

**Diversification of Hox gene clusters in osteoglossomorph fish
in comparison to other teleosts and the spotted gar outgroup**

Kyle J Martin ^{1,2*} and Peter W H Holland ¹

¹ Department of Zoology, University of Oxford,

South Parks Road, Oxford OX1 3PS, UK

² Department of Animal and Plant Sciences, University of Sheffield,

Western Bank, Sheffield S10 2TN, UK

kyle.john.martin@gmail.com; peter.holland@zoo.ox.ac.uk

*Corresponding author

1 Table

1 Figure

2 Supplementary Data

Running head: Osteoglossomorph Hox genes

Correspondence to: Kyle Martin kyle.john.martin@gmail.com

Grant sponsor: European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013 ERC grant 268513).

Abstract

An ancient genome duplication (TGD or 3R) occurred in teleost fish after divergence from the lineage leading to gar. This genome duplication is shared by the three extant teleost lineages: Osteoglossomorpha (bony-tongues), Elopomorpha (eels and tarpons) and Clupeocephala (a large clade including salmon, carp, medaka, zebrafish, cichlids, pufferfish, stickleback and ~26,000 other species). After TGD, different clupeocephalan species retained different gene duplicates; this is seen clearly in Hox gene clusters but extends to all genes. Since divergent resolution of TGD paralogs is a potential driving force for speciation, it is possible this contributed to diversification of this clade. The extent to which divergent resolution of TGD paralogs occurred within Osteoglossomorpha has not been investigated in detail, and Hox cluster organisation has been reported for just two species: *Pantodon buchholzi* (Pantodontidae) and *Scleropages formosus* (Osteoglossidae). We applied survey-scale genome sequencing and *de novo* assembly to three further osteoglossomorph taxa: *Osteoglossum bicirrhosum* (Osteoglossidae), *Chitala ornata* (Notopteridae), and *Gnathonemus petersii* (Mormyridae). We find that each retained more Hox genes than clupeocephalan taxa (excluding those that underwent additional genome duplication), but fewer than eels. Several Hox genes are missing in all teleosts, including duplicates of two Hox genes present in the slow evolving pre-TGD genome of the spotted gar. We find divergent resolution through individual gene losses and whole cluster losses has been rampant across osteoglossomorphs, despite their extant species paucity. We suggest that reciprocal gene loss following TGD were probably insufficient to drive the exceptional diversification of teleosts.

Keywords

Homeobox; arowana; mormyrid; knifefish; genome

Introduction

It has been hypothesized that whole genome duplication (WGD) can accelerate speciation through different populations resolving genetic redundancy in distinct ways, leading to hybrid dysfunction or sterility (Lynch and Force, 2000; Presgraves, 2010). An ancient WGD occurred in the teleost lineage ~350 Ma (teleost genome duplication, TGD, also called 3R) leading to speculation that this mechanism played a role in the radiation of teleosts into over 26,000 species (Taylor et al., 2001a; Hoegg et al., 2004; Jaillon et al., 2004). Divergent resolution of redundancy can occur by reciprocal non-functionalization (gene loss) between lineages or by reciprocal subfunctionalization (partitioning of parental gene functions between daughter genes) for example through the duplication-degeneration-complementation (DDC) pathway (Force et al., '99; Postlethwait et al., 2004). Previous studies have investigated the degree of divergent resolution within clupecocephalans, the largest teleost subdivision of teleosts consisting of ostarioclupeomorphs (zebrafish, carp, cavefish etc.) and euteleosts (medaka, cichlids, pufferfish etc.), and have revealed abundant genome-wide reciprocal gene loss (Sémon and Wolfe, 2007; Garcia de la Serrana et al., 2014). For example, different patterns of post-TGD gene loss can be seen very clearly by comparing Hox gene clusters of zebrafish, pufferfish, cichlids, stickleback and medaka (Hoegg et al., 2007). It should be noted that the postulated link between WGD and teleost diversification is controversial, and has recently been challenged by the demonstration that extinct 'pre-TGD' holosteans were as phenotypically diverse as early 'post-TGD' teleosts (Clarke et al. 2016). However, recent genomic analyses have shown that paralogue divergence can be delayed for tens of millions of years after polyploidy so close temporal correlation between WGD and diversity should be not be expected even if WGD promotes diversification (Martin and Holland, 2014; Macqueen and Johnston 2014).

While the TGD is unambiguously a synapomorphy of all teleosts (Taylor et al., 2001a,b; Henkel et al., 2012; Martin and Holland, 2014), species richness is not evenly spread across teleost phylogeny. Indeed, the Ostariophysii (~9,000 species) and Percomorpha (~16,000 species), both within the Clupecocephala, account for most of extant teleost biodiversity (Santini et al., 2009). In addition to the Clupecocephala, two other major teleost subdivisions are recognized: Elopomorpha (eels and tarpons) and Osteoglossomorpha (bony tongues). Each has relatively low extant diversity, with just ~800 and ~220 described species respectively. There is disagreement about the phylogenetic relationships between osteoglossomorphs, elopomorphs and clupecocephalans, but the monophyly of each is consistently well supported (Le et al., '93; Inoue et al., 2001; Inoue et al., 2004; Broughton, 2010; Broughton et al., 2013). It is unclear to what extent divergent resolution of TGD paralogs occurred within these species-poor clades. One way to investigate this issue is to compare Hox gene cluster

organisation across the Osteoglossomorpha, while recognizing that Hox genes are not necessarily a proxy for the whole genome. Hox genes have well recognized roles in developmental patterning, and their evolution has been studied extensively. They can be readily compared between vertebrate species facilitating analysis of gene loss, thereby enabling divergent resolution of paralogs to be examined. Furthermore, Hox gene clusters are well annotated in the genome of the spotted gar, *Lepisosteus oculatus*, a non-teleost fish that diverged before the TGD event (Braasch et al., 2016). Comparison to gar will be informative because differences to the Hox gene clusters of osteoglossomorphs will reveal the extent of gene loss in each lineage since their common ancestor.

The first study to investigate Hox clusters of an osteoglossomorph fish was a PCR study of *Hiodon alosoides* (Chambers et al., 2009). This identified 41 Hox gene fragments that could be provisionally assigned to eight Hox gene clusters, suggesting that *Hiodon* might have suffered no Hox cluster losses following TGD. Assignment to specific clusters, however, was compromised by the short length of PCR fragments obtained and absence of data from other osteoglossomorphs. More recently, we used genome sequencing and BAC cloning to analyse Hox genes in an osteoglossomorph from the monotypic family Pantodontidae, *Pantodon buchholzi*, finding a very different Hox cluster organization (Martin and Holland, 2014). Only five Hox gene clusters were found in *Pantodon*, plus a probable pseudogene remnant of a sixth cluster, making it the teleost with the fewest Hox clusters described to date (Martin and Holland, 2014). Despite having fewer Hox clusters than *Hiodon* or clupeocephalans, *Pantodon* has 45 Hox genes. These include four genes (*hoxb2y*, *hoxb4y*, *hoxb9y*, *hoxb13y*) not present in the genomes of clupeocephalan teleosts; one, *hoxb13y*, is also absent from the otherwise Hox-rich genomes of eels (Guo et al., 2010; Henkel et al., 2012). These findings demonstrate divergent resolution at the level of individual genes and entire clusters between the three major teleost subdivisions and hints at diversity amongst the osteoglossomorphs. Another demonstration of differences amongst osteoglossomorphs came from genome sequence of the Asian arowana (*Scleropages formosus*) which revealed 59 Hox genes in eight gene clusters (Bian et al., 2016).

To further compare Hox gene cluster organisation across the Osteoglossomorpha, here we investigate Hox gene diversity in three additional species: *Osteoglossum bicirrhosum* (Silver arowana), *Chitala ornata* (Clown knifefish) and *Gnathonemus petersii* (Peters' elephantnose fish). Addition of these species extends the gene cluster analysis to five extant families of Osteoglossomorpha: Hiodontidae (*H. alosoides*, Chambers et al., 2009), Pantodontidae (*P. buchholzi*, Martin and Holland 2014), Osteoglossidae (*S. formosus*, Bian et al., 2016; *O. bicirrhosum*, this study), Notopteridae (*C. ornata*; this study) and Mormyridae (*G. petersii*; this study). The strategy deployed was low coverage *de novo* draft genome sequencing and assembly, followed by bioinformatic survey of Hox gene complements.

Materials and Methods

Juvenile specimens of *Osteoglossum bicirrhosum*, *Chitala ornata* and *Gnathonemus petersii* were obtained commercially; heterozygosity was not reduced by deliberate inbreeding. Fish were killed humanely using overdose of anaesthetic (MS222) followed by confirmation of death. High quality genomic DNA was extracted from tissue and used to generate paired-end libraries with 200 bp average insert size according to the Illumina HiSeq2000 protocol. Three barcoded libraries were pooled and sequenced with 100 bp read-length on a single lane of the HiSeq2000 platform yielding 12.9 to 14.1 Gb sequence data per species (available at NCBI SRA via BioProject PRJNA347616). Data were assembled using the short read assembler Velvet (Zerbino and Birney, 2008) with a range of kmers (27, 37, 47, 57); assembly metrics compared included N50, total assembly size and representation of 248 conserved core eukaryotic genes (CEGs) (Parra et al., 2007). The assembly with the highest N50 was used, as CEG representation did not differ by more than one gene between assemblies (available at ORA-Data, doi 10.5287/bodleian:mvNpmVGZD). The *Scleropages* genome (Bian et al., 2016) was not used to assist *Osteoglossum* assembly, as it was published subsequent to the work described here. Genome size for each species was estimated using kmer counting (Li et al., 2010; Liu et al., 2013).

We devised a progressive reciprocal BLAST-based screening method for surveying draft genome assemblies for Hox genes. For initial search sequences we used the 43 Hox genes of spotted gar, *Lepisosteus oculatus*, because this species apparently experienced no Hox gene losses since divergence from its latest common ancestor with teleosts and its Hox protein coding sequences have been evolving slowly (Braasch et al. 2016). These were used as query input using tblastn (Altschul et al., '90) and all contigs returning significantly similar hits were recovered ($E=10^{-6}$). The resultant 363, 312 and 321 contigs from *Osteoglossum*, *Chitala* and *Gnathonemus* were then interrogated by blastx against the NCBI nr database using Blast2GO (Götz et al., 2008). As expected, many contigs contained non-Hox homeobox genes; this gives confidence in the sensitivity of the gar-seeded searches to detect all Hox genes, because divergent sequences were also returned. To help identify paralogy group and cluster, we also used tblastx searches using individual deduced exons of osteoglossomorph Hox contigs as reciprocal queries against gar Hox genes. In most cases we could confidently assign osteoglossomorph genes or exons to the gnathostome-level Hox paralogy group and cluster first letter (a, b, c, d); a few could only be assigned to paralogy group (1-13). We could not confidently assign post-TGD cluster identity (e.g. *hoxaa* vs *hoxab* cluster); indeed, as previously shown, it can be difficult or theoretically impossible to reconstruct 1:1 orthology relationships between many duplicates in osteoglossomorphs and clupeocephalans, even with genomic linkage data, due to differing timings of

post-TGD rediploidisation (Martin and Holland, 2014). Sequences Hox genes are available as Supplementary File 1.

Results

We sequenced genomic DNA from three osteoglossomorph species and assembled low-coverage genomes. These are fragmentary assemblies, not intended to be used as resources for chromosomal level analyses, but as tools for exploring the diversity and representation of Hox genes. Assembly metrics are presented in table 1. We estimate the genome size of *Osteoglossum bicirrhosum* at 1142 Mb, slightly larger than the fluorometric estimate of 978 Mb (Hinegardner and Rosen, '72). We estimate the genome size of *Chitala ornata* to be 910 Mb, slightly below than the flow cytometry estimate of 1056 Mb for a closely related species (Ojima and Yamamoto, '90). Finally, we estimate the *Gnathonemus petersii* genome to be 1091 Mb, close to the previously published estimate of 1174 Mb (Hinegardner and Rosen, '72). Using our genome size estimates, actual sequence coverage ranges from 11X to 14X. The total assembly size for each genome is lower than the expected genome size (52-66%), indicating that many sequence reads were not assembled. It is likely that much of the missing assembly is repetitive DNA because short read sequences will not be able to distinguish between DNA repeat copies, even with paired-end data. Searches for conserved eukaryotic genes (CEGs) revealed at least partial models for 76-83% of genes suggesting reasonably good coverage of the protein-coding fraction of the genomes. The coverage for Hox genes is expected to be higher because in vertebrates these are generally small two-exon genes and teleost introns are small (median lengths 120-250 bp; Moss et al., 2011). We therefore expect our estimates of Hox gene numbers to be only slight underestimates. Inferences about gene losses must be made with caution, as subsequent deeper coverage sequencing may reveal genes that have been missed in the current study.

A test of accuracy was undertaken by comparing of our data inferred for Silver arowana (*Osteoglossum bicirrhosum*) with the complete Hox gene clusters published for the Asian arowana (*Scleropages formosus*; Bian et al., 2016). These species are from different genera but both belong to the family Osteoglossidae. We find close concordance between the two datasets, with at least 59 Hox genes inferred by our analysis of Silver arowana and 59 published for Asian arowana (Supplementary File 2). The datasets differ by presence of a *Hoxd8* gene in Asian arowana which was not found in our analyses; this could represent a biological difference or slight undersampling. We also find an additional Hox4-like exon 1 that could not be confidently assigned to a particular cluster. We conclude, therefore, that

the draft genome approach is a reasonably accurate and useful approach to estimating Hox gene family complexity.

We identified putative full coding sequence of 15, 25, and 29 Hox genes in *Osteoglossum*, *Chitala* and *Gnathonemus* respectively, plus numerous exon 1 and exon 2 sequences from Hox genes. In *Osteoglossum*, we find evidence for (minimally) 59 Hox genes, representing at least seven clusters. We do not detect strong evidence of a second Hoxd cluster, however it is possible that one or more unclassified exons not confidently assigned to cluster level might belong to a second Hoxd cluster. In *Gnathonemus* we find at least 52 Hox genes, representing at least six clusters including two Hoxa and two Hoxc clusters. In *Chitala* we find at least 59 Hox genes and evidence for probable eight clusters with two copies of at least one paralogy group in each of the Hoxa, Hoxb, Hoxc, and Hoxd clusters. Interestingly, whereas in *Pantodon* Hoxb was the only cluster to be retained in duplicate, with both copies extremely replete in Hox genes (Martin and Holland 2014), in *Gnathonemus* one Hoxb cluster may have been lost completely while in *Chitala* it appears that only one Hoxb gene has both duplicates preserved. Hence, the constraints that preserve so many Hoxb duplicates in Pantodontidae are not globally present across Osteoglossomorpha. We find no putative Hox14 paralogs in any of the osteoglossomorph species examined. With low coverage sequence data, we did not attempt to link genes physically; however, we did detect *hoxb8-hoxb7* and *hoxb6-hoxb5* linkages in *Gnathonemus*, and a *hoxa2-hoxa1* linkage in *Chitala*.

Figure 1 depicts a summary of all putative Hox genes discovered in the genome analyses undertaken in this study plus previously published results from *Pantodon*, along with data from an elopomorph (*Anguilla*), two clupeocephalans (*Danio* and *Gasterosteus*) and the gar outgroup (*Lepisosteus*), mapped onto a phylogeny based on Lavoué et al. (2010). Colour coding denotes confident cluster assignment.

The data reveal a high degree of evolutionary lability within Hox gene clusters of osteoglossomorphs. This is particularly apparent at the level of Hox cluster number. At one extreme, our data suggest that *Chitala* has retained all eight clusters generated by the TGD event; this contrasts with the five Hox clusters retained by *Pantodon* (which lost one copy each of the Hoxa, Hoxc and Hoxd clusters). Independent loss of one Hoxd cluster may have taken place in the lineage leading to *Osteoglossum*. We note that a closely related genus *Scleropages* retains only a single gene from its second Hoxd cluster and therefore loss of this cluster likely occurred progressively rather than through whole-cluster deletion. There is also possible independent loss of one Hoxd cluster and one Hoxb cluster in *Gnathonemus*. In all cases, further genomic sequencing will be necessary to confirm cluster losses.

Dramatic differences between osteoglossomorph species are also present at the level of individual genes, particularly relating to which post-TGD duplicates have returned to single copy (Figure 1).

Comparison between osteoglossomorphs and gar are also informative (Figure 1). We note two examples of Hox genes present in gar but not yet been detected in any osteoglossomorph species: these are *Hoxa6* and *Hoxd2*. These two genes are also missing from the novel osteoglossomorph teleost genomes we examined, and could reflect gene losses that occurred before TGD or at least before the radiation of the extant teleost lineages. In contrast, we find no examples of Hox genes that are absent in gar but present in osteoglossomorph fish. This is consistent with, and further strengthens, the developing view that the gar lineage experienced no Hox gene losses in the ~352 Ma since its divergence from teleosts (Hurley et al., 2007), demonstrating remarkable evolutionary stability.

Discussion

The spotted gar *Lepisosteus oculatus* has 43 Hox genes in four genomic clusters, and is proposed to have lost no Hox genes since the common ancestor of gar and teleost fish (Braasch et al., 2016), just over 350 million years ago (Hurley et al., 2007). If we assume that no Hox genes were lost between this ancestral state and the TGD event, then the teleost ancestor immediately post-TGD will have possessed 86 (=2x43) Hox genes in eight clusters. This figure drops to 82 if *Hoxa6* and *Hoxd2* were lost before TGD.

It is striking how variable the subsequent pathways of Hox gene evolution have been. A small number of Hox genes were most likely lost between TGD and divergence of the three major teleost clades (Osteoglossomorpha, Elopomorpha, Clupeocephala). Deducing the precise number is dependent on sampling, but it is certainly few (Henkel et al., 2012; Bian et al., 2016; Braasch et al. 2016). In the Elopomorpha, current data suggest rather little gene loss subsequent to this state; for example, the European eel *Anguilla anguilla* retains 73 Hox genes and all eight clusters (Henkel et al. 2012). In contrast, both clupeocephalans and osteoglossomorphs have generally lost many more Hox genes. For example, within Clupeocephala the pufferfish *Takifugu rubripes* has 45 Hox genes and zebrafish *Danio rerio* has 49. Within Osteoglossomorpha, *Pantodon buchholzi* has 45 Hox genes, Asian arowana *Scleropages formosus* has 59, and we estimate other osteoglossomorphs each have 52 to 59 Hox genes. More striking than the gene numbers are the differences between species, both at the level of

individual genes and whole gene clusters. This has long been recognized within clupeocephalans, where *Takifugu* is missing the entire Hoxcb cluster but *Danio* lacks the Hoxdb cluster (apart from a miRNA gene; Woltering and Durston, 2006). Similarly, in osteoglossomorphs we previously found that *Pantodon* lacks three clusters (Martin and Holland, 2014), and in the current study we find evidence suggesting independent loss of one Hox cluster in *Osteoglossum* and probable loss of two in *Gnathonemus*. There are also numerous differences in presence or absence of particular Hox genes.

These data reveal dynamic patterns of Hox gene loss across Osteoglossomorpha, including divergent resolution of duplicated genes and clusters in a comparable manner to previously reported patterns in Clupeocephala. It has previously been suggested that divergent resolution of gene duplicates was a driving force for the extensive radiation of clupeocephalans into over 26,000 species (Taylor et al., 2001a; Hoegg et al., 2004; Jaillon et al., 2004). While this is an attractive hypothesis, here we note extensive divergent resolution of Hox gene duplicates without extensive speciation in the Osteoglossomorpha. It remains to be determined if these patterns extend to other genes in osteoglossomorph genomes, and ultimately how other factors combine with genomic processes to effect speciation in each major clade of ray-finned fish. However, although Hox gene loss has been rampant throughout Clupeocephala and Osteoglossomorpha, this contrasts markedly with the European eel, an elopomorph. The contrast to gar is even more striking, and the data reported here add further support to the contention that gar retains ancestral gene complements to an unusual and exceptional degree.

Acknowledgements

The authors thank Ingo Braasch, John Postlethwait, Jordi Paps, Ignacio Maeso, Ferdinand Marletaz, Jerome Hui, Adam Hargreaves and Sebastian Shimeld for helpful discussions. Sequencing was undertaken at the High-Throughput Genomics Group at the Wellcome Trust Centre for Human Genetics funded by Wellcome Trust grant 090532/Z/09/Z. Raw sequence data is available from NCBI SRA (BioProject PRJNA347616); assemblies are available from ORA-Data (doi 10.5287/bodleian:mvNpmVGZD). K.J.M. acknowledges support from the National Sciences and Engineering Research Council (NSERC) of Canada and the Oxford University Press Clarendon Fund. P.W.H.H. and K.J.M. acknowledge support from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013)/ERC grant (268513). The authors declare no conflicts of interest.

References

- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic Local Alignment Search Tool. *J. Mol. Biol.* 215:403–410.
- Bian C, Hu Y, Ravi V, Kuznetsova IS, Shen X, Mu X, Sun Y, You X, Li J, Li X, Qiu Y, Tay BH, Thevasagayam NM, Komissarov AS, Trifonov V, Kabilov M, Tupikin A, Luo J, Liu Y, Song H, Liu C, Wang X, Gu D, Yang Y, Li W, Polgar G, Fan G, Zeng P, Zhang H, Xiong Z, Tang Z, Peng C, Ruan Z, Yu H, Chen J, Fan M, Huang Y, Wang M, Zhao X, Hu G, Yang H, Wang J, Wang J, Xu X, Song L, Xu G, Xu P, Xu J, O'Brien SJ, Orbán L, Venkatesh B, Shi, Q. 2016. The Asian arowana (*Scleropages formosus*) genome provides new insights into the evolution of an early lineage of teleosts. *Scientific Reports* 6:24501.
- Braasch I, Gehrke AR, Smith JJ, Kawasaki K, Manousaki T, Pasquier J, Amores A, Desvignes T, Batzel P, Catchen J, Berlin AM, Campbell MS, Barrell D, Martin KJ, Mulley JF, Ravi V, Lee AP, Nakamura T, Chalopin D, Fan S, Wcisel D, Cañestro C, Sydes J, Beaudry FE, Sun Y, Hertel J, Beam MJ, Fasold M, Ishiyama M, Johnson J, Kehr S, Lara M, Letaw JH, Litman GW, Litman RT, Mikami M, Ota T, Saha NR, Williams L, Stadler PF, Wang H, Taylor JS, Fontenot Q, Ferrara A, Searle SM, Aken B, Yandell M, Schneider I, Yoder JA, Volff JN, Meyer A, Amemiya CT, Venkatesh B, Holland PW, Guiguen Y, Bobe J, Shubin NH, Di Palma F, Alföldi J, Lindblad-Toh K, Postlethwait JH. 2016. The spotted gar genome illuminates vertebrate evolution and facilitates human-teleost comparisons. *Nature Genet* 48:427-437.
- Chambers KE, McDaniel R, Raincrow JD, Deshmukh M, Stadler PF, Chiu C. 2009. Hox cluster duplication in the basal teleost *Hiodon alosoides* (Osteoglossomorpha). *Theory Biosci* 128:109–120.
- Clarke JT, Lloyds GT, Friedman M. 2016. Little evidence for enhanced phenotypic evolution in early teleosts relative to their living fossil sister group. *Proc Natl Acad Sci USA* 113: 11531-11536.
- Force A, Lynch M, Pickett FB, Amores a, Yan YL, Postlethwait J. 1999. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* 151:1531–1545.
- Garcia de la Serrana D, Mareco E A, Johnston I A. 2014. Systematic variation in the pattern of gene paralog retention between the teleost superorders Ostariophysi and Acanthopterygii. *Genome Biol Evol* 6:981–987.
- Götz S, García-Gómez JM, Terol J, Williams TD, Nagaraj SH, Nueda MJ, Robles M, Talón M, Dopazo J, Conesa A. 2008. High-throughput functional annotation and data mining with the Blast2GO suite. *Nucl Acids Res* 36:3420–3435.
- Guo B, Gan X, He S. 2010. Hox genes of the Japanese eel *Anguilla japonica* and Hox cluster evolution in teleosts. *J Exp Zool B Mol Dev Evol* 314:135–147.

Henkel CV, Burgerhout E, de Wijze DL, Dirks RP, Minegishi Y, Jansen HJ, Spaink HP, Dufour S, Weltzien FA, Tsukamoto K, van den Thillart GE. 2012. Primitive duplicate Hox clusters in the European eel's genome. PLoS ONE 7:e32231.

Hinegardner R, Rosen DE. 1972. Cellular DNA content and the evolution of teleostean fishes. Amer Nat 106:621–644.

Hoegg S, Brinkmann H, Taylor JS, Meyer A. 2004. Phylogenetic timing of the fish-specific genome duplication correlates with the diversification of teleost fish. J Mol Evol 59:190–203.

Hurley IA, Mueller RL, Dunn KA, Schmidt EJ, Friedman M, Ho RK, Prince VE, Yang Z, Thomas MG, Coates MI. 2007. A new time-scale for ray-finned fish evolution. Proc Roy Soc B 274:489–498.

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

Inoue JG, Miya M, Tsukamoto K, Nishida M. 2004. Mitogenomic evidence for the monophyly of elopomorph fishes (Teleostei) and the evolutionary origin of the leptocephalus larva. Mol Phylo Evol 32:274–286.

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

Lavoué S, Miya M, Arnegard ME, McIntyre PB, Mamonekene V, Nishida M. 2010. Remarkable morphological stasis in an extant vertebrate despite tens of millions of years of divergence. Proc Roy Soc B 278:1003–1008.

Le HLV, Lecointre G, Perasso R. 1993. A 28S rRNA-based phylogeny of the gnathostomes: first steps in the analysis of conflict and congruence with morphologically based cladograms. Mol Phylo Evol 2:31–51.

Li R, Fan W, Tian G, Zhu H, He L, Cai J, Huang Q, Cai Q, Li B, Bai Y, Zhang Z, Zhang Y, Wang W, Li J, Wei F, Li H, Jian M, Li J, Zhang Z, Nielsen R, Li D, Gu W, Yang Z, Xuan Z, Ryder OA, Leung FC, Zhou Y, Cao J, Sun X, Fu Y, Fang X, Guo X, Wang B, Hou R, Shen F, Mu B, Ni P, Lin R, Qian W, Wang G, Yu C, Nie W, Wang J, Wu Z, Liang H, Min J, Wu Q, Cheng S, Ruan J, Wang M, Shi Z, Wen M, Liu B, Ren X, Zheng H,

Dong D, Cook K, Shan G, Zhang H, Kosiol C, Xie X, Lu Z, Zheng H, Li Y, Steiner CC, Lam TT, Lin S, Zhang Q, Li G, Tian J, Gong T, Liu H, Zhang D, Fang L, Ye C, Zhang J, Hu W, Xu A, Ren Y, Zhang G, Bruford MW, Li Q, Ma L, Guo Y, An N, Hu Y, Zheng Y, Shi Y, Li Z, Liu Q, Chen Y, Zhao J, Qu N, Zhao S, Tian F, Wang X, Wang H, Xu L, Liu X, Vinar T, Wang Y, Lam TW, Yiu SM, Liu S, Zhang H, Li D, Huang Y, Wang X, Yang G, Jiang Z, Wang J, Qin N, Li L, Li J, Bolund L, Kristiansen K, Wong GK, Olson M, Zhang X, Li S, Yang H, Wang J, Wang J. 2010. The sequence and de novo assembly of the giant panda genome. *Nature* 463:311–317.

Liu B, Shi Y, Yuan J, Hu X, Zhang H, Li N, Li Z, Chen Y, Mu D, Fan W. 2013. Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome projects. *arXiv* 1308.2012.

Macqueen DJ, Johnston IA. 2014. A well-constrained estimate for the timing of the salmonid whole genome duplication reveals major decoupling from species diversification. *Proc Roy Soc B* 281: 20132881.

Martin KJ, Holland PWH. 2014. Enigmatic orthology relationships between Hox clusters of the African butterfly fish and other teleosts following ancient whole-genome duplication. *Mol Biol Evol* 31:2592–2611.

Moss SP, Joyce D a, Humphries S, Tindall KJ, Lunt DH. 2011. Comparative analysis of teleost genome sequences reveals an ancient intron size expansion in the zebrafish lineage. *Genome Biol Evol* 3:1187–1196.

Ojima Y, Yamamoto K. 1990. Cellular DNA contents of fishes determined by flow cytometry. *La Kromosomo* 57:1871-1888.

Para G, Bradnam K, Korf I. 2007. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* 23:1061-1067.

Postlethwait J, Amores A, Cresko W, Singer A, Yan Y-L. 2004. Subfunction partitioning, the teleost radiation and the annotation of the human genome. *Trends Genet* 20:481–490.

Presgraves DC. 2010. The molecular evolutionary basis of species formation. *Nature Rev Genet* 11:175–180.

Sémon M, Wolfe KH. 2007. Reciprocal gene loss between *Tetraodon* and zebrafish after whole genome duplication in their ancestor. *Trends Genet* 23:108–112.

Taylor JS, Van de Peer Y, Meyer A. 2001a. Genome duplication, divergent resolution and speciation. *Trends Genet* 17:299–301.

Taylor JS, Van de Peer Y, Meyer A. 2001b. Revisiting recent challenges to the ancient fish-specific genome duplication hypothesis. *Curr Biol* 11:R1005–8.

Zerbino DR, Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* 18:821–829.

Table 1

Osteoglossomorph sequence data metrics

	<i>Osteoglossum bicirrhosum</i>	<i>Chitala ornata</i>	<i>Gnathonemus petersii</i>
Paired-end reads	64,562,111	70,583,995	69,889,639
Data (Gb)	12.91	14.12	13.98
Genome coverage estimate	13.1X	14.3X	11.1X
Genome size estimate (Mb)	1142	910	1091
Best kmer	47	27	47
Contig N50 (bp)	1,936	2,880	2,816
Contigs	436,545	338,181	398,683
Total Assembly (Mb)	594	604	669
% GC	45.76	41.86	43.82
Full CEGs	109 (44%)	103 (41%)	120 (48%)
Partial CEGs	189 (76%)	204 (82%)	205 (83%)

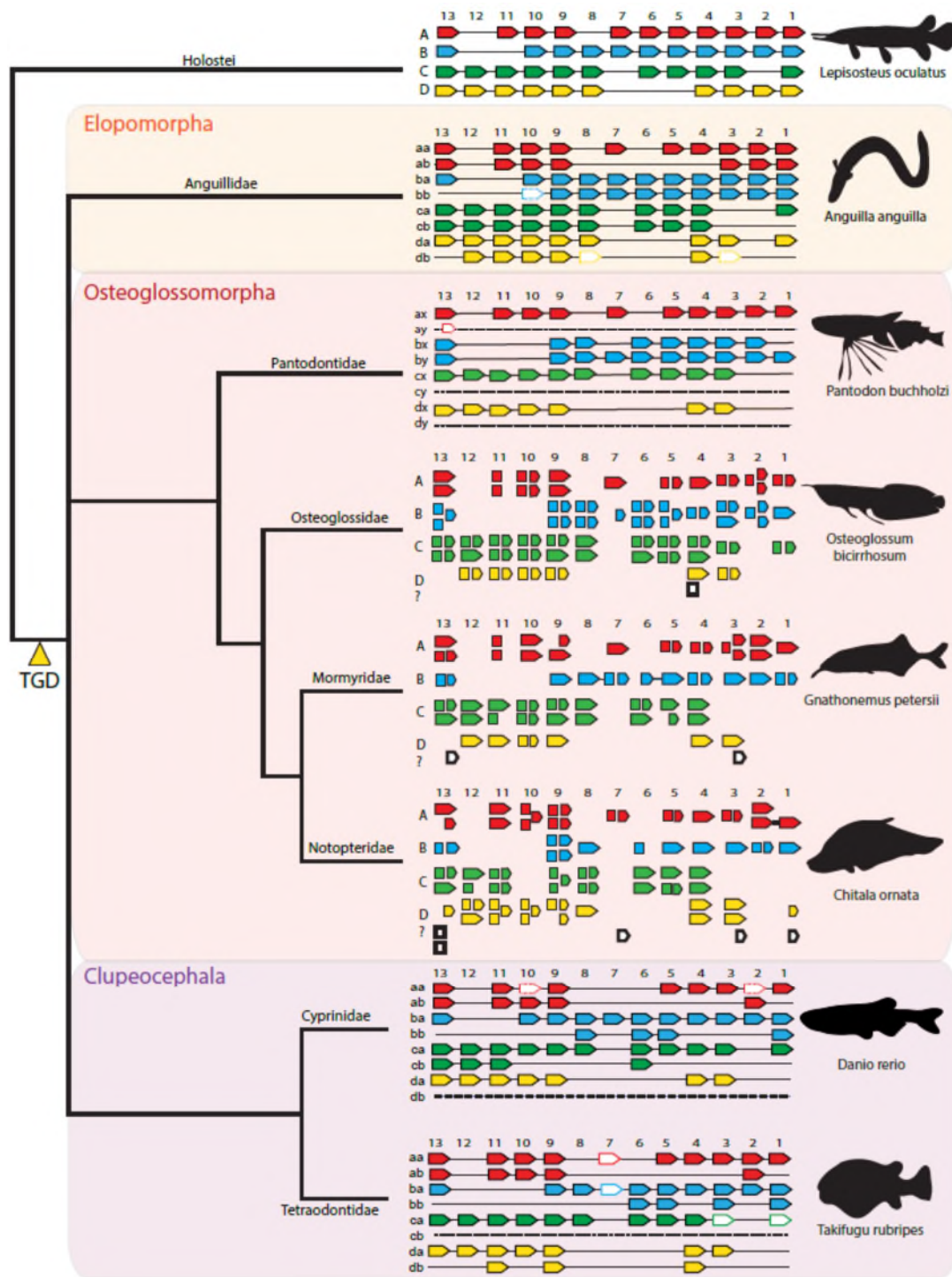


Figure1: Hox cluster complements in five osteoglossomorph families. Hox gene complements from *Osteoglossum*, *Gnathonemus*, and *Chitala* recovered in this study are compared to previously surveyed Hox gene complements from *Pantodon*, *Anguilla* and two clupeocephalans, and the outgroup spotted gar *Lepisosteus oculatus*, to give a first estimate of changes in Hox gene content across the Osteoglossomorpha. Hoxa clusters in red, Hoxb clusters in blue, Hoxc clusters in green, Hoxd clusters in yellow. Dashed lines indicate definitively missing clusters. Pseudogenes are shown as coloured unfilled boxes. Long arrows denote genes with both exons in one contig; shorter blocks or arrows denote exon 1 or exon 2 on a contig. Genes not assigned to particular groups are shown as unfilled boxes in the corresponding column of their most likely paralogy group.

Supplementary File 1: Table of osteoglossomorph Hox gene contigs giving DNA sequence and putative identity.

Supplementary File 2: Figure comparing Hox gene complements of two arowana species: Asian arowana *Scleropages formosus* (Bian et al., 2016) and Silver arowana *Osteoglossum bichirrhosum* (this study).

