM. 2008. Taxon sampling affects inferences of macroevolutionary processes from phylogenetic trees. *Syst Biol.* 57:160–166.
Hillis DM. 1998. Taxonomic sampling, phylogenetic accuracy, and investigator bias. *Syst Biol.* 47:3–8.
Hulsen T, Huynen MA, de Vlieg J, Groenen PMA. 2006. Benchmarking ortholog identification methods using functional genomics data. *Genome Biol.* 7:R31.
Iwabe N, Kuma K, Hasegawa M, Osawa S, Miyata T. 1989. Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. *Proc Natl Acad Sci U S A.* 86:9355–9359.
Jaillon O, Aury JM, Brunet F, et al. (61 co-authors). 2004. Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. *Nature* 431:946–957.
Jeffroy O, Brinkmann H, Delsuc F, Philippe H. 2006. Phylogenomics: the beginning of incongruence? *Trends Genet.* 22:225–231.
Jobb G, von Haeseler A, Strimmer K. 2004. TREEFINDER: a powerful graphical analysis environment for molecular phylogenetics. *BMC Evol Biol.* 4:18.
Johnson GD, Patterson C. 1993. Percomorph phylogeny: a survey of acanthomorphs and a new proposal. *Bull Mar Sci.* 52:554–626.
Kasahara M, Naruse K, Sasaki S, et al. (38 co-authors). 2007. The medaka draft genome and insights into vertebrate genome evolution. *Nature* 447:714–719.
Kassahn KS, Dang VT, Wilkins SJ, Perkins AC, Ragan MA. 2009. Evolution of gene function and regulatory control after whole-genome duplication: comparative analyses in vertebrates. *Genome Res.* 19:1404–1418.
Katoh K, Kuma K, Toh H, Miyata T. 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.* 33:511–518.
Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transformation. *Nucleic Acids Res.* 30:3059–3066.
Kawahara R, Miya M, Mabuchi K, Lavoué S, Inoue JG, Satoh TP, Kawaguchi A, Nishida M. 2008. Interrelationships of the 11 gasterosteiform families (sticklebacks, pipefishes, and their relatives): a new perspective based on whole mitogenome sequences from 75 higher teleosts. *Mol Phylogenet Evol.* 46:224–236.
Kent WJ. 2002. BLAT—the BLAST-like alignment tool. *Genome Res.* 12:656–664.
Kollman JM, Doolittle RF. 2000. Determining the relative rates of change for prokaryotic and eukaryotic proteins with anciently duplicated paralogs. *J Mol Evol.* 51:173–181.
Koski LB, Golding GB. 2001. The closest BLAST hit is often not the nearest neighbor. *J Mol Evol.* 52:540–542.
Kuhl H, Beck A, Wozniak G, Canario AV, Volckaert FA, Reinhardt R. 2010. The European sea bass *Dicentrarchus labrax* genome puzzle: comparative BAC-mapping and low coverage shotgun sequencing. *BMC Genomics*. 11:68.
Kumar S, Filipski A. 2007. Multiple sequence alignment: in pursuit of homologous DNA positions. *Genome Res.* 17:127–135.
Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL. 2004. Versatile and open software for comparing large genomes. *Genome Biol.* 5:R12.
Kuzniar A, van Ham RCHJ, Pongor S, Leunissen JAM. 2008. The quest for orthologs: finding the corresponding gene across genomes. *Trends Genet.* 24:539–551.
Lartillot N, Lepage T, Blanquart S. 2009. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics* 25:2286–2288.
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.
Li B, Dettai A, Cruaud C, Couloux A, Desoutter-Meniger M, Lecointre G. 2009. RNF213, a new nuclear marker for acanthomorph phylogeny. *Mol Phylogenet Evol.* 50:345–363.
Mabuchi K, Miya M, Azuma Y, Nishida M. 2007. Independent evolution of the specialized pharyngeal jaw apparatus in cichlid and labrid fishes. *BMC Evol Biol.* 7:10.
Margulies EH, Chen CW, Green ED. 2006. Differences between pairwise and multi-sequence alignment methods affect vertebrate genome comparisons. *Trends Genet.* 22:187–193.
Mitchell A, Mitter C, Regier JC. 2000. More taxa or more characters revisited: combining data from nuclear protein-encoding genes for phylogenetic analyses of Noctuoidea (Insecta: Lepidoptera). *Syst Biol.* 49:202–224.
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, Satoh TP, Nishida M. 2005. The phylogenetic position of toadfishes (order Batrachoidiformes) in the higher ray-finned fish as inferred from partitioned Baysian analysis of 102 whole mitochondrial genome sequences. *Biol J Linn Soc Lond.* 85:289–306.
Miya M, Takeshima H, Endo H, et al. (12 co-authors). 2003. Major patterns of higher teleostean phylogenies: a new perspective based on 100 complete mitochondrial DNA sequences. *Mol Phylogenet Evol.* 26:121–138.
Mott R. 1997. EST_GENOME: a program to align spliced DNA sequences to unspliced genomic DNA. *Comput Applic.* 13:477–478.
Nakatani Y, Takeda H, Kohara Y, Morishita S. 2007. Reconstruction of the vertebrate ancestral genome reveals dynamic genome reorganization in early vertebrates. *Genome Res.* 17:1254–1265.
Nei M. 2005. Selectionism and neutralism in molecular evolution. *Mol Biol Evol.* 22:2318–2342.
Nelson JS. 2006. Fishes of the world, 4th ed. Hoboken (NJ): John Wiley & Sons.
Nishihara H, Okada N, Hasegawa M. 2007. Rooting the eutherian tree: the power and pitfalls of phylogenomics. *Genome Biol.* 8:R199.
Nuin PA, Wang Z, Tillier ER. 2006. The accuracy of several multiple sequence alignment programs for proteins. *BMC Bioinformatics* 7:471.
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.

Parra G, Blanco E, Guigó R. 2000. GeneID in *Drosophila*. *Genome Res.* 10:511–515.
Philippe H, Delsuc F, Brinkmann H, Lartillot N. 2005. Phylogenomics. *Annu Rev Ecol Evol Syst*. 36:541–562.
Philippe H, Lartillot N, Brinkmann H. 2005. Multigene analyses of bilaterian animals corroborate the monophyly of Ecdysozoa, Lophotrochozoa, and Protostomia. *Mol Biol Evol.* 22: 1246–1253.
Philippe H, Telford MJ. 2006. Large-scale sequencing and the new animal phylogeny. *Trends Ecol Evol.* 21:614–620.
Phillips MJ, Delsuc F, Penny D. 2004. Genome-scale phylogeny and the detection of systematic biases. *Mol Biol Evol.* 21:1455–1458.
Pollard DA, Iyer VN, Moses AM, Eisen MB. 2006. Widespread discordance of gene trees with species tree in Drosophila: evidence for incomplete lineage sorting. *PLoS Genet*. 2: e173.
Posada D, Crandall KA. 1998. Modeltest: testing the model of DNA substitution. *Bioinformatics* 14:817–818.
Rodriguez-Ezpeleta N, Brinkman H, Roure B, Lartillot N, Philippe H. 2007. Detecting and overcoming systematic errors in genome-scale phylogenies. *Syst Biol.* 56:389–399.
Rokas A, Carroll SB. 2005. More genes or more taxa? The relative contribution of gene number and taxon number to phylogenetic accuracy. *Mol Biol Evol.* 22:1337–1344.
Rokas A, Kruger D, Carroll SB. 2005. Animal evolution and the molecular signature of radiations compressed in time. *Science* 310:1933–1938.
Rokas A, Williams BL, King K, Carroll SB. 2003. Genome scale approaches to resolving incongruence in molecular phylogenies. *Nature* 425:798–804.
Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19:1572–1574.
Rosen DE. 1973. Interrelationships of higher euteleostean fishes. In: Greenwood PH, Miles RS, Patterson CP, editors. Interrelationships of fishes. New York: J Lin Soc (Zool). 53 (Suppl 1) p. 397–513.
Rosenberg MS. 2009. Sequence alignment. Methods, models, concepts and strategies. Berkeley: University of California Press.
Rosenberg MS, Kumar S. 2001. Incomplete taxon sampling is not a problem for phylogenetic inference. *Proc Natl Acad Sci U S A.* 98:10751–10756.
Rosenberg MS, Kumar S. 2003. Taxon sampling, bioinformatics, and phylogenomics. *Syst Biol.* 52:119–124.
Ruiz-Trillo I, Riutort M, Littlewood DTJ, Herniou EA, Baguñà J. 1999. Acoel flatworms: earliest extant bilaterian metazoans, not members of Platyhelminthes. *Science* 283:1919–1923.
Ruiz-Trillo I, Roger AJ, Burger G, Gray MW, Lang BF. 2008. A phylogenomic investigation into the origin of Metazoa. *Mol Biol Evol.* 25:664–672.
Savard J, Tautz D, Richards S, Weinstock GM, Gibbs RA, Werren JH, Tettelin H, Lercher MJ. 2006. Phylogenomic analysis reveals bees and wasps (Hymenoptera) at the base of the radiation of Holometabolous insects. *Genome Res.* 16:1334–1338.
Schmidt HA, Strimmer K, Vingron M, von Haeseler A. 2002. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics* 18:502–504.
Setiamarga DHE, Miya M, Yamanoue Y, Mabuchi K, Satoh TP, Inoue JG, Nishida M. 2008. Interrelationships of Atherinomorpha (medakas, flyingfishes, killifishes, silversides, and their relatives): the first evidence based on whole mitogenome sequences. *Mol Phylogenet Evol.* 49:598–605.
Shimodaira H. 2002. An approximately unbiased test of phylogenetic tree selection. *Syst Biol.* 51:492–508.
Smith WL, Craig MT. 2007. Casting the percomorph net widely: the importance of broad taxonomic sampling in the search for the placement of serranid and percid fishes. *Copeia* 2007:35–55.
Smith WL, Wheeler WC. 2004. Polyphyly of the mail-cheeked fishes (Teleostei: Scorpaeniformes): evidence from mitochondrial and nuclear sequence data. *Mol Phylogenet Evol.* 32:627–646.
Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22:2688–2690.
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.
Swofford DL. 2002. PAUP*, phylogenetic analysis using parsimony (*and other methods). Version 4.10. Sunderland (MA): Sinauer Associates.
Swofford DL, Olsen GJ, Wadell PJ, Hillis DM. 1996. Phylogenetic inference. In: Hillis DM, Moritz C, Mable BK, editors. Molecular systematics, 2nd ed. Sunderland (MA): Sinauer Associates. p. 407–514.
Tamura K, Dudley J, Nei M, Kumar S. 2007. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol Biol Evol.* 24:1596–1599.
Taylor JS, Braasch I, Frickey T, Meyer A, Van de Peer Y. 2003. Genome duplication, a trait shared by 22,000 species of ray-finned fish. *Genome Res.* 13:382–390.
Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTALW: 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.
Whitaker HA, McAndrew BJ, Taggart JB. 2006. Construction and characterization of a BAC library for the European sea bass *Dicentrarchus labrax*. *Anim Genet.* 37:526.
Wilm A, Mainz I, Steger G. 2006. An enhanced RNA alignment benchmark for sequence alignment programs. *Algorithms Mol Biol.* 1:19.
Wong KM, Suchard MA, Huelsenbeck JP. 2008. Alignment uncertainty and genomic analysis. *Science* 319:473–476.
Xia X, Xie Z. 2001. DAMBE: data analysis in molecular biology and evolution. *J Hered*. 92:371–373.
Yamanoue Y, Miya M, Inoue JC, Matsuura K, Nishida M. 2006. The mitochondrial genome of spotted green pufferfish *Tetraodon nigroviridis* (Teleostei: Tetraodontiformes) and divergence time estimation among model organisms in fishes. *Genes Genet Syst*. 81:29–39.
Yamanoue Y, Miya M, Matsuura K, Yagishita N, Mabuchi K, Sakai H, Katoh M, Nishida M. 2007. Phylogenetic position of tetraodontiform fishes within the higher teleosts: Bayesian inferences based on 44 whole mitochondrial genome sequences. *Mol Phylogenet Evol.* 45:89–101.
Zhaxybayeva O, Lapierre P, Gogarten JP. 2005. Ancient gene duplications and the root(s) of the tree of life. *Protoplasma*. 227:53–64.