MSc Bionformatics
SCHOOL OF MEDICINE
UNIVERSITY OF CRETE

DISSERTATION

‘A PHYLOGENOMIC PERSPECTIVE ON SPARIDAE (TELEOSTEI: SPARIFORMES) POSITIONING WITHIN THE TREE OF TELEOSTS: CHALLENGES AND NEW INSIGHTS’

PASCHALIS NATSIDIS

Primary Advisor: Tereza Manousaki (IMBBC, HCMR)

Thesis Commitee Members:
Costas Tsigenopoulos (IMBBC, HCMR)
Pavlos Pavlidis (ICS, FORTH)
Christoforos Nikolaou (Biology Department, UoC)

HERAKLION 2018

ΠΜΣ ΒΙΟΠΛΗΡΟΦΟΡΙΚΗ
ΙΑΤΡΙΚΗ ΣΧΟΛΗ
ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

‘φυλογονιδιωματικη προσεγγιση στην θεση των sparidae(teleostei: spariformes) μεσα στο δεντρο των τελεοστεων: δυσκολιεσ και νεα ευρηματα’

Πασχάλης Νατσίδης

Βασική επιβλέπουσα: Τερέζα Μανουσάκη (IMBBC, HCMR)

Μέλη τριμελούς επιτροπής:
Κώστας Τσιγγενόπουλος (IMBBC, HCMR)
Παύλος Παυλίδης (ICS, FORTH)
Χριστόφορος Νικολάου (Biology Department, UoC)

Ηράκλειο 2018

TABLE OF CONTENTS

ABSTRACT

Sparidae (Teleostei: Spariformes) are a family of fish constituted by approximately 150 species with high popularity and commercial value, such as porgies and seabreams. Although the phylogeny of this family has been under investigation multiple times, many controversies are present within the existing literature. Most studies have used a single or few genes to decipher the phylogenetic relationships of sparids. Here, I use a phylogenomic approach to resolve the position of the family, using five recently available Sparidae gene sets and 26 well-curated available fish proteomes. Through a vigorous phylogenomic analysis I suggest Tetraodontiformes (puffer fish, sunfish) as the most closely related group to Sparidae. This contrasts to the findings of a previous phylogenomic analysis involving the gilthead seabream genome that proposed the yellow croaker and the european seabass as sister taxa of Sparidae. By analytically comparing the methodologies applied in both cases, I show that this discordance is not caused by the use of different orthology algorithms and pipelines; on the contrary, I prove that it is caused by the increased taxon sampling of the present study, outlining the great importance of this aspect in phylogenomic analyses in general.

Keywords: Sparidae, fish phylogenomics, orthology assignment, taxon sampling

INTRODUCTION

Teleostei represent the dominant group within ray-finned fish (Actinopterygii), with more than 26,000 extant species. Their evolution has been extensively studied through past decades, using a variety of data including fossil records, morphological characters and molecular data, leading to a gradual resolution of teleost phylogeny (Betancur-R. et al., 2013 & 2017).

With the continuous emergence of new whole genome sequences, phylogenomic techniques are applied to characterise the evolutionary relationships among species. Whole-genome information can help in resolving uncertain nodes, as well as provide stronger evidence on already established relationships. Regarding fish phylogeny, several genome-wide approaches have been implemented so far. One of the first efforts to study ray-finned fish phylogenomics was conducted by (Li et al., 2007). Since then multiple studies have been published using not only gene markers but also noncoding elements such as the work of (Faircloth et al., 2013) who used UCE (ultra-conserved elements) to investigate the diversification of basal clades in ray-finned fish. Most genome papers include a phylogenomic analysis albeit with limited taxon sampling (e.g. Vij et al., 2016 and Xu et al., 2017), while the use of whole transcriptome data is also being employed to uncover phylogenetic relationships of specific taxonomic groups as well (Dai et al., 2018; Rodgers et al., 2018). With the emergence of new genomes and the possibilities of modern sequencing technologies, bigger datasets are becoming the norm. For example, a supermatrix of 1,110 genes from 22 actinopterygians was assembled to resolve controversies regarding the evolution of the Otocephalan group (Dai et al., 2018). Recently, the international project "Transcriptomes of 1,000 Fishes" (Fish-T1K) (Sun et al., 2016) published a massive phylogenomic analysis including more than 300 fish species (Hughes et al., 2018).

Sparidae (Teleostei: Spariformes), the focal family of this study, is a family of teleosts with high popularity and commercial value, constituted by approximately 150 species such as porgies and seabreams. The phylogenetic relationship of species within the family and among Sparidae and other teleost families has been tackled by multiple studies. However, most of them use satellite DNA or single gene markers with controversial findings (see Hanel & Tsigenopoulos, 2011 for a review). Studies that focused on the relationships among sparids have reached various conclusions. Firstly, a close relation between the genera *Pagrus* and *Pagellus* has been proposed based on microsatellite DNA (Garrido-Ramos et al., 1995). A few years later, de la Herran et al. (2001) presented an unrooted phylogeny of Sparidae using two microsatellite DNA families, which divided the family into 2 major lineages: one containing the blackspot seabream (*Pagellus bogaraveo*) and the genera *Sparus, Diplodus* and *Boops*, and the other with common pandora (*Pagellus erythrinus*) and the genera *Pagrus* and *Dentex.* In this tree, common pandora was placed closer to common dentex (*Dentex dentex*) rather than red porgy (*Pagrus pagrus*)*,* a relationship also proposed by a recent tree including 1229 percomorphs from 23 concatenated genes (Sanciancgo et al., 2016). In contrast to the aforementioned findings, two other studies placed common pandora together with *P. pagrus,* leaving common dentex outside; the first study included 66 Sparidae species and 18 mitochondrial loci (Chiba et al., 2009) and the second 91 Sparidae species and five loci (Santini et al., 2014). This relationship is supported also by a recent single-gene approach using mitochondrial COI samples from sparids inhabiting the Egyptian waters (Abbas et al., 2017). Thus, even though multiple studies have been conducted so far, the evolutionary relationship of major Sparidae genera remains unclear.

The relationship of Sparidae to other fish families is another field of controversy. In the tree proposed by Orrel & Carpenter (2004), the sister clade of Sparidae contained four

species from two different families (Lutjanidae & Haemulidae), however with the inclusion of only two loci from 48 species. More recent papers employed larger datasets such as a mitogenome data analysis from 75 teleosts (Kawahara et al., 2008) that placed Tetraodontiformes (puffer fish, sunfish) as the sister family of Sparidae, and another analysis using a six-loci supermatrix from 363 Mediterranean teleosts (Meynard et al., 2012), which proposed *Scarus ghobban* (Family: Labridae) as the immediate Sparidae relative. Another tree of 44 actinopterygii mitogenome sequences placed two Lethrinidae (emperor fish) species, *Lethrinus obsoletus* and *Monotaxis grandoculis,* next to two Sparidae, *Pagrus major* and *Spicara maena* (Yamanoue et al., 2007). Lethrinidae are also reported as the closest relatives of Sparidae in an investigation of Acanthomorpha (a subgroup of Teleostei) divergence times using a 10-gene dataset (Near et al., 2013), and in the 1229-percomorph tree of Sanciancgo et al (2016). A very recent and large-scale (303 fish species) phylogenomics study including four Sparidae transcriptomes (*Evynnis cardinalis, Spondyliosoma cantharus, Acanthopagrus latus* and *Acanthopagrus schlegelii*) presented a tree from 1,105 loci that recovered the spinefoot *Siganus guttatus* (family: Siganidae) as the sister taxon to Sparidae, although with low support (Hughes et al., 2018). The testing of these last hypotheses using whole-genome information is not feasible yet due to lack of high quality reference-based gene prediction of Lethrinidae or Siganidae genes.

Sparidae genetic data have been recently greatly enriched by transcriptomic studies (Tsakogiannis et al., 2018, Manousaki et al., 2014) and two whole-genome sequencing datasets, those of gilthead seabream (Pauletto et al., 2018) and Chinese black porgy (Zhang et al., 2018). In the phylogenomic analysis of Pauletto et al. using 2,032 genes from 14 species, the large yellow croaker (*Larimichthys crocea*, family: Sciaenidae) and the European seabass (*Dicentrarchus labrax*, family: Moronidae) were placed as sister groups to gilthead seabream

with high confidence. This scenario has been previously supported only by Orrell & Carpenter (2004) and Chiba (2009).

Here, I revisited the phylogenetic relationship of major Sparidae genera and other teleost fish. A comprehensive phylogenomic dataset was built using the genomic data from species spanning a wide spectrum of teleost subgroups, and five species from the family of Sparidae. The inferred phylogenetic tree addressed with high confidence the phylogenetic position of Sparidae and major teleost groups, as well as fish phylogeny in general using the most recent, well-curated, whole-genome based gene datasets.

MATERIALS AND METHODS

Sparidae data preprocessing, taxon sampling and quality assessment

The transcriptomes from brains and gonads of common dentex (*Dentex dentex*), sharpsnout seabream (*Diplodus puntazzo*), common pandora (*Pagellus erythrinus*) and red porgy (*Pagrus pagrus*) were obtained from previous studies (Tsakogiannis et al., 2018, Manousaki et al., 2014). The four Sparidae transcriptomes were processed using the EMBOSS v6.6.0.0 software "getorf" (Rice et al., 2000) with the option '-minsize 150', to recover all open reading frames (ORFs) of length >= 50 amino acids. The longest ORF was kept for each gene using a Python script. For gilthead seabream (*Sparus aurata*), the full gene set was obtained from its genome sequence publication (Pauletto et al., 2018).

The selection of non-Sparidae taxa included in the analysis was based on: i) the availability of a well-annotated predicted gene set, ii) the availability of a genome paper that describes an elaborated gene prediction pipeline, and iii) representation of a wide range of the different teleost groups. The selection process resulted to the inclusion of 26 fish gene sets, recovered mainly from NCBI (NCBI Resource Coordinators, 2018), Ensembl (Hubbard et al., 2002) and GigaDB (Sneddon et al., 2012) databases (Table 2), in addition to the five Sparidae species, forming a 31 taxa dataset (Fig. 1A). The spotted gar (*Lepisosteus oculatus*), a member of Holostei, was selected as an outgroup for the analysis. Most of the fish that came with a genome sequence but were dismissed from the analysis had either low assembly statistics, or their inclusion was redundant, since other closely related taxa were selected. Proteomes retrieved from NCBI and Ensembl databases had multiple isoforms for some genes. Those proteomes were processed to keep the longest isoform per gene using in-house scripting.

To assess the quality of the retrieved gene sets, BUSCO v3 (Simão et al., 2015) software was employed using the '-l actinopterygii' option to enable the proper lineage library for the data (Fig. 1A). This library consists of 4,584 genes that are expected to be present in at least 90% of the species in the actinopterygii lineage. So, a high representation of the BUSCO genes in each of the datasets is an indicator of quality and completeness of the gene sets. BUSCO provides statistics for genes found in complete form, fragmented or duplicated in the tested datasets.

Orthology assignment and superalignments construction

To investigate orthology relationships among the Sparidae transcriptomes and the downloaded gene sets, two different tools were employed(Fig. 1B), OrthoFinder v2.1.2 (Emms & Kelly, 2015) and PorthoMCL (Tabari & Su, 2017). The OrthoFinder algorithm solves a previously undetected gene length bias in orthogroup inference, by normalising the BLAST bit scores. The OrthoMCL algorithm (Li et al., 2003) uses Markov clustering to group (putative) orthologs and paralogs. PorthoMCL is a parallel implementation of the OrthoMCL algorithm, making genome-scale orthology assignment computationally feasible.

We discarded all ortholog groups with more than one sequence per species to avoid potential paralogies. From the resulted single-copy groups, those with representation of at least 27 of the 31 taxa were selected, so that every group contained at least 1 of the 5 Sparidae. I used Python scripting to retrieve the amino acid sequences of each orthogroup and use them for downstream analyses.

MAFFT v7 (Katoh and Standley, 2013) was employed to align the sequences of each orthogroup separately, allowing the "--auto" parameter to determine the most suitable alignment method. The alignments of OrthoFinder and PorthoMCL groups were

concatenated into two distinct superalignments using a Python script (Fig. 1C). The superalignments were then filtered with Gblocks v0.91b (Castresana, 2000) to remove poorly aligned sites, changing the parameter "Allowed Gap Positions" to "half" and leaving all other parameters at default values.

Phylogenomic analysis

Each of the two filtered superalignments, one from OrthoFinder groups and one from PorthoMCL groups, was provided as input to RAxML v8.2.9 (Stamatakis, 2014) to search for the maximum likelihood tree. The parameter "-m PROTGAMMAAUTO" was selected to automatically select the model that best fits the dataset. One hundred rapid bootstrap replicates were drawn from the input alignment during each RAxML run. Apart from bootstrap resampling, maximum likelihood was run on 100 jackknifed datasets. For that, a random 30% of the orthogroups was excluded each time from the supermatrix, keeping the rest 70% of the orthogroups. The random split was achieved using a series of bash and Python scripts. A majority rule consensus tree was built to summarize the bipartition information of the 100 jackknifed trees using the RAxML "-J MRE" option.

Bayesian Inference was performed using ExaBayes v1.4.1. (Aberer et al., 2014). Two independent chains were initiated in parallel using the "-R 2" option. The Markov chain Monte Carlo (MCMC) sampling of trees was automatically stopped after 1,000,000 generations due to convergence of the two chains, after discarding the default 25% burn-in. The sampled distributions of the parameters were inspected and the sufficiency of the effective sample sizes (ESS > 200) of all sampled parameters was confirmed with the "postProcParam" utility. Finally, a consensus tree was built from the two sets of trees using the "consense" utility of ExaBayes.

We employed RogueNaRok v1.0 (Aberer et al., 2013) to identify potential rogue taxa in the 100 bootstrap replicates of the two maximum likelihood trees. The RogueNaRok algorithm optimizes the relative bipartition information criterion (RBIC), which is defined as the sum of all support values divided by the maximum possible support in a fully bifurcating tree with the initial (i.e., before pruning any rogues) set of taxa. The algorithm prunes taxa until RBIC cannot be further improved.

CONSEL v0.20 (Shimodaira et al., 2001) was employed to compare the placement of Sparidae in my trees with the one suggested by the gilthead seabream genome paper. CONSEL calculates p-values for various statistical tests based on the per-site log likelihoods for the candidate trees given a sequence alignment. These tests include the approximate unbiased (AU) test (Shimodaira, 2002), the K-H test (Kishino & Hasegawa, 1989), the S-H (Shimodaira & Hasegawa, 1999) test, and others. The output of CONSEL allows to determine which of the candidate topologies is most likely to be the true one. The per-site log likelihoods were obtained using the RAxML option '-f G'. CONSEL analysis was applied on both OrthoFinder and PorthoMCL datasets.

We also wanted to test if the aforementioned discordence is due to the selection of the orthology assignment algorithm. To that end, I kept only the 14 species that were used by Pauletto et al., and ran OrthoFinder and PorthoMCL anew. Single-copy groups with at least 13 of the 14 species were selected and then aligned them using MAFFT. The alignments of each tool were separately concatenated into two superalignments using custom Python scripts, and then filtered with Gblocks. Two corresponding maximum likelihood trees were constructed using RAxML.

Gene tree incongruence

To check the (in)congruence of individual tree phylogenies with the recovered trees I performed gene tree analysis. Only the alignments of the single-copy groups that included sequence information for all 31 species were kept and processed them with Gblocks, to keep sites that were aligned properly. The filtered alignments were used to construct individual gene trees using RAxML with "-m PROTGAMMAAUTO" option for automatic selection of the best fitting model and 100 rapid bootstrap replicates. A majority rule consensus tree was built to summarize the bipartition information of the resulted gene trees using RAxML "-J MR" option. Also, internode certainty (IC) of each node and the extended internode certainty (ICA) were calculated, as well as the tree certainty (TC) and the extended tree certainty (TCA) values (Salichos & Rokas, 2013). IC and TC are calculated based on the most prevalent confilicting bipartition, while ICA and TCA take into account all prevalent conflicting bipartitions. Those metrics were calculated using the RAxML "-f i" option (Salichos et al., 2014) under the JTT +F+Γ4, model.

RESULTS

Sparidae data preprocessing, taxon sampling and quality assessment

The four Sparidae transcriptomes included from 98,012 to 129,012 transcripts (Table 1). After keeping the longest ORF per gene, the largest set of sequences was that of common pandora, with 89,124 genes and the smallest that of red porgy with 62,116 genes.

Regarding the other teleost species, following a careful investigation of all the available sources for fish genomes, I formed a comprehensive dataset containing 31 species (Table 2). Apart from the five Sparidae gene sets, another 23 proteomes from NCBI were collected, Ensembl and GigaDB databases, and 3 proteomes from other sources (species-specific databases, communication with paper authors). Almost half of all teleost fish with published whole-genome sequences (Ravi & Venkatesh, 2018) were included in the final dataset. Percomorphs (subdivision: Percomorphaceae) are well-represented in the dataset with 27 species, spanning 7 out of their 9 major series, as defined by Betancur-R. et al (2017). The two unincluded series, Ophidiaria and Batrachoidiaria are missing because none of their members has its genome sequence published. Note that members of Salmonidae family with whole genome sequence available (*Salmo salar, Oncorhynchus mykiss*) were not included in the final dataset because their extra whole genome duplication (Lien et al., 2016) might hamper the orthology inference algorithms. Some species were excluded due to their low genome assembly statistics, such as the scaffold N50. The inclusion of multiple closely related species was avoided, for example only one (*Boleophthalmus pectinirostris,* family: Gobiidae) out of the four available mudskipper genomes (You et al., 2014) was kept using the assembly statistics as selection criterion. Apart from the 27 percomorphs, the remaining four species of the final dataset were the Paracanthopterygii member Atlantic cod (*Gadus morhua*, order: Gadiformes), two members of the Ostariophysi superorder, the zebrafish

(*Danio rerio,* order: Cypriniformes) and the blind cavefish (*Astyanax mexicanus,* order: Characiformes) and the Holostei spotted gar (*Lepisosteus oculatus*) as an outgroup.

To assess the quality of each gene set BUSCO analysis was run. The results showed that the gene set of sharpsnout seabream has the lowest number of BUSCO library sequences with 3,347 (73%) out of the 4,584 genes (Fig. 2, Supplementary table 1). The other Sparidae proteomes scored higher BUSCO statistics, outperforming even some of the 26 datasets from online sources. The common dentex dataset contained 3,876 (84.5%) BUSCO genes, common pandora had 3,954 (86.3%), while red porgy had 3,945 (86.1%) genes. The geneset of gilthead seabream contained 3,910 (85.3%) genes, but had the fewest missing genes among the five Sparidae. As for the publicly available proteomes, the ones downloaded from the Ensembl database presented the smallest amount of missing genes (from <10-100), while datasets obtained from NCBI contained the most duplicated genes.

Orthology assignment and superalignments construction

The total number of genes from all 31 proteomes included in the orthology assignment analysis was 974,940. OrthoFinder and PorthoMCL identified 45,730 and 42,693 groups of orthologous genes, respectively (Table 3). Following filtering, 793 and 533 groups from each dataset were kept to construct the two superalignments. 56 identical orthology groups were found shared between the results of the two software tools (Fig. 3). The superalignment of OrthoFinder groups consisted of 468,718 amino acids and the one of PorthoMCL groups of 321,695. Gblocks filtering retained 231,078 (49%) and 141,608 (44%) sites, respectively.

Phylogenomic analysis

All RAxML runs selected the JTT (Jones et al., 1992) as the model of evolution that best explains the dataset, with gamma distribution on rates and empirical base frequencies (noted as PROTGAMMAJTTF). Maximum likelihood trees for both OrthoFinder (Fig. 4) and PorthoMCL (Supplementary Fig. 1) superalignments resulted in similar topologies for most species. Firstly, they agreed on the monophyly of the five Sparidae species. The common pandora and the red porgy were grouped together, with common dentex as their closest relative. The gilthead seabream and the sharpsnout seabream, were placed together in the clade that diverged first within the Sparidae lineage. All intrafamilial relationships of Sparidae were supported by a 100 bootstrap value in both OrthoFinder and PorthoMCL maximum likelihood trees.

The closest group to Sparidae was Tetraodontiformes. The green spotted puffer (*T. rubripes,* family: Tetraodontidae) and the Japanese puffer (*T. nigrovirdis*, family: Tetraodontidae) were assigned longer branch lengths than the third Tetraodontiformes member, the ocean sunfish (*M. mola*, family: Molidae). RAxML tree searches using different subsets of Tetraodontiformes and Sparidae taxa (Supplementary Fig. 2) agreed on their proposed relationship, with maximum bootstrap support at all times. The large yellow croaker and the European seabass were grouped together, as the immediately closest group to the Sparidae/Tetraodontiformes clade.

The two species that reside in the Antarctic waters, the dragonfish (*P. charcoti*, Family: Bathydraconidae) and the bullhead (*N. coriiceps*, Family: Nototheniidae) were placed in the same clade, with stickleback (*G. aculeatus*, Family: Gasterosteidae) as their closest relative. These three fish are all members of the order Perciformes (Betancur-R. et al., 2017).

The OrthoFinder tree had maximum bootstrap support values (100) assigned in all nodes of the above findings, that describe the phylogenetic relationships of the 14 Eupercaria (Eu) fish of the dataset. PorthoMCL tree recovered identical topology for the Eu, with all nodes presenting maximum bootstrap support values, except from the croaker/seabass ancestor (93).

The monophyly of each of the Carangaria (C), Anabantaria (A) and Ovalentaria (O) series was supported by both OrthoFinder and PorthoMCL maximum likelihood trees with high intra-series support values. However, the inter-series relationships of these three groups are ambiguously recovered by the two trees. OrthoFinder tree suggested the grouping of C/A cluster together with the Eu, while PorthoMCL tree placed C/A and O in the same clade, although with low support (49).

Another point of discordance between the two maximum likelihood trees was the position of the pacific bluefin tuna (*T. orientalis,* family: Scombridae), which is a member of the Pelagiaria series (Betancur-R. et al., 2017). In the OrthoFinder tree, tuna was placed next to the Eu clade, while in PorthoMCL tree it is placed outside the Eu/C/A/O cluster. Both of these placements were supported by a relatively not so high bootstrap proportion (73 and 71 respectively).

For the non-percomorph fish, the two maximum likelihood trees converged on grouping the two Ostariophysi members, the zebrafish and the blind cavefish, together. These two fish were the first ones that diverged from the rest of the teleosts, with the next divergence giving the Atlantic cod clade, followed by the mudskippe. The seahorse, of the Syngnatharia series, is also one of the lineages that was recovered identically in both trees. All nodes describing the divergences mentioned here were assigned maximum bootstrap value in both OrthoFinder and PorthoMCL maximum likelihood trees.

Two consensus trees were built to summarize the information of the 100 OrthoFinder and the 100 PorthoMCL jackknifed trees. The first tree (Supplementary Fig. 3A) presented identical topology with the main OrthoFinder RAxML tree. Support for the controversial nodes (BS < 85) of the main tree were increased to 93, except for the tongue sole split that was present in 83 out of the 100 jackknifed trees. The consensus tree of the 100 PorthoMCL jackknifed trees (Supplementary Fig. 3B) presented identical topology with the main PorthoMCL RAxML tree. The controversial nodes (BS < 85) of the main tree maintained their low support in the jackknife consensus tree as well. However, the split of common pandora and red porgy received a support value of 79 in the jackknife consensus tree, while it was recovered with 100 bootstrap support on the main PorthoMCL tree.

The two consensus trees from OrthoFinder and PorthoMCL Bayesian analyses recovered identical topologies, except for the relationships among the three Carangaria species. The OrthoFinder tree (Supplementary Fig. 4A) proposed the grouping of Asian seabass (*L. calcarifer*, family: Latidae) and greater amberjack (*S. dumerili,* family: Carangidae) group leaving the tongue sole (*C. semilaevis,* family: Cynoglossidae) outside, while the PorthoMCL tree (Supplementary Figs. 4B) grouped the tongue sole together with the greater amberjack. The tongue sole was assigned a longer branch than its two relatives. Both Bayesian trees presented posterior probabilities equal to 1.0 in all of their nodes.

To identify any possibly rogue taxa, RogueNaRok was run on the bootstrap replicates of each maximum likelihood tree search. The results did not drop any taxa as rogue, with RBIC scores calculated at 0.966 and 0.939 for OrthoFinder and PorthoMCL maximum likelihood trees, respectively. Nevertheless, I tested how the removal of some possibly ambiguous taxa affected the topology and the support values. For this analysis, the 793 orthogroups of OrthoFinder were used.

To check how the long branch of tongue sole affected the proposed phylogeny, I discarded its sequences from all OrthoFinder groups and built maximum likelihood tree anew. This tree suggested identical topology to the one with all 31 species, but with a slight increase of the bootstrap support values (Supplementary Fig. 5A).

We also examined whether the pacific bluefin tuna dataset is related to the low bootstrap support values of the tree. To that end, the OrthoFinder groups were furtherly reduced to 29 species by removing tuna sequences as well. The resulting trees proposed the same topology as the initial trees for the remaining species, but this time with all nodes at maximum support value (Supplementary Fig. 5B).

To check how my result compares to the tree suggested in the gilthead seabream genome paper, CONSEL was used. The results strongly supported the topology with Tetraodontiformes as most closely related group to Sparidae, as opposed to the topology of Pauletto et al., that suggested the croaker/seabass clade as sister to Sparidae. The p-values of all tests were equal to 1 (Table 4) for both OrthoFinder and PorthoMCL datasets. Specifically for the approximate unbiased (au) test, which is the main result of a CONSEL run, one may reject the possibility that a tree is the most likely tree among all candidates when $AU < 0.05$ at the significance level 0.05. Thus, $AU = 1.0$ provides very strong evidence for Tetraodontiformes against croaker/seabass as the closest group of Sparidae, based on this dataset.

To test if the selection of the orthology assignment algorithm is responsible for the discordance between the present study and the gilthead seabream genome paper, OrthoFinder and PorthoMCL were run using the same 14 species as they did. After filtering for single-copy groups with maximum one species missing, 2,192 and 1,366 genes were left for the two tools respectively. After concatenating them into two separate superalignments,

RAxML was employed. The two resulting trees were identical both with each other, and with the tree presented by Pauletto et al. The European seabass and the large yellow croaker were placed as sister taxa to the gilthead seabream, while the two puffer fish were recovered as immediate relatives to these three fish.

Gene tree incongruence

To assess the (in)congruence between the gene trees and the estimated species trees, individual trees were constructed for the groups of orthologs that contained sequence information from all 31 species. 135 OrthoFinder and 78 PorthoMCL groups satisfied the above criterion, and their trees were used to build a consensus tree, and to calculate internode certainty (IC and ICA) and tree certainty (TC and TCA) values , related to the corresponding species tree. The two consensus trees contained multiple multifurcating nodes (Supplementary Figs. 6A and 6B). The results of IC/TC analysis suggested low conflict in shallow nodes of the trees, i.e. at family level (Supplementary Figs. 6C and 6D) compared to deeper nodes. For example, the Sparidae monophyly was highly supported (IC=0.865 in the OrthoFinder dataset). However, high conflict was observed in deeper divergences, with even negative IC values at some ambiguous nodes such as the tuna placement. Negative IC values show that the most represented topology within the gene trees is not the one recovered in the species tree. Relative TC values were reported to be 0.295 and 0.212 for OrthoFinder and PorthoMCL datasets respectively.

DISCUSSION

Here, I analysed a comprehensive teleost phylogenomic dataset and questioned the position of Sparids within the tree of teleosts using high quality gene prediction datasets. The results suggested Tetraodontiformes as the sister group to Sparidae and grouped the analysed sparids according to their reproduction modes.

Regarding within sparids relationships, all trees that were built in the present analysis recovered a single topology (Fig. 4) for the five species used. The resulted topology agrees with previous studies (Chiba et al., 2009; Santini et al., 2014; Abbas et al., 2017). Interestingly, the species have been grouped according to their reproductive mode, i.e. red porgy and common pandora are protogynous, the protandrous gilthead seabream is grouped with the rudimentary protandrous sharpsnout seabream and the gonochoristic common dentex falls in between the two groups. In general, the members of Sparidae family exhibit a variety of reproduction methods (Mylonas et al., 2011). These findings may be linked to how these different modes of reproduction have emerged during the evolution of Sparidae lineages. For example, protogyny may have been evolved in the red porgy/common pandora ancestor, after its divergence from the rest of Sparidae. However, further investigation including more species is necessary for this hypothesis to be confirmed.

As for the relationships of Sparidae and other teleost groups, the results showed that, from the species included in the analysis, Tetraodontiformes is the closest group to Sparidae. This has been frequently reported in the literature as well (Kawahara et al., 2007 & Meynard et al., 2012). However, the very recent phylogenomics study presented in the gilthead seabream genome paper (Pauletto et al., 2018), the first thorough analysis including a Sparidae species and 14 other taxa, proposed with high confidence the yellow croaker and the European seabass as more closely related to sparids and not the Tetraodontiformes. To

understand why the two phylogenomic analyses find such controversial results, I tracked down the main differences of the present work to that of Pauletto et al. The main differences are: i) the algorithm used for identifying the orthology groups and ii) the denser taxon sampling of the present study. Regarding the first, the groups of ortholog genes in Pauletto et al. were recovered using the OMA standalone (Train et al., 2017), a software considered to have high specificity, but low sensitivity in finding the true orthologous clusters (Linard et al., 2011). To see whether the selection of orthology inference algorithm affected the resulted phylogeny, I repeated my orthology assignment employing OrthoFinder and PorthoMCL using only the 14 taxa used in Pauletto et al, and conducted the phylogenomic analysis. Interestingly, the analysis of this reduced dataset (Supplementary Fig. 7) was in total agreement with the one reported by Pauletto et al. This suggests that the discordance with my results is not due to the selection of different orthology inference algorithm, and might be explained by the more ample taxon sampling of the present study, both within Sparidae species (5 vs 1 by Pauletto et al.) and in the rest of teleost taxa (26 vs 14 by Pauletto et al.). Another hypothesis potentially explaining the discordance of the two analyses could be that in this study a third species of Tetraodontiformes was included, the ocean sunfish, that might have overcome a potential long branch attraction in Pauletto's tree. To test this hypothesis the ocean sunfish was removed from the 31-species dataset and rebuilt tree and Tetraodontiformes remained as the sister taxa (Supplementary Fig. 2B). This result remained the same even when the analysis was rerunusing only gilthead seabream from the five Sparidae (Supplementary Fig. 2A). Finally, CONSEL analysis, given my superalignment, strongly supported my suggested topology against the one of Pauletto et al. All these pieces of evidence corroborate the robustness of the results presented here and at the same time underline how critical dense taxon sampling is.

The positioning of the non-sparid teleosts in the resulted trees arose a noteworthy issue as well. The OrthoFinder tree placed tuna as sister taxon to the Eupercaria clade, while the PorthoMCL tree proposed that tuna diverged right after the seahorse divergence. Both trees assigned relatively low support on the tuna split and some other nodes close to it. When tuna sequences were removed from the dataset, all support values of the trees were increased to 100. Resolving the position of tuna within the fish phylogeny has been an object of contradiction in the existing literature. In the tree of Meynard et al. (2012), the Scombriformes order was placed very close to the Gobiiformes, the order that muskippers belong into. However, the 1410-species review of Betancur-R. et al. (2013) grouped together the orders of Scombriformes and Syngnathiformes, suggesting a closer relationship of tunas to seahorses, rather than mudskippers. This relationship was confirmed by Sanciangco et al. (2016) and Betancur-R (2017), that proposed Syngnatharia, the series of seahorses, as closest relatives of Pelagiaria, the series of tunas. In both studies though, the Syngnatharia/Pelagiaria branch was assigned a moderate support value (<89). Only very recently, the relationship between seahorses and tunas was recovered with high confidence (Hughes et al., 2018). This relationship remains to be confirmed by future studies.

Apart from the tuna positioning, most of the other findings on phylogenetic placement of the non-sparid fish are in agreement with the existing literature. Indicatively, the two Antarctic fishes (dragonfish and bullhead) and the stickleback were placed in the same clade in the present study. This is in agreement with the results presented in the dragonfish genome paper (Ahn et al., 2017) and the study of Hughes et al. (2018) as well.

Technically speaking, taxon sampling is a crucial part of a phylogenetic analysis. This has been shown by multiple studies (e.g. Zwickl & Hillis, 2002; Hedtke et al., 2006; Heath et al., 2008) and by a recent review tackling the impact of taxon sampling on phylogenetic

inference (Nabhan & Sarkar, 2012). Incongruence in molecular phylogenies can also be resolved by increasing the number of genes included in the analysis (Rokas et al., 2003). Therefore, it is necessary for any phylogenomic analysis to include as many taxa as possible, without reducing the amount and the quality of the loci used to build the tree. In the present work, I have a much denser taxon sampling compared to Pauletto et al. but reduced number of genes, which is normal when taxa inclusion increases. However, Pauletto et al phylogeny was recovered even with the dataset of the present study, when keeping only the species used in that study. Thus, in this case it seems that although they used many more genes the taxon sampling was the crucial factor.

Another important factor in phylogenomic studies is the selection of the orthology inference algorithm. Various issues for most tools, regarding computational time and accuracy have been described and reviewed by Nichio et al. (2017); however, recently developed promising tools improve greatly the orthology inference in both of the above aspects. Here, I chose to employ two recently developed graph-based orthology inference software tools, OrthoFinder and PorthoMCL (parallel implementation of OrthoMCL). OrthoFinder has improved accuracy compared to other algorithms, while PorthoMCL is the fastest option for genome-scale analyses. They both use BLAST results to infer orthology, but OrthoFinder steps include a normalization of the BLAST bit scores according to the length of the genes Emms & Kelly, 2015). This normalization solves a previously unadressed bias that favoured longer genes, as they were assigned greater bit scores. In the present study, this led to an increased number of orthogroups returned by OrthoFinder both initially (45,730 vs 42,693 in PorthoMCL) and after filtering for 1-1 groups with at least 27 taxa (793 vs 533 in PorthoMCL). The average length of these groups, however, was slightly smaller in the OrthoFinder groups (591.06 sites/group vs 603.56 in PorthoMCL). Moreover, the results of

the jackknifed trees analysis suggested that OrthoFinder groups were more robust, and a tree with 70% of them at random will most likely recover the topology of the whole dataset. On the other hand, a random 70% of the PorthoMCL groups was not always enough to fully recover the relationship between common pandora and red porgy, as well as some of the deeper splits.

Gene tree analysis was unable to recover the topologies that resulted from the supermatrix approach, suggesting that phylogenetic signal in gene trees is inadequate. The consensus trees for both OrthoFinder and PorthoMCL 31-species gene trees presented mostly polytomies, while IC/ICA values were low, even negative, in some deeper nodes. The amount of discordance among the gene trees, as well as the conflict between the gene trees and the species tree recovered via supermatrix approach indicates that this type of analysis is not suitable for the present dataset. This is an innate property in cases where gene trees are used to infer species phylogeny (Degnan et al., 2009).

CONCLUSIONS

We investigated the phylogeny of the family Sparidae (Teleostei: Spariformes) by incorporating five recently available Sparidae gene sets into a comprehensive teleost phylogenomic dataset together with 26 more species. My findings suggested that, from all teleosts with high quality reference-based gene prediction, Tetraodontiformes were the most closely related group to Sparidae. This finding has been rigorously tested using jackknife resampling, gene tree incongruence and tree selection tests. Comparison with previous phylogenomic studies has revealed a consistent incongruence among different phylogenomic datasets turning this question to a paradigm of phylogenomics highlighting the importance of

taxon sampling as a critical factor compared to other aspects of the pipeline such as the selection of orthology assignment algorithm.

REFERENCES

Abbas EM, et al. 2017. Phylogeny and DNA Barcoding of the Family Sparidae Inferred from Mitochondrial DNA of the Egyptian Waters. J. Fish. Aquat. Sci. 12:73-81

Aberer AJ, Krompass D, Stamatakis A. 2013. Pruning rogue taxa improves phylogenetic accuracy: an efficient algorithm and webservice. Syst. Biol. 62:162-6

Aberer AJ, Kobert K, Stamatakis A. 2014. ExaBayes: massively parallel Bayesian tree inference for the whole-genome era. Mol. Biol. Evol. 31(10):2553-6

Ahn DH, et al. 2017. Draft genome of the Antarctic dragonfish, *Parachaenichthys charcoti*. Gigascience 6(8):1-6

Ao J, et al. 2015. Genome sequencing of the perciform fish *Larimichthys crocea* provides insights into molecular and genetic mechanisms of stress adaptation. PLoS Genet. 11(4):e1005118

Aparicio S, et al. 2002. Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. Science 297:1301-10

Austin CM, et al. 2017. De novo genome assembly and annotation of Australia's largest freshwater fish, the Murray cod (*Macchullochella peelii*), from Illumina and Nanopore sequencing read. Gigascience 6(8):1-6

Betancur-R R, et al. 2013. The tree of life and a new classification of bony fishes. PLoS Curr 5:ecurrents.tol.53ba26640df0ccaee75bb165c8c26288

Betancur-R R, et al. 2017. "Phylogenetic classification of bony fishes. BMC Evol. Biol. 17(1):162

Braasch I, et al. 2016. The spotted gar genome illustrates vertebrate evolution and facilitates human-teleost comparisons. Nat. Genet. 48(4):427-37

Brawand D, et al. 2014. The genomic substrate for adaptive radiation in African cichlid fish. Nature 513(7518):375-81

Castresana, J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol. Biol. Evol. 17:540-52

Chen S, et al. 2014. Whole-genome sequence of a flatfish provides insights into ZW sex chromosome evolution and adaptation to a benthic lifestyle. Nat. Genet. 46(3):253-60

Chiba SN, Iwatsuki Y, Yoshino T, Hanzawa N. 2009. Comprehensive phylogeny of the family Sparidae (Perciformes: Teleostei) inferred from mitochondrial gene analyses. Genes Genet. Syst. 84(2):153-70

Dai W, et al. 2018. Phylogenomic Perspective on the Relationships and Evolutionary History of the Major Otocephalan Lineages. Sci. Rep. 8(1):205

de la Herrán R, Rejón CR, Rejón MR, Garrido-Ramos MA. 2001. The molecular phylogeny of the Sparidae (Pisces, Perciformes) based on two satellite DNA families. Heredity (Edinb) 87(Pt 6):691-7

Degnan JH, DeGiorgio M, Bryant D, Rosenberg NA. 2009. Properties of consensus methods for inferring species trees from gene trees. Syst. Biol. 58(1); 35-54

Emms D, Kelly S. 2015. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol. 16:157

Faircloth BC, Sorenson L, Santini F, Alfaro ME. 2013. A Phylogenomic Perspective on the Radiation of Ray-Finned Fishes Based upon Targeted Sequencing of Ultraconserved Elements (UCEs). PLoS One 8(6):e65923

Garrido-Ramos MA, et al. 1995. Phylogenetic relationships of the Sparidae family (Pisces, Perciformes) inferred from satellite-DNA. Hereditas 122:1-6

Hanel R, Tsigenopoulos CS. 2011. Phylogeny, Evolution and Taxonomy of Sparids with Some Notes on their Ecology and Biology" In: Sparidae: Biology and Aquaculture of Gilthead seabream and Other Species. Chapter 2, Publisher: Blackwell Publishing Ltd.

Heath T, Hedtke SM, Hillis DM. 2008. Taxon sampling and the accuracy of phylogenetic analyses. J. Syst. Evol. 46:239-57

Hedtke SM, Townsend TM, Hillis DM. 2006. Resolution of phylogenetic conflict in large data sets by increasing taxon sampling. Syst. Biol. 55(3):522-9

Howe K,et al. 2013. The zebrafish reference genome sequence and its relationship to the human genome. Nature 496(7446):498-503

Hubbard T, et al. 2002. The Ensembl genome database project. Nucleic Acids Res. 30(1):38-41

Hughes LC, et al. 2018. Comprehensive phylogeny of ray-finned fishes (Actinopterygii) based on transcriptomic and genomic data. Proc. Natl. Acad. Sci. U. S. A., pii: 201719358

Jaillon O, et al. 2004. Genome duplication in the teleost fish *Tetraodon nigrovirdis* reveals the early vertebrate proto-karyotype. Nature 431(7011):946-57

Jones DT, Taylor WR, Thornton JM. 1992. The rapid generation of mutation data matrices from protein sequences. Comput. Appl. Biosci. 8:275–282.

Jones FC, et al. 2012. The genomic basis of adaptive evolution in threespine sticklebacks. Nature 484(7392):55-61

Kasahara M, et al. 2007. The medaka draft genome and insights into vertebrate genome evolution. Nature 447(7145):714-9

Katoh K, Standley DM. 2013. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol. Biol. Evol. 30(4):772-80

Kawahara R, et al. 2008. Interrelationships of the 11 gasterosteiform families (sticklebacks, pipefishes, and their relatives): new perspective based on whole mitogenome sequences from 75 higher teleosts. Mol. Phylogenet. Evol. 46(1):224-36

Kelley JL, et al. 2016. The Genome of the Self-Fertilizing Mangrove Rivulus Fish, Kryptolebias marmoratus: A Model for Studying Phenotypic Plasticity and Adaptations to Extreme Environments. Genome Biol. Evol. 8(7):2145-54

Kishino H, Hasegawa M. 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order of hominoidea. J. Mol. Evol. 29(2):170-9

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

Li L, Stoeckert CJ, Roos DS. 2003. OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. Genome Res. 13:2178-89

Lien S, et al. 2016. The Atlantic salmon genome provides insights into rediploidization. Nature 533(7602):200-5

Lin Q, et al. 2017. Draft genome of the lined seahorse, *Hippocampus erectus*. Gigascience 6(6):1-6

Linard B, Thompson JD, Poch O, Lecompte O. 2011. OrthoInspector: comprehensive orthology analysis and visual exploration. BMC Bioinformatics 12:11

Manousaki T, et al. 2014. The sex-specific transcriptome of the hermaphrodite sparid sharpsnout seabream (Diplodus puntazzo). BMC Genomics 15:655

McGaugh SE, et al. 2014. The cavefish genome reveals candidate genes for eye loss. Nat. Commun. 5:5307

Meynard CN, Mouillot D, Mouquet N, Douzery EJ. 2012. A phylogenetic perspective on the evolution of Mediterranean teleost fishes. PLoS One 7(5):e36443

Mylonas CC, Zohar Y, Pankhurst N, Kagara H. 2011. Reproduction and Broodstock Management. In: Sparidae: Biology and Aquaculture of Gilthead seabream and Other Species. Chapter 2, Publisher: Blackwell Publishing Ltd.

Nabhan AR, Sarkar IN. 2012. The impact of taxon sampling on phylogenetic inference: a review of two decades of controversy. Brief. Bioinform. 13(1):122-34

Nakamura Y, et al. 2013. Evolutionary changes of multiple visual pigment genes in the complete genome of Pacific blue tuna. Proc. Natl. Acad. Sci. U. S. A. 110(27):11061-6

NCBI Resource Coordinators. 2018. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 46(Database issue), D8–D13

Near TJ, et al. 2013. Phylogeny and tempo of diversification in the superradiation of spiny-rayed fishes. Proc. Natl. Acad. Sci. U. S. A. 110(30):12738-43

Nichio BTL, Marchaukoski JN, Raittz RT. 2017. New tools in orthology analysis: a brief review of promising perspectives. Front. Genet. 8:165

Orrell TM, Carpenter KE. 2004. A phylogeny of the fish family Sparidae (porgies) inferred from mitochondrial sequence data. Mol. Phylogenet. Evol. 32(2):425-34

Pan H, et al. 2016. The genome of the largest bony fish, ocean sunfish (*Mola mola*) provides insights into its fast growth rate. Gigascience 5(1):36

Ravi V, Venkatesh B. 2018. The Divergent Genomes of Teleosts. Annu. Rev. Anim. Biosci. 6:47-68

Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet 16(6):276-7

Rodgers R, Roach JL, Reid NM, Whitehead A, Duvernell DD. 2018. Phylogenomic analysis of Fundulidae (Teleostei: Cyprinodotiformes) using RNA-sequencing data. Mol. Phylogenet. Evol. 121:150-7

Rokas A, Williams BL, King N, Carroll SB. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425(6960):798-804

Rosenberg MS, Kumar S. 2001. Incomplete taxon sampling is not a problem for phylogenetic inference. Proc. Natl. Acad. Sci. U. S. A. 98(19):10751-6

Salichos L, Rokas A. 2013. Inferring ancient divergences requires genes with strong phylogenetic signals. Nature 497(7449):327-31

Salichos L, Stamatakis A, Rokas A. 2014. Novel information theory-based measures for quantifying incongruence among phylogenetic trees. Mol. Biol. Evol. 31(5):1261-71

Sanciancgo MD, Carpenter KE, Betancur-R R. 2016. Phylogenetic placement of enigmatic percomorph families (Teleostei: Percomorphaceae). Mol. Phylogenet. Evol. 94(Pt B):565-76

Santini F, Carnevale G, Sorenson L. 2014. First multi-locus timetree of seabreams and porgies (Percomorpha: Sparidae). Ital. J. Zool. 81(1):55-71

Schartl M, et al. 2013. The genome of the platyfish, *Xiphophorus maculatus*, provides insights into evolutionary adaptation and several complex traits. Nat Genet 45(5):567-72

Shimodaira H. 2002. An approximately unbiased test of phylogenetic tree selection. Syst Biol 51(3):492-508

Shimodaira H, Hasegawa M. 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol. Biol. Evol. 16(8):1114

Shimodaira H, Hasegawa M. 2001. "CONSEL: for assessing the confidence of phylogenetic tree selection" Bioinformatics 17(12):1246-7

Shin SC, et al (2014. The genome sequence of the Antarctic bullhead notothen reveals evolutionary adaptations to a cold environment. Genome Biol 15(9):468

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31(19):3210-2

Sneddon TP, Li P, Edmunds SC. 2012. GigaDB: announcing the GigaScience database. Gigascience 1(1):11

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30(9):1312-13

Star B, et al. 2011. The genome sequence of Atlantic cod reveals a unique immune system. Nature 477(7363):207-10

Sun Y, et al. 2016. Fish-T1K (Transcriptomes of 1,000 Fishes) Project: large-scale transcriptome data for fish evolution studies. Gigascience 5(1):18

Tabari E, Su Z. 2017. PorthoMCL: Parallel orthology prediction using MCL for the realm of massive genome availability. Big Data Analytics 2:4

Tine M, et al. 2014. European sea bass genome and its variation provide insights into adaptation to euryhalinity and speciation. Nat. Commun. 5:5770

Train CM, Glover NM, Gonnet GH, Altenhoff AM, Dessimoz C. 2017. Orthologous Matrix (OMA) algorithm 2.0: more robust to asymmetric evolutionary rates and more scalable hierarchical orthologous group inference. Bioinformatics 33(14):i75-i82

Tsakogiannis A, et al. 2018. The transcriptomic signature of different sexes in two protogynous hermaphrodites: Insights into the molecular network underlying sex phenotype in fish. Sci. Rep. 8(1):3564

Vij S, et al. 2016. Chromosomal-Level assembly of the Asian seabass genome using long sequence reads and multi-layered scaffolding. PLoS Genet 12(4):e1005954

Warren WC, et al. 2018. Clonal polymorphism and high heterozygosity in the celibate genome of the Amazon molly. Nat. Ecol. Evol. 2:669-79

Xu J, et al. 2017. Draft genome of the Northern snakehead, Channa argus. Gigascience 6(4):1-6

Yamanoue Y, et al. 2007. Phylogenetic position of tetraodontiform fishes within the higher teleosts: Bayesian inferences based on 44 whole mitochondrial genome sequences. Mol. Phylogenet. Evol. 45(1):89-101

Yi M, et al. 2014. Rapid evolution of piRNA pathway in the teleost fish: implication for an adaptation to transposon diversity. Genome Biol. Evol. 6(6):1393-407

You X, et al. 2014. Mudskipper genomes provide insights into the terrestrial adaptation of amphibious fishes. Nat. Commun. 5:5594

Zhang Z et al. 2018. Draft genome of the protandrous Chinese black porgy, *Acanthopagrus schlegelii*. Gigascience. 7:1-7

Zwickl DJ, Hillis DM. 2001. Increased taxon sampling greatly reduces phylogenetic error. Syst. Biol. 51(4):588-98

FIGURES AND TABLES

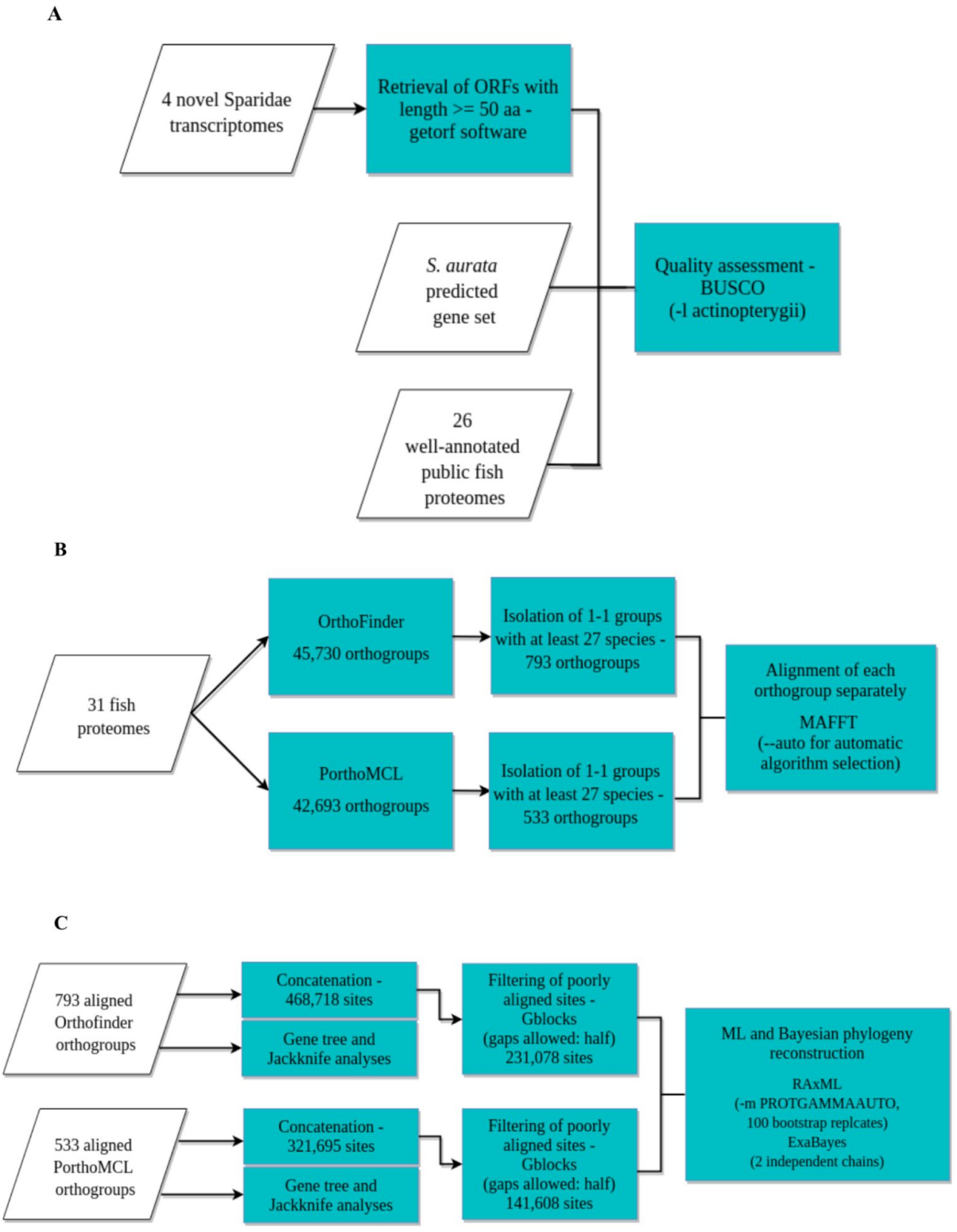

Fig. 1 Workflow: A) Taxon sampling/Quality assessment, B) Orthology assignment/MSA, C) Phylogeny reconstruction

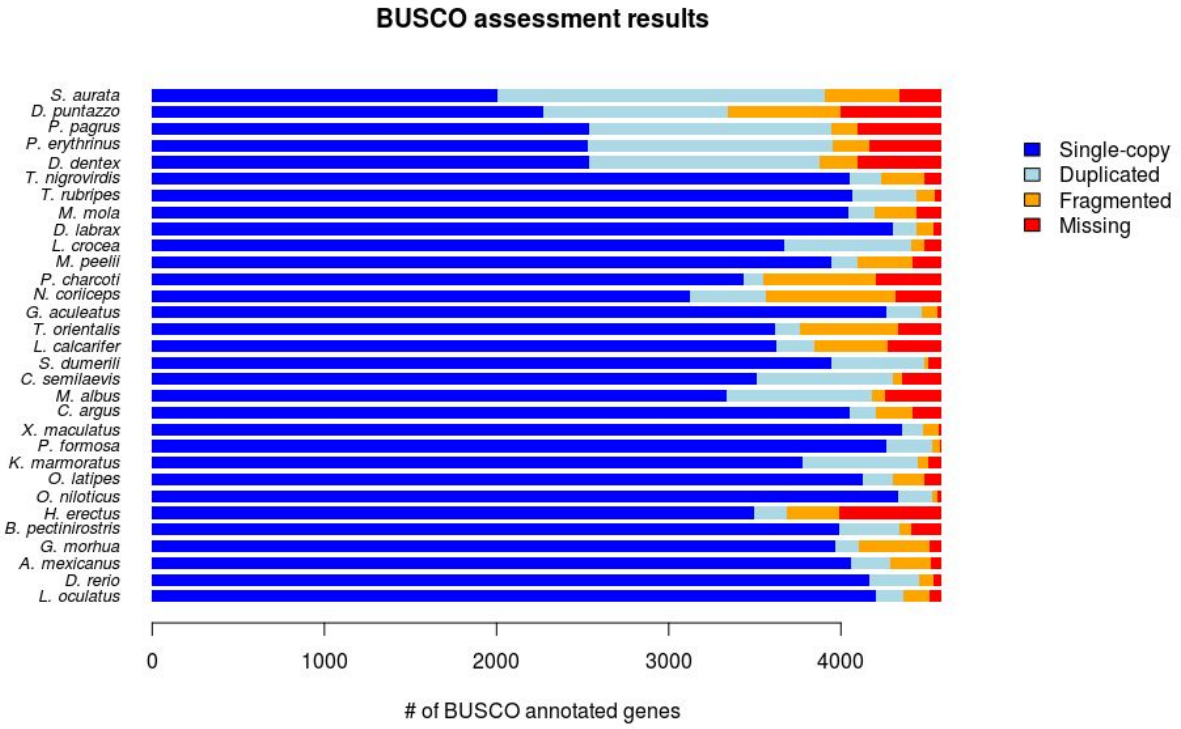

Fig. 2 Quality assessment of the 31 proteomes using BUSCO

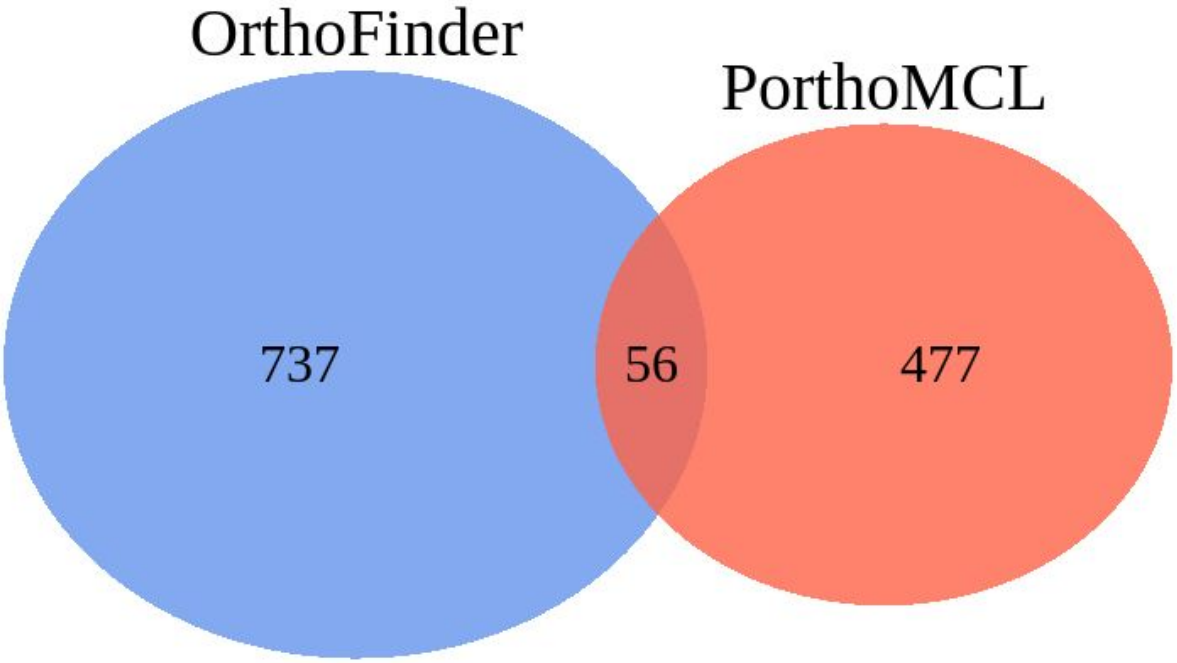

Fig. 3 Comparison of the two orthology inference results, in terms of '1-1' groups with at least 27 species present

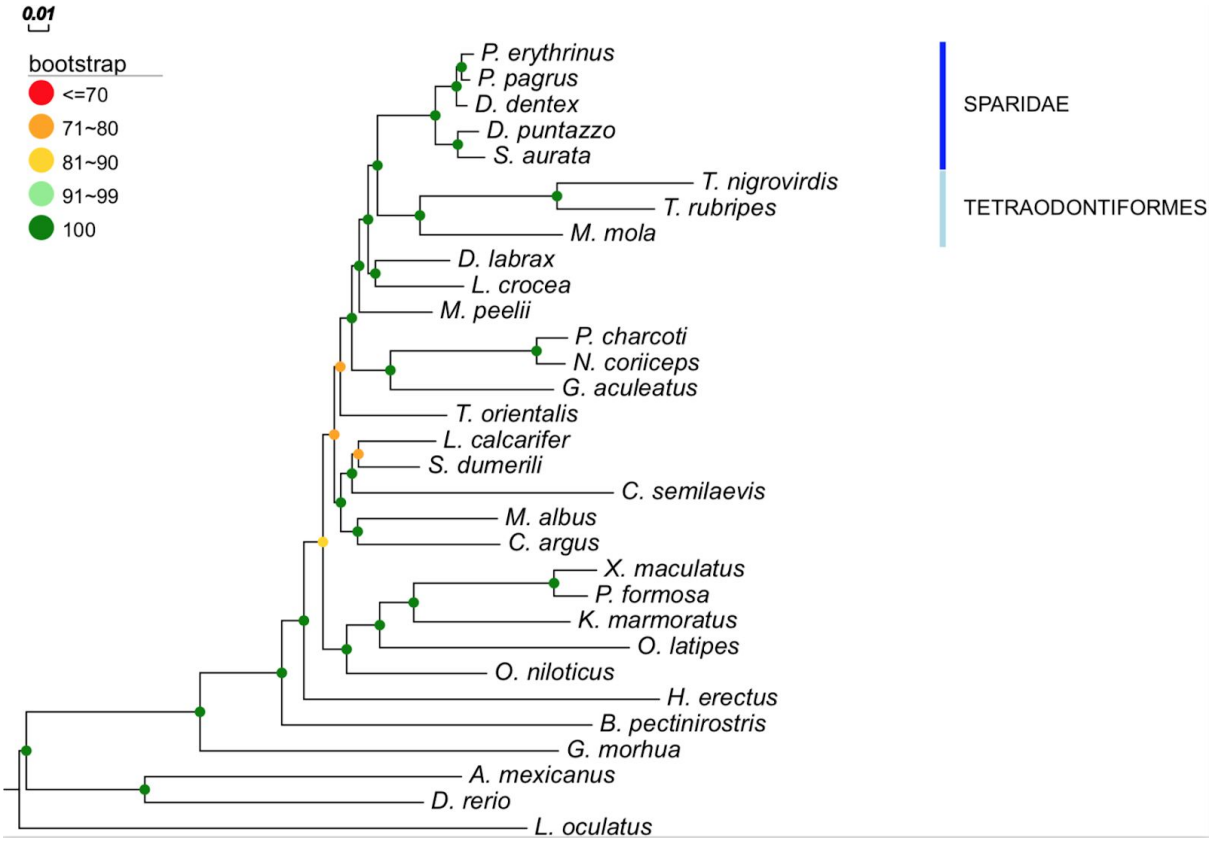

Fig. 4 Maximum likelihood (RAxML) tree of 793 concatenated OrthoFinder groups using JTT+F+Γ model and 100 bootstrap replicates. The spotted gar (L. oculatus) was used as an outgroup.

Table 1. Preprocessing of the four Sparidae transcriptomes. For each transcriptome I present the number of sequences contained, the number of open reading frames (ORF) found, the number of transcripts with at least one ORF and The final proteome included in the analysis after keeping the longest ORF per gene.

Species	**Transcripts in assemblies**	**Total number of ORFs found with length >50 a.a.**	**Transcripts with at least one ORF (% of transcripts with ORF)**	**Number of coding genes used in the final analysis**
D. puntazzo	129,012	1,272,493	113,208 (87.75%)	83,527
D. dentex	118,258	1,285,298	113,684 (96.13%)	78,451
P. erythrinus	141,309	1,416,980	129,523 (91.66%)	89,124
P. pagrus	98,012	1,264,706	91,787 (93.64%)	62,116

Table 2. List of species included in the phylogenomic analysis. For each species I indicate the series (or another distinct taxonomic group for the non-Percomorphaceae) they belong to, the sources of the proteomes used, the reference paper and the number of the protein sequences contained in each proteome.

Species	Series (for Percomorphaceae)	source	reference	#of proteins
A. mexicanus	(Ostariophysi)	Ensembl database	McGaugh et al., 2014	22,998
B. pectinirostris	Gobiaria	NCBI ftp server	You et al., 2014	21,541
C. argus	Anabantaria	GigaDB	Xu et al., 2017	20,541
C. semilaevis	Carangaria	NCBI ftp server	Chen et al., 2014	24,489
D. rerio	(Ostariophysi)	Ensembl database	Howe et al., 2013	25,644
D. dentex	Eupercaria	in-house sequenced	Tsakogiannis et al., submitted	83,527
D. labrax	Eupercaria	species database	Tine et al., 2014	26,719
D. puntazzo	Eupercaria	in-house sequenced	Manousaki et al., 2014	78,451
G. morhua	(Paracanthopterygii)	Ensembl database	Star et al., 2011	19,978
G. aculeatus	Eupercaria	Ensembl database	Jones et al., 2012	20,625
H. erectus	Syngnatharia	GigaDB	Lin et al., 2017	20,788
K. marmoratus	Ovalentaria	NCBI ftp server	Kelley et al., 2016	25,257

L. crocea	Eupercaria	NCBI ftp server	Ao et al., 2013	28.009
L. calcarifer	Carangaria	NCBI ftp server	Vij et al., 2016	22,221
L. oculatus	(Holostei)	Ensembl database	Braasch et al., 2016	18,304
M. peelii	Eupercaria	GigaDB	Austin et al., 2017	26,539
M. mola	Eupercaria	GigaDB	Pan et al., 2016	19,605
M. albus	Anabantaria	NCBI ftp server	Yi et al., 2014	24,943
N. coriiceps	Eupercaria	NCBI ftp server	Shin et al., 2014	25,937
O. niloticus	Ovalentaria	Ensembl database	Brawand et al., 2014	21,383
O. latipes	Ovalentaria	Ensembl database	Kasahara et al., 2007	19,603
P. erythrinus	Eupercaria	in-house sequenced	Tsakogiannis et al., 2018	89,124
P. pagrus	Eupercaria	in-house sequenced	Tsakogiannis et al., 2018	62,116
P. charcoti	Eupercaria	provided by authors	Ahn et al., , 2017	32,713
P. formosa	Ovalentaria	Ensembl database	Warren et al., 2018	23,315
S. dumerili	Carangaria	NCBI ftp server	Araki et al., unpublished	24,000
S. aurata	Eupercaria	in-house sequenced	Pauletto et al., 2018	61,850
T. rubripes	Eupercaria	Ensembl database	Aparicio et al., 2002	18,433
T. nigrovirdis	Eupercaria	Ensembl database	Jaillon et al., 2004	19,511

T. orientalis	Pelagiaria	species database	Nakamura et al., 2013	26,433
X. maculatus	Ovalentaria	Ensembl database	Schartl et al., 2013	20,343

Table 3. Comparison of the two orthology inference tools and the respective superalignments. OrthoFinder provided greater number of orthogroups than PorthoMCL both initially and after filtering for 1-1 groups with representation from at least 27 species.

Software	Groups of orthologs returned	Single-copy groups with at least 27 taxa	average aligned group length (a.a.)	concatenated alignment length (a.a.)	filtered alignment length (a.a.)
OrthoFinder	45,730	793	591.06	468,718	231,078
PorthoMCL	42,693	533	603.56	321,695	141,608

Table 4. Comparison of the topology presented here, with Tetraodontiformes as sister group to Sparidae, and the topology suggested by Pauletto et al., with croaker and seabass as sister group to Sparidae, using CONSEL. The table shows the p-values of various statistical tests. One may reject the possibility that a topology is the most likely to be the true when au < 0.05 at the significance level 0.05 (Shimodaira et al., 2001). Nats: present study; Paul: Pauletto et al., in press; obs: observed log-likelihood difference; au: approximately unbiased test; np: multiscale bootstrap probability; bp: usual bootstrap probability; kh: Kishino-Hasegawa test; sh: Shimodaira-Hasegawa test; wkh: weighted Kishino-Hasegawa test; wsh: weighted Shimodaira-Hasegawa test

OrthoFinder								
tree	**obs**	**au**	**np**	**bp**	**kh**	**sh**	**wkh**	**wsh**
Nats	-558.7	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Paul	558.7	4e-07	2e-06	0	0	0	0	0
PorthoMCL								
tree	**obs**	**au**	**np**	**bp**	**kh**	**sh**	**wkh**	**wsh**
Nats	-345.8	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Paul	345.8	1e-50	2e-17	0	0	0	0	0

Fig. 1. The main workflow divided into three main components: A) Taxon sampling & quality assessment, B) Orthology assignment & MSA, C) Phylogenomics analysis

Fig. 2. Quality assessment using BUSCO. The five Sparidae proteomes are shown in the five top bars.

Fig. 3. Number of single-copy groups with at least 27 species from each orthology inference software, and their intersection.

Fig. 4. Maximum likelihood (RAxML) tree of 793 concatenated OrthoFinder groups using JTT+F+Γ model and 100 bootstrap replicates. The spotted gar (*L. oculatus*) was used as an outgroup.

Supplementary table 1. Quality assessment of the 31 proteomes using BUSCO. The 5 Sparidae gene sets are depicted with bold characters. The actinopterygii-specific library of BUSCO contains 4,584 annotated genes. The *D. puntazzo* dataset contained the smallest amount of BUSCO-annotated genes (3,347).

Species	**Complete (single-copy)**	**Duplicated**	**Fragmented**	**Missing**
A. mexicanus	4,285 (4,062)	223	238	61
B. pectinirostris	4,339 (3,987)	352	70	175
C. argus	4,202 (4,052)	150	214	168
C. semilaevis	4,300 (3,512)	788	58	226
D. rerio	4,451 (4,164)	287	87	46
D. dentex	**3,876 (2,536)**	**1,340**	**218**	**490**
D. labrax	4,441 (4,302)	139	94	49
D. puntazzo	**3,347 (2,274)**	**1,073**	**650**	**587**
G. morhua	4,101 (3,971)	130	413	70
G. aculeatus	4,466 (4,267)	199	92	26
H. erectus	3,688 (3,499)	189	302	594
K. marmoratus	4,449 (3,775)	674	60	75

L. crocea	4,411 (3,674)	737	71	102
L. calcarifer	3,848 (3,622)	226	423	313
L. oculatus	4,360 (4,203)	157	156	68
M. peelii	4,095 (3,941)	154	318	171
M. mola	4,194 (4,043)	151	246	144
M. albus	4,183 (3,335)	848	73	328
N. coriiceps	3,568 (3,124)	444	751	265
O. niloticus	4,531 (4,331)	200	30	23
O. latipes	4,304 (4,127)	177	182	98
P. erythrinus	**3,954 (2,533)**	**1,421**	**210**	**420**
P. pagrus	**3,945 (2,538)**	**1,407**	**152**	**487**
P. charcoti	3,552 (3,432)	120	651	381
P. formosa	4,529 (4,267)	262	45	10
S. dumerili	4,486 (3,943)	543	22	76
S. aurata	**3,910 (2,004)**	**1,906**	**428**	**246**
T. rubripes	4,440 (4,065)	375	105	39
T. nigrovirdis	4,231 (4,049)	182	257	96

T. orientalis	3,762 (3,614)	148	571	251
X. maculatus	4,475 (4,353)	122	94	15

Supplementary Fig. 1. Maximum likelihood tree of 533 concatenated PorthoMCL groups using RAxML

Supplementary Fig. 2. Maximum likelihood trees of 793 OrthoFinder groups using different subsets of Sparidae/Tetraodontiformes species: A) only seabream from Sparidae, B) sunfish removed from Tetraodontiformes and C) only sunfish from Tetraodontiformes

Supplementary Fig. 3. Consensus trees of 100 jackknifed replicates (70% of groups kept) for: A) OrthoFinder and B) PorthoMCL groups

Supplementary Fig. 4. Bayesian consensus trees after 25% burn-in for two parallel MCMC chains for: A) OrthoFinder and B) PorthoMCL groups

Supplementary Fig. 5. Maximum likelihood tree of 793 OrthoFinder groups A) without tongue sole and B) without tongue sole and pacific bluefin tuna

Supplementary Fig. 6. Gene tree analysis consensus trees for: A) 135 OrthoFinder and B) 78 PorthoMCL groups with all 31 species present. Figures C and D show the results of IC/ICA calculation by RAxML, for OrthoFinder an PorthoMCL respectively

Supplementary Fig. 7. Maximum likelihood trees with the 14 species used in gilthead seabream genome paper for: A) 2,192 OrthoFinder and B) 1,366 PorthoMCL single-copy groups with at least 13 species present.

* Supplementary Figures are available upon request to the author at pnatsidis@hotmail.com