Kyle Martin
Merton College

Title:

# Investigating the evolutionary impact of the teleost genome duplication through comparative genomics and phylogenetic analysis of homeobox genes in the Osteoglossomorpha

Submitted in partial fulfillment for the degree of DPhil

University of Oxford

Department of Zoology

Candidate Number: 415432

Submitted Michaelmas Term, 2015

Revisions Submitted Hilary Term, 2016

## Acknowledgements

First and foremost I would like to thank my supervisor, Professor Peter W. H. Holland for instilling in me a sense of wonder at the amazing diversity of life for and encouraging me to pursue DPhil in Evo-Devo with him at the University of Oxford only moments after delivering the keynote address at the Developmental Basis of Evolutionary Change symposium at the University of Chicago in 2005 when I was still only a wide-eyed and ignorant undergraduate student at McGill University. After an MSc. at the University of Ottawa and finally a DPhil at Oxford, I feel as wide-eyed and ignorant as I ever have but it is thanks to Peter's strong example and encouragement that the same sense of wonder that inspired me a decade ago continues to push me forward in my research career now. Peter taught me to 'think outside the box' and to explore theoretical models beyond those considered 'consensus' in the field, and animal models which less daring groups would consider out of reach. Peter also demonstrated throughout his career and certainly while I was a member of the lab that your research activities need not be confined by your past work or even your present expertise by simultaneously leading projects in choanoflagellate, lepidopteran, teleostean, mammalian, cephalochordate, amphibian, reptilian and human biology linking the overarching research theme of gene and whole genome duplications to the fields of microRNA evolution, gene family expansion, phylogenetics, reproductive biology and medicine, and morphological evolution. Peter gave me the rare freedom to explore my own interest even as a DPhil student and from this I developed (somewhat at least…) as an independent researcher.

So many other members of the Department of Zoology team contributed either directly or indirectly to my personal development and the work in this thesis and I will single out just a few. I want to thank Dr. Sebastian Shimeld for his participation as a member of my transfer committee and for his continuous advice and support during all stages of this DPhil during shared lab meetings between our groups. Dr. John Mulley contributed significantly to this work by proposing *Pantodon* as a model early on and developing the BAC library. Dr. Jordi Paps lent his advice and support for all phylogenetics related questions. Dr. Ferdinand Marlétaz was indispensable in his role as the administrator of our precious lab server 'paralog' and also for theoretical discussions in evolutionary genomics. Furthermore amongst many with whom I shared lab meetings or less formal discussions over coffees or beers I wish to acknowledge Dr. Ignacio Maeso, Dr. Thomas Butts, Dr. Jerome Hui, Dr. Laura Ferguson¸ Dr. Nathan Kenny and Dr. Cédric Patthey who all participated in helpful discussions relevant to this thesis and created an environment of exceptional diversity and learning that was refreshing and encouraging every day we were together whether in the lab or the Turf Tavern.

I wish to thank my parents Arleen and Raymond and my sister Kristin for their lifelong support and for enabling me to pursue my passion through their tremendous demonstration of strength and tenacity in the face of challenging circumstances. Without their selfless efforts during troubling times it is unlikely this work could have continued. Finally, I would like to thank my wife Miriam for her steadfast encouragement, warm affections and for giving me a home away from home in her heart. Without a doubt I could not have completed this endeavour without her.

This DPhil was funded jointly by the Natural Sciences and Engineering Research Council of Canada (NSERC), the University of Oxford Press Clarendon Fund, and I received supplementary financial aid from the Department of Zoology Elizabeth Hannah Jenkinson Fund and the Merton College Hardship Fund.

## Table of Contents


General Abstract ..... 1

Chapter 1: General Introduction ..... 4

Gene Duplication ..... 4

Homology, Orthology and Paralogy ..... 8

Phenetics and Phylogenomics in the era of 'Next-Generation' Genomics ..... 11

Whole Genome Duplication – Polyploidy ..... 15

Whole Genome Duplication – Palaeopolyploidy ..... 16

The Vertebrate Genome Duplications (VGD1/VGD2 ..... 17

The Teleost Genome Duplication (TGD ..... 17

Diploidization ..... 18

The Influence of the TGD on Teleost Evolution ..... 20

The Key Phylogenetic Position of Osteoglossomorpha ..... 24

Overall Aims and Objectives ..... 28

Figure 1.1 Whole genome duplications and the key phylogenetic position of Osteoglossomorpha ..... 30

Figure 1.2 Proposed phylogenetic relationships between the three major Teleost subdivisions ..... 31

Figure 1.3 Model osteoglossomorph taxa under investigation ..... 32

Chapter 2: Enigmatic orthology relationships between Hox clusters of the African butterfly fish and other teleosts following ancient whole genome duplication ..... 33

Abstract ..... 33

Introduction ..... 35

Materials and Methods ..... 39
Results ..... 53
Extensive Hox cluster loss in *Pantodon* ..... 53
Discordant Hox gene genealogies in *Pantodon* ..... 58
Detecting selection in Hox genes following the TGD ..... 63
Heterogeneous phylogenetic signal within concatenated Hox clusters ..... 65
Ancient CNE retention patterns in *Pantodon* Hox clusters ..... 67
Discussion ..... 69
Considering diploidization following the TGD ..... 71
Hox diploidization dynamics following the TGD ..... 74
Tetralogy: a new homology subtype ..... 76
Conclusion ..... 79
Table 2.1 454 BAC sequencing of candidate Hox cluster-containing BACs ..... 80
Table 2.2 Assembly of 454 sequencing results from candidate Hox cluster-Containing BACs ..... 81
Table 2.3 Test for positive selection on the stem of the teleost ‘a’ and ‘b’ paralogy groups ..... 82
Figure 2.1 The Hox gene complement of *Pantodon buchholzi* ..... 83
Figure 2.2 The Hox genes of *Pantodon buchholzi* do not segregate consistently with the post-TGD ‘a’ and ‘b’ orthology groups of other teleosts ..... 84
Figure 2.3 Individual Hox gene trees between *Pantodon buchholzi* and other vertebrates ..... 85-90
Figure 2.4 3D site-likelihood ratio plots of concatenated Hox clusters reveals strong cluster-wide conflict between alternative constrained hypotheses of *Pantodon buchholzi* Hox cluster homology. ..... 91
Figure 2.5 Ancient CNE conservation within Hox clusters ..... 92

Figure 2.6 The evolution of teleost *Hox* gene clusters following the TGD outlining the relative timing of diploidization events and the speciation of the major teleost subdivisions ..... 93

Figure 2.7 Types of homology relationships following WGD ..... 94-95

Chapter 3: Characterization of a *de novo* draft assembly of the genome of the osteoglossomorph fish *Pantodon buchholzi* ..... 96

Abstract ..... 96

Introduction ..... 98

Materials and Methods ..... 104

Results ..... 108

*P. buchholzi* genome assembly and quality assessment ..... 108

Single nucleotide variants in the *P. buchholzi* genome ..... 112

Transposable element content in the *P. buchholzi* genome ..... 112

Transposable element diversity in the *P. buchholzi* genome ..... 117

Transposable element content in *Pantodon* Hox clusters ..... 122

Genome-wide protein coding gene predictions ..... 119

Discussion ..... 128

Conclusions ..... 133

Table 3.1 *P. buchholzi* genome assembly metrics ..... 134

Table 3.2 *P. buchholzi* genome assembly version Pbuc_v37 read mapping statistics ..... 135

Table 3.3. Transposable element content in the *P. buchholzi* genome ..... 136

Table 3.4 Transposable element content in the Hox clusters of *P. buchholzi* ..... 137

Figure 3.1 kmer depth plot and genome size estimation ..... 138

Figure 3.2 *P. buchholzi* genome assembly Pbuc_v37 contig length and GC content ..... 139

Figure 3.3 Proportion of transposable elements in the genome of *P. buchholzi* and other chordates ..... 140

Figure 3.4 Diversity of transposable elements in the genome of *P. buchholzi* and other chordates ..... 141

Figure 3.5 Transposable elements are excluded from the *P. buchholzi* Hox clusters ..... 142

Figure 3.6 Transcriptome assembly of *P. buchholzi* brain and muscle total RNA ..... 143

Figure 3.7 Reference genome BLAST analysis of the Pbuc_v37 final MAKER gene model set ..... 144

Chapter 4: Ancient duplicated ANTP homeobox gene paralogs in the genome of *Pantodon buchholzi* reveals delayed gene loss following the teleost genome duplication ..... 145

Abstract ..... 145

Introduction ..... 147

Materials and Methods ..... 150

Results ..... 153

ANTP homeobox gene annotation ..... 153

Phylogenetic analysis and divergent resolution of ANTP class homeobox genes ..... 156

Ancestral state reconstructions and rates of gene loss ..... 161

Discussion ..... 164

Conclusions ..... 167

Table 4.1 Classification of all ANTP proteins in *P. buchholzi,* human, gar, and zebrafish ..... 168-171

Figure 4.1 TGD paralogs retained only in *P. buchholzi* and lost in clupeocephalan teleosts ..... 172-179

Figure 4.2 TGD paralogs retained by both *P. buchholzi* and clupeocephalan

teleosts ..... 180-182

Figure 4.3 Evolution of ANTP class homeobox genes in teleosts after the TGD ..... ……183-184

Figure 4.4 Relative rates of ANTP class homeobox gene paralog loss along key branches following the TGD..... 185

Chapter 5: A bioinformatic survey of Hox genes in the draft genome sequences of three osteoglossomorph teleosts reveals extensive divergent resolution of Hox clusters following the teleost genome duplication..... 186

Abstract ..... 186

Introduction..... 188

Materials and Methods ..... 193

Results ..... 196

Genome assembly metrics ..... 196

Fishing for Hox in fish: the next-generation ..... 198

Ancestral state reconstruction ..... 201

Discussion ..... 206

Conclusions ..... 212

Table 5.1 Osteoglossomorph genome assembly metrics ..... 213

Figure 5.1 Hox cluster complements in five osteoglossomorph families ..... 214-215

Figure 5.2 Reconstruction of ancestral Hox complements ..... 216

Figure 5.3 Rate of Hox gene losses mapped onto basal teleost phylogeny ..... 217

Chapter 6: General Discussion..... 218

Appendix 1: Publication of chapter 2 “Enigmatic orthology relationships between Hox clusters of African butterfly fish and other teleosts following ancient whole-genome duplication” ..... 227-247

Appendix 2: Contributions to published manuscript “Linking human biology to fish models: Sequence of the spotted gar genome” ..... 248-253

References ..... 254-278

## General Abstract

Multiple rounds of whole genome duplication (WGD) have played a pivotal role in the expansion, elaboration, and evolutionary diversification of vertebrate genomes. In addition to sharing two rounds of whole genome duplication with all other vertebrates, a teleost-specific genome duplication (TGD) occurred in the stem of the teleost lineage ~350 million years ago (MYA) and is thus a genomic synapomorphy shared by all ~26,000 extant species. The TGD has variously been implicated in accelerated speciation, evolution of morphological complexity, increased rates of molecular evolution, and the evolution of novelty, and therefore is therefore of significant interest for its impact on teleost evolution and also as a model for understanding the evolutionary patterns and processes which accompany WGDs more generally. Investigation of the TGD has contributed extensively to the general understanding of WGDs however, until the present work, a relatively narrow taxonomic sampling of species within a single teleost subdivision, Clupeocephala, have been investigated. This taxonomic bias has left potentially relevant evolutionary changes to the teleost genome in the immediate wake of the TGD obscured. Due to their deeply branching ancestry, species belonging to the two other major teleost subdivisions, Osteoglossomorpha and Elopomorpha, are well positioned for deeper comparative genomic analyses of the TGD and the accompanying phenomenon of diploidization. The focus of the present work has been to develop the first genomic resources specifically for osteoglossomorphs and to investigate the evolutionary patterns and processes which accompanied diploidization prior the deep divergence of the three extant teleost subdivisions. To this end, I have generated *de novo* genome and

transcriptome data from four osteoglossomorph taxa (*Pantodon buchholzi*, *Osteoglossum bicirrhosum*, *Chitala ornata*, and *Gnathonemus petersii*) and conducted comparative genomic and phylogenetic analysis with other teleosts and pre-TGD vertebrates including the gar *Lepisosteus oculeatus*. With a focus on Hox and other ANTP class homeobox-containing transcription factor families I provide evidence that speciation of the major teleost subdivisions occurred prior to the termination of the diploidization process following TGD and discuss the evolutionary implications of this model. Beginning with an analysis of the Hox clusters in *P. buchholzi* I show that divergent resolution of TGD-generated Hox duplicates occurred both at the individual gene level as well as at the level of whole cluster losses. Detailed phylogenetic analyses of the *P. buchholzi* Hox clusters further revealed that the transition from polyploid alleles to full paralogs during the diploidization process can occur independently in different lineages when speciation rapidly follows WGDs, causing duplicated genes to exhibit a special case of four-way gene homology which I have termed 'tetralogy'. A genome-wide survey of ANTP class homeobox genes in a *de novo* assembly of the *P. buchholzi* genome revealed that ancient TGD duplicates of at least 14 subfamilies were preserved uniquely in the *P. buchholzi* genome and lost from clupeocephalan teleosts. Finally, by comparing the Hox complements in gar and *P. buchholzi* with three additional osteoglossomorphs I show that the diversity in potential duplicate resolution patterns is also highly variable between osteoglossomorph families. Overall, this work highlights the importance of considering not only the relative timing of gene duplication and speciation in comparative genomic analyses but also their timing relative to diploidization. Going forward, the research community will need to carefully

evaluate the effects differences in diploidization rate and pattern, both between lineages and across the genome, have had in influencing the fate of individual gene duplicates as well as upon the macroevolutionary phenomena frequently correlated with WGDs more generally.

# Chapter 1

## General Introduction

Here, I attempt to give a very brief summary of some of the theoretical and experimental background relevant to this work suitable for reading as a primer to the field of teleost molecular evolution. Subsequent chapters are each in turn accompanied by their own more specialized subject-specific introductory materials.

### Gene Duplication

Gene duplication is today considered to be a major mechanism for the generation of novel genetic substrates upon which evolutionary forces including neutral genetic drift and directional selection may act, potentially driving an increase in genomic complexity through the evolution of modularity (Wagner 2007) and resulting in novel and diverse phenotypic outcomes in all three of life's domains (Bacteria, Archaea, Eukarya) (Zhang 2003; Conant & Wolfe 2008). However, it was not until 1970 when the now classic book "Evolution by Gene Duplication" was published by Susumo Ohno (Ohno 1970), that the large potential for impact and the widespread occurrence of gene duplication, particularly in vertebrates, was widely acknowledged by the research community. It was Ohno's contention that gene duplication could facilitate a degree of macroevolutionary change beyond the range of transformations which could occur through mutations in existing genes alone and he argued that all new genes arose through duplication (Ohno 1970).

Despite this potential, empirical research and modelling have shown that the most common long-term fate for a new gene duplicate is nonfunctionalization (silencing followed by loss though either deletion or pseudogenization) of one member of a new gene pair (Maere et al. 2005; Inoue et al. 2015; Sankoff et al. 2010) and it is thought that the average time gene duplicates survive in the genome prior to the loss of one member of the pair is only ~4 million years in animals (Lynch 2000). Importantly however, where new duplicates are not lost they have the potential to drive evolution in significant ways by contributing novel genetic material and relaxing functional constraints such as pleiotropy by creating genetic redundancy (Prince & Pickett 2002).

If simple genetic redundancy was the only result of gene duplication, nonfunctionalization through neutral genetic drift would be far and away the most plausible expected outcome, however the observation that a significant proportion of eukaryotic genomes (8-20% in the model fruitfly *Drosophila melanogaster*, nematode *Caenorhabditis elegans* and yeast *Saccharomyces cervisiae* and approximately 15% in humans) (Lynch 2000; Lynch & Conery 2003b) are still comprised of readily recognizable gene duplicates suggests that other forces are at play which may drive their long-term retention. Indeed ,Lynch and Conery first calculated that the fixed duplication rate (calculated as a factor of both the mutation and fixation rates) was in the order of 0.01 per gene per million years in many eukaryotes which is comparable to the single nucleotide substitution rate of 0.001-0.005 per site per million years in vertebrates (Lynch & Conery 2003a) and this result has largely been upheld by more recent studies (reviewed in Demuth & Hahn 2009) .

A variety of models have been proposed to explain retention of both members of a duplicated gene pair (reviewed extensively in Innan & Kondrashov 2010). While a wide diversity of subtly, but importantly, different models have been proposed they all largely fit into three principle categories: (1) neofunctionalization, (2) subfunctionalization, and (3) dosage balance. Neofunctionalization of one of the duplicates, that is a gain of a new function by one of the duplicates, was first suggested by Ohno as the single most significant factor driving duplicate gene retention and requires one of the duplicates to retain the full ancestral function of the parent gene while the other duplicate diverges (Ohno 1970). Subfunctionalization implies changes to both duplicates and involves a splitting of ancestral subfunctions originally performed by the single parental gene between the two new duplicates as outlined in the duplication-degeneration-complementation (DDC) model (Force et al. 1999; Postlethwait et al. 2004). Dosage/gene-balance type models are diverse (Edger & Pires 2009) but generally suggest that both duplicates can be retained either where there is a beneficial effect of retaining both copies such as selection for higher expression levels, or a deleterious effect of reducing expression levels from one member of an interacting group of genes which may occur when complex gene regulatory networks including positive and negative feedback loops are involved. This latter model is particularly relevant where multiple genes are duplicated simultaneously including large segmental duplications and whole genome duplications (WGDs).

Neofunctionalization may occur following either an accumulation of initially neutral mutations permitted through a relaxation of negative selective pressures due to redundancy, followed by a change in conditions such that the originally neutral

mutations lead to a new function, or through the accumulation and fixation of mutations through positive selection for new beneficial functions (Hahn 2009). Historically, neofunctionalization has been difficult to demonstrate experimentally as it requires precise knowledge of the ancestral function of a gene prior to duplication, something which is not always available especially when no closely related unduplicated outgroups are available. Nevertheless some interesting cases have been described including an apparent case of convergent neofunctionalization of a sodium channel gene (*scn4aa*) in two independent lineages of electrogenic fishes during the evolution of their electric organs (Zakon 2012; Thompson et al. 2014; Arnegard et al. 2010). In both Mormyrid and Gymnotiform electric fish duplicated paralogous copies (*scn4aa*/*scn4ab*) generated via duplication of a single ancestral *scn4a* gene during WGD, originally expressed in muscle diverged in both their coding sequences, with a signature of positive selection, and in their expression domains such that *scn4aa* became adapted for and localized to the unique electrophysiological profiles of the novel muscle-derived electric organ, while *scn4ab* maintained the original functions of the ancestral *scna4* gene in muscle.

Neofunctionalization may also occur through the evolution of novel protein domains. On a genome-wide scale, one study has indicated that novel protein domains have evolved in up to a quarter of paralogous duplicates in teleost fishes following the teleost-specific genome duplication (TGD), a rate which exceeded the acquisition rate in singletons (Kassahn et al. 2009).

Subfunctionalization may occur at the level of expression domains, timing of expression, or biochemical activities. The seminal example of subfunctionalization is a case of regulatory subfunctionalization and partitioning of expression domains which

occurred following the duplication of the engrailed-1 (*eng1*) gene in zebrafish with the TGD (Force et al. 1999). In the mouse and chicken *Eng1* is expressed in both the forelimbs and in a specific set of hindbrain interneurons (Joyner & Martin 1987; Gardner & Barald 1992), while following duplication in teleosts, the zebrafish co-orthologs *eng1a* and *eng1b* partitioned these two expression domains between them, with eng1a becoming restricted to the forelimbs, while *eng1b* became confined to the hindbrain interneurons (Force et al. 1999).

**Homology, Orthology and Paralogy**

The concept of homology is one of the most fundamental and powerful constructs in modern evolutionary theory. While a comprehensive review is beyond the scope of this thesis, a brief outline of some key categories of homology and their definitions should help to contextualize some of the key findings presented later. Richard Owen coined the term 'homology' in the 19$^{th}$ century and defined it as "the same organ in different animals under every variety of form and function" (Owen 1843). In lay terms, homology describes a specific type of biological 'sameness' which is a result of descent from a common evolutionary ancestor. This contrasts with homoplasy which describes biological 'sameness' as a result of independent origins but convergent processes. Prior to the revolution in evolutionary theory brought about by the reunion of genetics, development and evolution under the umbrella of Evolutionary Developmental Biology, or "Evo-Devo" (Raff 2007; Carroll 2008), the concept of homology was largely confined to as strictly historical, anatomical, interpretation *sensu* Owen 1843. Today in

addition to historical homology of anatomical structures, evolutionary biologists have expanded the framework of homology to encompass many different levels of biological organization all the way from individual base pairs or amino acids, to genes, gene regulatory networks (GRNs), tissues, developmental mechanisms and ultimately the expressed phenotype including, but not limited to, anatomy (Wagner 2007). As each of these levels of biological organization are intrinsically interrelated, so too the concept of homology as applied across these levels can be viewed in a partially nested, hierarchical organizational structure (Abouheif 1997). Ultimately, the goal of any homology concept at is to guide the orientation and interpretation of biological transformations which accompany the process of evolution in a helpful organizational framework. In the context of this thesis, the most relevant level of biological organization to which the homology concept applies is the unit of the gene or molecular homology more broadly.

The vast majority of protein-coding genes can be organized into multigene families whose history has been shaped by successive duplication, loss and speciation (linear descent with modification) events (Demuth & Hahn 2009; Dayhoff 1976; Nei & Rooney 2005). When comparing genes either within individual genomes or between species, genes are classified by sub-categories of molecular homology first introduced by Walter Fitch over 40 years ago: orthology (the relationship between genes present in a common ancestor and separated through a speciation event) or paralogy (the relationship between sister genes generated through a gene duplication event) (Fitch 1970). Because gene duplication and speciation have both been pervasive throughout evolutionary history, no genes can be considered absolute ‘orthologs’ or ‘paralogs’ in all

species across all levels of comparison phylogenetic tree. A single gene may be simultaneously considered a paralog relative to a duplicate within the very same genome but an ortholog relative to a gene in other species which shared a direct line of descent with modification where speciation followed a shared gene duplication event. Additional levels of sub-classification may be applied to both orthology and paralogy concepts in order to qualify the level of the homology relationships relative to both duplication and speciation events between lineages. Co-orthology describes a condition where two or more genes in one species are equally related to one gene in another species as a result of lineage-specific gene duplication. Furthermore, when comparing genes between species which have all shared a gene duplication event prior to their radiation the term outparalogs may be used, while lineage-specific gene duplication events relative to a particular speciation event may be considered inparalogs (Koonin 2005; Hillis 1994; Sonnhammer & Koonin 2002). Additional molecular evolutionary processes beyond duplication, speciation, and loss which shape the history of gene families and do not fit neatly into the above categories of gene homology include horizontal gene transfer (HGT) where the appearance of novel sequences in a particular lineage actually the result of non-linear transfer across a species barrier (hence xenologs) and rearrangements (fusion/fission/exon-shuffling) between genes as a result of concerted evolution or transposition which can result in a complex combination of homology relationships at the level of individual domains within a gene (Koonin 2005). Recent genomic studies have also revealed several cases of *de novo* evolution of genes for example via the lineage specific expansion and co-option of transposable elements into protein-coding sequences (Zhou et al. 2008; Abascal et al.

2015). In this case outside of the lineage where the novel co-option event occurred it would not be possible to define the level of genetic homology with other sequences.

Some authors have also proposed a specific sub-type of homology to describe sequences related by duplication through WGD events: ohnologs (Postlethwait 2007; Makino & McLysaght 2010; Wolfe 2000). Named after Susumo Ohno who. more than 40 years ago, first proposed that WGD events occurred in the vertebrate lineage based almost entirely on allozyme data (Ohno et al. 1968; Ohno 1970), the term ohnologs encompasses the complete set of duplicated homologous sequences within a lineage which has shared an ancestral WGD event (i.e. out-paralogs generated via WGD). However, for the purpose of this thesis the term ohnolog will not be adopted as it is less specific than orthology and paralogy and does not define a distinct level of biological organization but rather simply reflects the mechanism of gene duplication. Nevertheless, as the range of known types mutational processes affecting genome evolution increases and more refined models accurately reflecting the evolutionary history of individual genes subjected to these mutations are developed, the concept of homology may also need to adapt in order to reflect novel levels of biological organization and continue to facilitate the orientation and interpretation of evolutionary changes.

**Phenetics and Phylogenetics in the era of 'Next-Generation' Genomics**

Just as comparative analyses at an organismal level cannot be carried out outside the context of a phylogenetic framework, comparative genomic analyses require a phylogenetic reconstruction of the history of individual genes in order to be biologically

meaningful. Establishing correct gene homology relationships is the first step in this process but is far from trivial and requires careful phylogenetic, rather than simple phenetic evaluation (Thornton & DeSalle 2000). Phenetic approaches to classifying homologs rely on a single quantitative measure of pairwise sequence similarity (derived from either a DNA or amino-acid multi-sequence alignment) which is converted to a distance measurement in a matrix. The Basic Local Alignment Search Tool (BLAST) (Altschul et al. 1990) is perhaps the most widespread bioinformatic algorithm in use for identifying homologous sequences in different genomes and distance matrices built using the reciprocal-best-hit (RBH) scores in all-vs-all comparisons of protein sequences from diverse genomes and forms the basis of defining so-called 'Clusters of Orthologous Groups' (COGs) (Tatusov et al. 2000). As the availability of genomic data has increased, the desire for rapidly classifying newly sequenced genes according to their homology with previously known families has spawned a wave of new high-throughput phenetic classification tools including those that, like COGs, rely on BLAST scores and others which take advantage of newer MCL (Markov Cluster) algorithms (Tekaia 2016). However, while powerful in the sense of computational efficiency and therefore range and scale of potential applications, phenetic methods often fail to accurately reconstruct homology down to the level of orthology and paralogy between genes with complex histories of duplication and speciation. This is because by collapsing all of the information contained in a multigene alignment into a single quantitative measure (pairwise distance) potentially informative phylogenetic information is lost and the only criterion with which to evaluate homology which remains is overall similarity score. Reconstruction of homology relationships with phenetic tools therefore

relies solely on the definitively false assumption of equal rates of molecular evolution between different genes and different lineages (Echave et al. 2016; Ho & Duchêne 2014) and simply posits that the most similar sequences are the most closely related. It is for this reason that phylogenetics has moved beyond distance-based methods in favour of parsimony, likelihood, and Bayesian approaches to more accurately estimate both species and gene trees (Yang & Rannala 2012). While genome-wide approaches for clustering homologous sequences employing these phylogenetic principles are currently being developed (Tekaia 2016) for the time being it is necessary to resort to more laborious classical but ultimately reliable and accurate phylogenetic reconstructions of individual gene trees in order evaluate homology at the biologically informative levels of orthology and paralogy. This is especially true where the overall aim is to develop an understanding of the patterns and processes associated with the phenomenon of gene and WGD as in the case of this thesis.

Distinguishing between orthologous and paralogous genes is not simply a theoretical exercise and has real biological significance both for understanding the impact of gene and WGD to genome evolution but also for related disciplines which require accurate reconstructions of species-level phylogenies. For example, only orthologous genes related through direct vertical descent from a common ancestor can be used to infer species-level relationships using phylogenetic methods. Mistakenly comparing paralogous sequences in a multisequence alignment in order to reconstruct a species tree would lead to the assumption that synapomorphies (fixed substitutions) which accumulated between them reflect the evolutionary history of the species rather than the duplication history of the genes and result in an erroneous species tree. This

becomes especially difficult when patterns of reciprocal gene loss in different lineages following WGDs confound assessment of orthology and paralogy causing paralogs to appear as orthologs in a gene tree, a phenomenon which has been described as 'hidden paralogy'(Kuraku 2010).

The level of homology between two genes may also bear on the degree of their functional evolution. The classical view is that orthologs tend to be more similar in function than paralogs, which are predisposed to diverge following duplication (Dolinski & Botstein 2007; Koonin 2005). However, the generality of this phenomenon, recently dubbed the 'orthology-conjecture', has been called into question (Gabaldón & Koonin 2013; Studer & Robinson-Rechavi 2009; Nehrt et al. 2011). Clearly, any hypotheses correlating orthology, paralogy and functional evolution will require reliable homology relationships to be drawn before the functional comparisons to test them can possibly be made.

With the explosion in the availability of both genome and transcriptome sequence data, largely facilitated by the advent of 'Next-Generation' sequencing technologies (Metzker 2009; Goodwin et al. 2016), both the demand for and the importance of accurate homology assessment has grown. This is especially true in lineages where WGD events have occurred, resulting in the simultaneous duplication of all loci. Special consideration must be paid to genes which duplicated via a WGD beyond the scope of the more general evolutionary patterns and processes experienced following single-gene duplications due both to the large scale of the phenomenon and the potential non-independence of duplicated loci, both relative to their physical positions on a chromosome as well as their functional niche. Some of these challenges are addressed

in the following chapters. Despite these additional considerations, the range of eventual outcomes or fates which await duplicated genes, whether they have arisen via single gene duplications or WGD, are identical and it is only their relative proportions and perhaps the rate at which they occur which varies depending on the method of duplication.

**Whole Genome Duplication - Polyploidy**

The high prevalence of extant polyploids suggests that WGD is both a pervasive and contemporary influence in animal genome evolution (Otto 2007; Leitch et al. 2004; Otto & Whitton 2000). In comparison to plants, polyploidization is rare in the animal kingdom, but it is also probably true that our current counts are an underestimate. Within the vertebrate lineage, polyploids are most often found among amphibians and actinopterygian fishes (Mable et al. 2011; Otto 2007; Comai 2005; Braasch & Postlethwait 2012). Rather than an increased propensity for polyploidy in amphibians and fishes it may be that the apparently low rate of polyploidy in vertebrates in general is a bias introduced by the paucity of fixed polyploidization events in amniotes. While some authors have argued that WGD is prohibitive in mammals due to their precisely dosage-balanced XX/XY sex determination mechanism an intriguing case of apparent mammalian polyploidy has been described in the plains viscacha rat *Tympanoctomys barerae* (Gallardo et al. 2006). Furthermore, it seems that the tolerance of polyploidization is a heritable trait in mouse and the developmental stage to which an artificially tetraploidized mouse embryo can progress before aborting is significantly different in different laboratory strains (Eakin & Behringer 2003). Polyploidy has arisen

multiple times across the full spectrum of taxonomic levels during the evolution of fishes. In some cases, stable populations of polyploids exist alongside diploid populations of the same species and retain the capacity for genetic exchange as in the pond loach *Misgurnus anguillicaudatus* (Yang et al. 2007; Li et al. 2011). In other cases the WGD resulting in the production of the polyploid lineage occurred along deeper branches of the phylogenetic tree and entire families may be polyploid. Salmonidae (Allendorf & Thorgaard 1984), Catostomidae (Uyeno & Smith 1972), Callichthyidae (Oliveira et al. 1993) and members of the Cyprinidae including common carp, goldfish and their various domesticated varieties (Larhammar & Risinger 1994) also share a deep polyploid ancestry. The salmonid-specific genome duplication event (SSGD, or 4R), has been timed to no later than 88 Mya (Macqueen & Johnston 2014). Recently, whole genome sequencing of two salmonids – the rainbow trout *Oncorhynchus mykiss* (Berthelot et al. 2014) and the Atlantic salmon *Salmo salar* (Lien et al. 2016) has yielded an unprecedented view of the evolutionary changes to the genome structure and gene content which follow polyploidization and confirmed the hypothesis from earlier karyotypic analyses (Allendorf & Danzmann 1997) that the salmonid genome is 'pseudotetraploid' with most of the genome behaving in a diploid fashion while around 10% still exhibits characteristic signatures of polyploidy (Lien et al. 2016).

**Whole Genome Duplication - Palaeopolyploidy**

Even more ancient WGD events have occurred across eukaryotic evolution. In the present day, these lineages are considered palaeopolyploids (reviewed in (Van de Peer et al. 2009)). Lineages which have experienced ancient WGDs include most

eudicots (Jaillon et al. 2007), *Arabidopsis* (Simillion et al. 2002), *Saccromyces* (Wolfe & Shields 1997), *Paramecium* (Aury et al. 2006) and all vertebrates (Putnam et al. 2008; Furlong & Holland 2002; Sidow 1996; Li et al. 2001).

**The Vertebrate Genome Duplications (VGD1/VGD2)**

Within vertebrates two rounds of WGD are thought to have occurred in the stem lineage ~525 Mya around the time of the divergence between the living jawless (agnathan) fishes and jawed (gnathostomate) vertebrates (Figure 1.1; node 3); these are known alternately as 'Vertebrate Genome Duplication 1 and 2' (VGD1/VGD2) or the 1$^{st}$ and 2$^{nd}$ Rounds of WGD (1R/2R) (Holland et al. 1994; Li et al. 2001; Kuraku 2008; Putnam et al. 2008; Van de Peer et al. 2010; Smith et al. 2013). For the purpose of this thesis I will use the terms VGD1/VGD2 as they convey both a taxonomic and sequential meaning as opposed to 1R/2R which carry no taxonomic information about the lineages in which they have occurred. The VGD1/VGD2 duplications were pivotal in the expansion of developmental gene families; following WGD there was a preferential retention of duplicates from this functional group, while most other gene families reverted to single copy (Freeling and Thomas 2006; Putnam et al. 2008; Edger and Pires 2009; Makino and McLysaght 2010).

**The Teleost Genome Duplication (TGD)**

It is now also well established that the common ancestor of all teleost fish experienced an additional ancient WGD event beyond VGD1/VGD2 approximately 226-350 million years ago, depending on differences in the data and calibration used in the

molecular clock analysis. This event, mapped to the cladeogram in Figure 1.1, is known alternately as the Teleost Genome Duplication (TGD) or 3rd Round of WGD (3R) (Jaillon et al. 2004; Meyer & Van de Peer 2005; Amores et al. 1998; Taylor et al. 2003; Hoegg et al. 2004). For the purpose of this thesis the term TGD will be given preference over 3R as simple enumeration of WGD events is unhelpful in cases where multiple vertebrate lineages have experienced independent rounds of WGD including other non-teleost actinopterygians (i.e. American paddlefishes (Crow et al. 2012)).

**Diploidization**

Following any polyploidization event whether ancient or recent the inevitable outcome is that the genome transitions from a karyotypically tetraploid state to a karyotypically diploid state, a phenomenon known as diploidization (Wolfe 2001). The types of complex genetic interactions between duplicated homeologous loci which are possible after polyploidization is best documented in plants, where allopolyploidy in particular is a pervasive phenomenon (Doyle & Egan 2010; Wendel 2000; Otto & Whitton 2000). The idea that sequence similarity between duplicated homeologous chromosomes might linger in allopolyploid plants causing tetrasomic inheritance was first proposed over 50 years ago (Stebbins 1947). Recent work in plants including artificially resynthesized allopolyploid lines in the genus *Brassica* (Gaeta et al. 2007), a natural assemblage of very recently (<80 yrs) formed allopolyploid species in the genus *Tragopogon* (Chester et al. 2012) and older (1-2 Myr) allopolyploid lineages such as the cotton *Gossypium hirsutum* (Salmon et al. 2010), have revealed molecular evidence for extensive genetic exchange between duplicated homeologous loci. In contrast to

autopolyploidy where intraspecific genome doubling provides ample opportunity for tetrasomic inheritance, allopolyploidy comes in two categories which differ in the extent of genetic exchange expected between homeologous chromosomes: “Genomic” and “segmental” allopolyploidy (Stebbins 1971). Genomic allopolyploids result from the interspecific hybridization of two highly divergent parental species genomes causing homeologous chromosomes to exhibit almost immediate bivalent segregation and disomic inheritance. In these allopolyploids the genome is effectively pre-diploidized and the coalescent time of duplicated genes predates the date of hybrid formation, and instead corresponds to the speciation date of the parental species. Segmental allopolyploids, however, result from hybridization where only a fraction of the parental genomes is divergent enough for immediate bivalent formation, and other regions continue to exhibit tetrasomic inheritance and eventually must diploidize. Maize (*Zea mays*) which experienced a polyploidization event ~11.4 Myr ago seems to be a segmental allotetraploid in origin (Gaut & Doebley 1997). Much less is known about the process of diploidization after ancient WGD in animal lineages, especially in vertebrates. Tetravalent formation during meiosis, which favours homeologous recombination and gene conversion, has been directly observed in recently polyploidized lineages including autotetraploid frogs (*Odontophrynus americanus*) (Becak et al. 1966) and allotetraploid loaches (*Misgurnus anguillicaudatus*) (Li et al. 2011). Interestingly, salmonid teleosts, which underwent an additional independent WGD (sometimes called 4R) after the TGD ~95 Mya (Allendorf & Thorgaard 1984; Macqueen & Johnston 2014), still do not seem to have completed the diploidization of their genomes (Lien et al. 2016). There is evidence in this lineage for ongoing

recombination between duplicated loci (Johnson et al. 1987; Allendorf and Danzmann 1997; Gharbi et al. 2006) and for occasional tetravalent formation during meiosis (Lee & Wright 1981; Allendorf & Thorgaard 1984), suggesting that the diploidization process can take a significant period of time to complete.

**The Influence of the TGD on Teleost Evolution**

WGD through polyploidization results in the simultaneous duplication of all genes and is a significant contributor to the expansion of gene families in vertebrates, and is even considered a potential driving force for their long term diversification and evolutionary success (Van de Peer et al. 2009; Braasch & Postlethwait 2012; Meyer & Schartl 1999; Holland et al. 1994).

Teleosts fishes comprise over half of all vertebrate biodiversity (Nelson 2006). Because of this, and their outstanding morphological complexity, there has been a longstanding interest in whether there is a causal relationship between the TGD and their morphological complexity and species richness (Meyer & Van de Peer 2005; Meyer & Schartl 1999). It has been proposed that gene-balanced duplications such as WGDs can drive the evolution of morphological complexity by the preferential retention of dosage-sensitive developmental genes (Freeling & Thomas 2006; Edger & Pires 2009). Correspondingly, it has also been suggested that reciprocal paralog loss following a WGD between different populations, leading to hybrid dysfunction, or sterility can promote speciation through the creation of genetic barriers (Presgraves 2010; Lynch & Force 2000) (Lynch & Force 2000; Taylor et al. 2001). It has been suggested that divergent resolution can occur both at the level of reciprocal non-functionalization

(gene-loss) between lineages or in the reciprocal subfunctionalization (partitioning of parental gene functions between two daughter genes) for example through the complementary degeneration of *cis*-regulatory elements in accordance with the duplication-degeneration-complementation (DDC) model (Force et al. 1999; Postlethwait et al. 2004). Previous studies have investigated the degree of divergent resolution which has occurred within clupeocephans, the largest teleost subdivision consisting of ostarioclupeomorphs (zebrafish, carp, cavefish etc.) and euteleosts (medaka, cichlids, pufferfishes etc.), and have shown that there was indeed abundant genome-wide reciprocal gene loss between two of the most speciose superorders, the Ostariophysii and the Percomorpha (Sémon & Wolfe 2007; Garcia de la Serrana et al. 2014). For example, the differential retention of gene duplicates between zebrafish and *Tetraodon* has been proposed as a potential mechanism for their reproductive isolation (Sémon & Wolfe 2007). However, it should be noted that the zebrafish and *Tetraodon* lineages diverged >100 million years following the TGD, and a thorough analysis of the fossil record suggests there was a significant time lag of approximately 150 million years between the origin of teleosts and the radiation of the most speciose lineages (Santini et al. 2009). Additionally, while today actinopterygian diversity is dominated by teleosts (comprising some 99% of extant species) diversity of so-called 'ancient fishes' was much greater and the present-day dominance of teleosts may be partially as a result of an extinction bias rather than increased speciation. (Donoghue & Purnell 2005; Hurley et al. 2007). Importantly, while the TGD is unambiguously a synapomorphy of all teleosts (Taylor et al. 2003; Henkel et al. 2012; Martin & Holland 2014), species richness is not evenly spread across the teleost tree. Furthermore, an analysis of

diversification rates in teleosts estimated that an acceleration in speciation rate on the branch between the last neopterygian common ancestor, and the last teleost common ancestor (the interval during which the TGD is believed to have occurred) could be responsible for as little as 10% of teleost diversity (Santini et al. 2009). This study concluded that the two most significant increases in speciation rates in the Ostariophysii (~9,000 species) and Percomorpha (~16,000 species), were responsible for generating ~88% of extant diversity, occurred significantly later in the ostariophysian and percomorph lineages (Santini et al. 2009). A parallel situation is observed in salmonids following the 4R WGD event. Although the high diversity of extant salmonids may at first appear to be correlated with the 4R WGD, a well constrained time tree analysis using mitochondrial data and a carefully selected set of WGD-generated paralogues revealed that although the 4R WGD occurred 88-105 My ago, most extant salmonid lineages only arose in the last 10 My revealing an extended time lag (Macqueen & Johnston 2014). It therefore appears plausible that if there was indeed an effect of the TGD on teleost speciation rate, the most significant increases only occurred after the divergence of the three major teleost subdivisions and not concomitant with the TGD itself in their common ancestor. Reciprocal loss of TGD-generated duplicates has also occurred much more recently in evolutionary history. There is also evidence of ongoing reciprocal gene loss of Hox genes within the percomorph family Cichlidae (Thomas-Chollier & Ledent 2008).

Evidence suggests that retention of gene duplicates following a WGD is more common than after tandem, segmental, or retrotransposition type duplications (Nadeau & Sankoff 1997). Even after the ~350 million years since the TGD and extant lineages

have unambiguously returned to a diploid karyotype, it has been observed that between 12%-24% of ancient TGD duplicates are retained in teleost genomes (Kassahn et al. 2009; Jaillon et al. 2004; Brunet et al. 2006; Postlethwait 2000). Consistent with the gene-dosage balance model of duplicate retention, and like the older vertebrate duplications VGD1/VGD2, there appears to be a bias for genes belonging to certain functional classes, including transcription factors and other developmental genes, to be retained after the TGD (Kassahn et al. 2009; Putnam et al. 2008; Brunet et al. 2006). There is even evidence which suggests that in humans, the functional class of genes biased towards retention following VGD1/VGD2 are in fact refractory to other types of duplication (Makino & McLysaght 2010). This suggests that WGD is the principal driving force behind the expansion of transcription factor and other developmental gene family sizes in vertebrates.

At the core of both hypotheses linking the TGD to either morphological complexity or species diversity is the timing and pattern of gene loss during the ensuing diploidization process. The most prevalent model is that 'gene-collapse', or rapid loss of redundant duplicates immediately follows WGD and that most duplicates are non-functionalized early (Sankoff et al. 2010). This has been forwarded as a likely model in the case in teleost genomes as well, even for genes belonging to functional classes biased towards retention (Sato et al. 2009; Braasch et al. 2009; Schartl et al. 2013). Importantly however, none of these studies addressing the dynamics of gene loss following the TGD had data available from either of the two earliest diverging teleost subdivisions, the Osteoglossomorpha or the Elopomorpha (Figure 1.1), and with taxa belonging only to the largest teleost subdivisions the Ostarioclupeomorpha (zebrafish,

carp etc.) and the Euteleostei (pufferfish, medaka, cichlids etc.), they could not answer the most pressing question of whether the majority of gene losses occurred in the common ancestor of all teleosts rapidly following the TGD, or only much later in the common ancestor of Clupeocephala (Ostarioclupeomorpha + Euteleostei).

Evidence from more recent vertebrate polyploidization events suggest the gene loss following WGD could be slower than previously anticipated. Polyploidization is widespread across eukaryotes and extant polyploids (i.e. having a karyotype with twice as many chromosomes as a close relative) are still found in every major eukaryotic lineage (Otto & Whitton 2000). In vertebrates, extant polyploids are especially prevalent in fish (Braasch & Postlethwait 2012), and many commercially important species (e.g. sturgeon, carp, salmon) have recent polyploid ancestry (Leitch et al. 2004; Mable et al. 2011; Piferrer et al. 2009). All members of the family Salmonidae (salmon, trout etc.) are descendent from a polyploidization event (also known as 4R or the salmonid-specific genome duplication SSGD) event dated at ~88-103 million years ago (Macqueen & Johnston 2014).

**The Key Phylogenetic Position of Osteoglossomorpha**

Besides the TGD, at least 27 morphological synapomorphies distinguish the teleosts from the basal actinopterygians (Helfman 2009). The name “teleost” itself means “perfect bone” and indeed many of these synapomorphies are skeletal in origin. A mobile premaxillary bone in their skull, the uroneural arches of their vertebrae, and basibranchial tooth plates are three such inventions of the teleosts which confer unto them robust and adaptable morphology for feeding and locomotion.

As outlined in Figure 1.2, molecular phylogenetic studies have disagreed on whether osteoglossomorphs, elopomorphs, or a clade containing both, is the sister group to clupeocephans (Le et al. 1993; Inoue et al. 2001; Inoue et al. 2003; Hurley et al. 2007; Alfaro et al. 2009; Near et al. 2012). Systematists do not agree on whether osteoglossomorphs, elopomorphs, or a clade containing both are the sister lineage to clupeocephalans, but the respective monophyly of each major subdivision is consistently well supported (Broughton et al. 2013; Hurley et al. 2007; Inoue et al. 2001; Le et al. 1993; Broughton 2010; Inoue et al. 2004). Osteoglossomorphs are also the most species-poor subdivision of teleosts, with only ~220 extant members occurring within 6 families: Hiodontidae (2 species), Pantodontidae (1 species), Osteoglossidae (8 species), Notopteridae (8 species), Gymnarchidae (1 species) and Mormyridae (~201 species) (Nelson 2006). This outstanding question is relevant to ichthyologists interested in orienting morphological character state changes in the teleost tree but it also has significant impact on the orientation of molecular genomic evolutionary processes relative to the TGD event.

Progress towards understanding the TGD event has been made by studying the relatively few teleost species with good genomic resources available such as the laboratory models zebrafish (*Danio rerio*) and medaka (*Oryzias latipes*), the genomic models Fugu (*Takifugu rubripes*) and green spotted pufferfish (*Tetraodon nigroviridis*), and evolutionary genetic model stickleback (*Gasterosteus aculeatus*). However, all of these species belong to the single largest teleost subdivision: the Clupeocephala (Figure 1.1; node 16), which in turn is comprised of the superorders Ostarioclupeomorpha (Ostariophysii + Clupeomorpha) (Figure 1.1; node 17) and

Euteleostei (Figure 1.1; node 18). The remaining two smaller but earlier branching teleost subdivisions, Osteoglossomorpha (Figure 1.1; node 14) and Elopomorpha (Figure 1.1; node 15) are studied less, but are poised to extend the range of comparative genomic analyses to a time closer to the TGD. Of special interest to this thesis is the timing of diploidization following the TGD relative to elopomorph and osteoglossomorph divergence. In order to understand molecular evolutionary processes which occurred deep in evolutionary history in long extinct ancestral lineages it is helpful to select representative taxa which diverged deeply within the phylogenetic tree and can therefore be used in the comparative context to perform relative (to speciation) dating and orient evolutionary changes including gene gains, losses or morphological transformations.

The Osteoglossomorpha (bony-tongues) consists of approximately 220 species (less than 1% of all teleosts) in six families: Hiodontidae, Osteoglossidae, Notopteridae, Gymnarchidae, Mormyridae, and Pantodontidae (Hilton 2003; Sullivan 2004). Pantodontidae consists of only a single species, a small (~12cm) tropical fish endemic to the fresh waters of Western Africa and common to the European aquarium trade since at least 1905 (White 1994): the African freshwater butterflyfish, *Pantodon buchholzi* (Figure 1.3). The small size and widespread availability of *P. buchholzi*, in addition to its relatively small genome size (~753Mb) (Hinegardner & Rosen 1972), makes it an attractive model for genomic investigations and it represents the focus model organism in this thesis. The morphological uniqueness of *P. buchholziP. buchholzi* amongst osteoglossomorphs has earned it recognition as a monotypic species within its own family, the Pantodontidae, well before molecular phylogenetic

data was available (Hilton 2003). Interestingly, despite being recognized as a single monotypic species, there are two allopatric populations of *P. buchholzi* endemic to either the Niger and Congo basins of western Africa which have recently been found to have diverged >50Ma despite presenting no meristic morphological differences, nor any statistically significant morphometric differences (Lavoué et al. 2010). The degree of morphological stasis observed between these two deeply diverging populations is of an order comparable to the most extreme cases of morphological stasis ever observed in the vertebrate fossil record. I have also selected three further osteoglossomorph species for analysis: Peter's elephantnose fish *Gnathonemus petersii* (Mormyridae), the silver arowana *Osteoglossum bichirrhosum* (Osteoglossidae), and the clown knifefish *Chitala ornate* (Notopteridae). Together with new data generated in concert with a large genome consortium for the pre-TGD neopterygian fish the spotted gar *Lepisosteus oculeatus* these model taxa, shown in Figure 1.3, promise to yield new insights into the evolution of the teleost genome in the immediate wake of the TGD.

## Overall Aims and Objectives

The overall goal of this thesis is to contribute novel insights into the diverse range of evolutionary patterns and processes which can occur in a lineage following WGD events and to challenge some previously held assumptions. I hope to accomplish this by using the TGD as a case study and by contributing novel genome-scale data from the previously underexploited early diverging teleost subdivision Osteoglossomorpha to help fill important phylogenetic gaps in previous comparative analyses. Although significant inroads have been made in understanding the evolution of the genome of the most successful subdivision of teleosts following the TGD, the Clupeocephala, it is uncertain whether phenomena observed in this clade, including widespread lineage specific loss of duplicates, subfunctionalization and neofunctionalization of preserved duplicates leading to functional genomic changes, also occurred in the Osteoglossomorpha. By determining if these phenomena have also occurred in the early diverging Osteoglossomorpha clade and to what extent they are shared with Clupeocephala or evolved independently it will be possible to make an estimate of the relative timing of the occurrence of these phenomena following the TGD. In particular the phenomenon of diploidization following WGD is poorly understood and currently difficult to study in the context of the TGD due to the lack of useful comparative reference genomes which diverged before and/or during the diploidization process. The aims of this thesis therefore are to: 1) generate a substantial amount of new genomic data from representatives of the Osteoglossomorpha for comparative genomic analyses with Clupeocephala and pre-TGD actinopterygians, 2) employ phylogenetic methods to reconstruct the diploidization history of well-studied gene families with a clear

evolutionary history and determine to what extent Osteoglossomorpha has a shared vs an independent history of diploidization following the TGD, 3) determine the extent of shared vs independent gene losses experienced by Osteoglossomorpha and Clupeocephala in order to develop a more accurate estimate of the rate of gene loss following the TGD, and 4) perform comparative genomic analyses between extant osteoglossomorphs to examine to what degree evolutionary phenomena which occurred in the clupeocephalan lineage after their divergence from osteoglossomorphs (such as lineage specific reciprocal gene loss) also occurred within the osteoglossomorph lineage and therefore may be a predictable, repeatable outcome following WGDs and not chance lineage-specific feature of clupeocephalans. By focusing on Hox and other ANTP homeobox-containing transcription factor families which have been instrumental in advancing the present understanding of WGDs in general, examples of novel evolutionary patterns and processes associated with WGDs and the often-neglected process of diploidization will be revealed. These examples can then serve to guide future genome-wide analyses of the TGD as well as other WGDs and advance our general understanding of the range of effects WGD can have on molecular and phenotypic evolution more broadly.

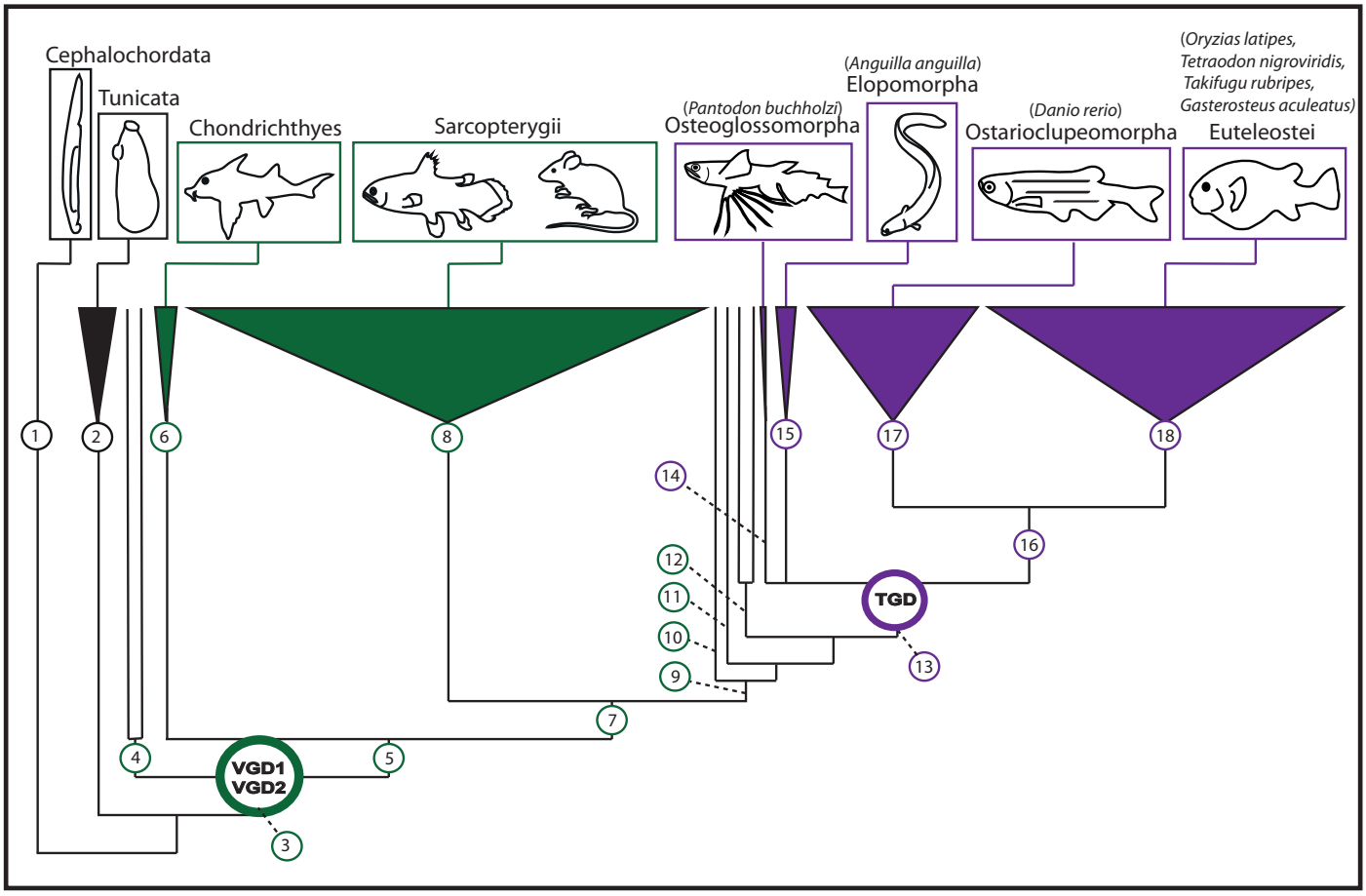



**Figure 1.1.** Whole Genome Duplications and the Key Phylogenetic Position of Osteoglossomorpha

Cladogram depicting chordate phylogeny and biodiversity with the position of whole genome duplication events shown: VGD1/VGD2 (Vertebrate Genome Duplication 1 and 2 also known as 1R/2R), TGD (Teleost Genome Duplication also known as 3R). The width of each clade corresponds to the number of species in the group. Osteoglossomorpha (Clade 14) branched from other teleosts near the root of the Teleostei (Clade 13) and is therefore a useful group for deep genomic comparisons of the TGD event with teleost species whose genomes have been well characterized within the Clupeocephala (Clade 16). Other major chordate clades are also illustrated for perspective. 1: Cephalochordata, 2: Tunicata, 3: Craniata, 4: Agnatha, 5: Gnathostomata; 6: Chondrichthyes, 7:Osteichthyes, 8: Sarcopterygii, 9: Actinopterygii, 10: Polypteriformes; 11: Acipenseriformes, 12: Holostei,13: Teleostei, 14: Osteoglossomorpha, 15: Elopomorpha, 16: Clupeocephala, 17: Ostarioclupeomorpha,18: Euteleostei. Representative species with available genomic data used in our analyses are listed.

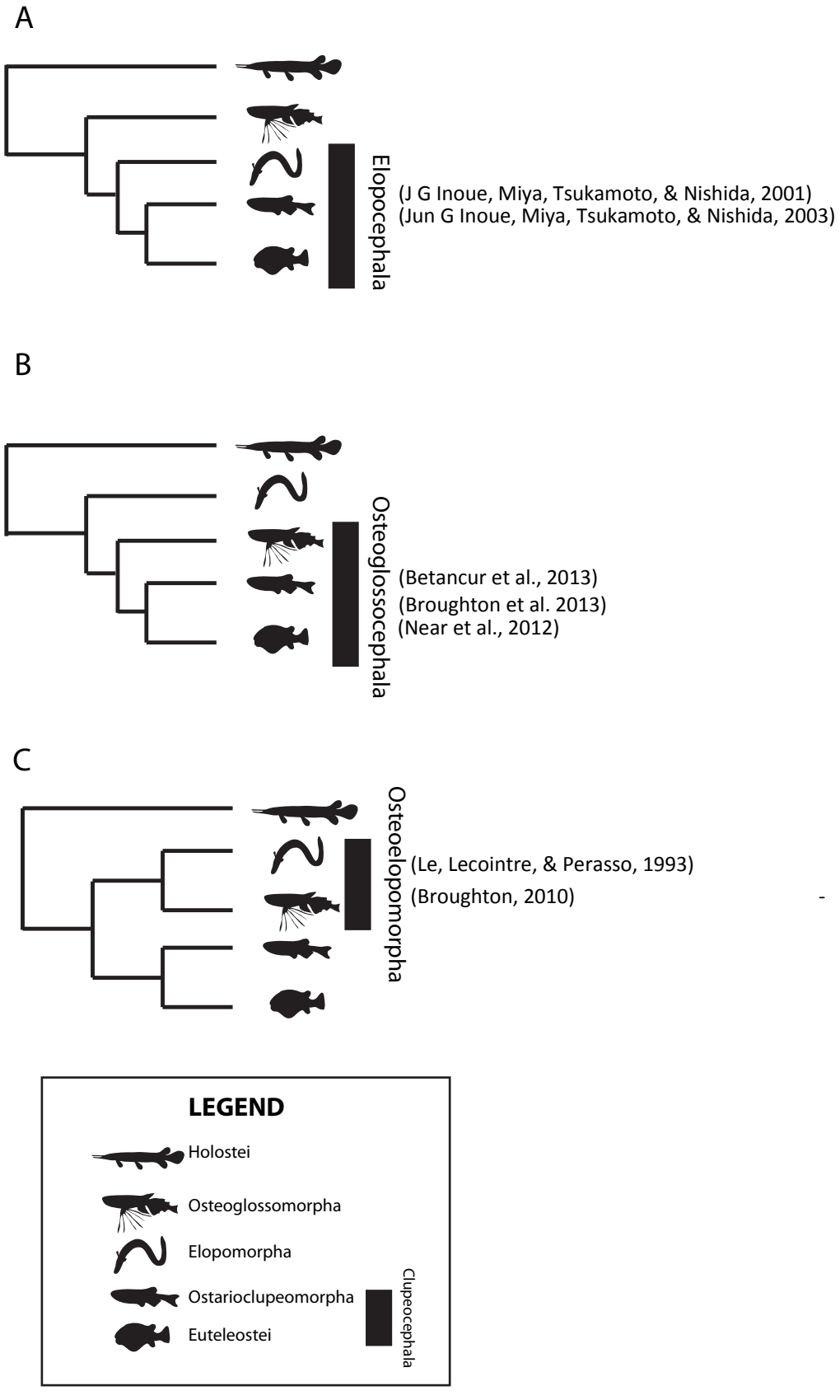



**Figure 1.2**. Proposed phylogenetic relationships between the three major teleost subdivisions. Three different phylogenetic relationships between Osteoglossomorpha, Elopomorpha, and Clupeocephala (Ostarioclupeomorpha + Euteleostei) have been proposed by molecular phylogenetic data. Key studies which support each hypothesis are listed.

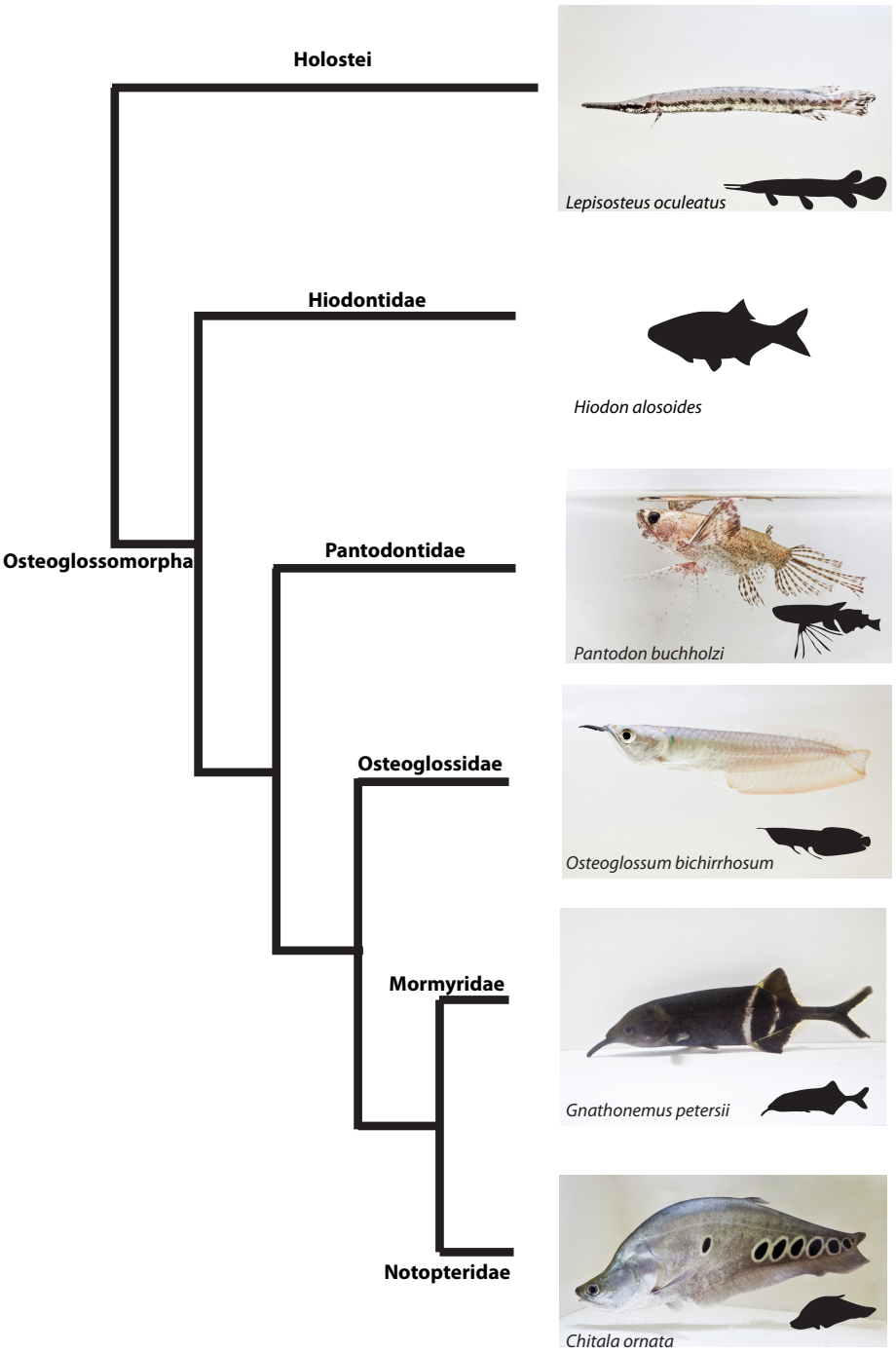



**Figure 1.3.** Model osteoglossomorph taxa under investigation.

Cladogram depicting the presently understood phylogenetic relationship between the species for which novel data was generated during this thesis. The spotted gar (*Lepisosteus oculeatus*) is a pre-TGD actinopterygian vertebrate. The goldeye (*Hiodon alosoides*) was not examined in this work but is the subject of ongoing investigation in other laboratories. The four osteoglossomorph fishes which are the subject of this thesis are the monotypic member of Pantodontidae, the African butterfly fish (*Pantodon buchholzi*), a representative member of Osteoglossidae, the silver arowana (*Osteoglossum bicirrhosum*), a representative Mormyridae, Peter's elephantnose fish (*Gnathonemus petersii*) and a member of the Notopteridae, the clown knifefish (*Chitala ornata*). All photographs were taken by myself of specimens maintained at the University of Oxford.

## Chapter 2

Published in *Molecular Biology and Evolution* (Martin and Holland 2014) – see Appendix 1

# Enigmatic orthology relationships between *Hox* clusters of the African butterfly fish and other teleosts following ancient whole genome duplication


## Abstract

Numerous ancient whole genome duplications (WGD) have occurred during eukaryote evolution. In vertebrates, duplicated developmental genes and their functional divergence have had important consequences for morphological evolution. While two vertebrate WGD events (VGD1/VGD2) occurred over 525 Myr ago, I have focussed on the more recent TGD (teleost genome duplication) event which occurred ~350 Myr ago in a common ancestor of over 26,000 species of teleost fishes. Through a combination of whole genome and BAC clone sequencing I characterized all *Hox* gene clusters of *Pantodon buchholzi*, a member of the early branching teleost subdivision Osteoglossomorpha. I found 45 *Hox* genes organized in only five clusters indicating that *P. buchholzi* has suffered more *Hox* cluster loss than other known species. Despite strong evidence for homology of the five *P. buchholzi* clusters to the four canonical pre-TGD vertebrate clusters (one HoxA, two HoxB, one HoxC and one HoxD), I was unable to confidently resolve 1:1 orthology relationships between four of the *P. buchholzi* clusters and the eight post-TGD clusters of other teleosts. Phylogenetic analysis revealed that many *P. buchholzi* genes segregate outside the conventional ‘a’ and ‘b’ post-TGD orthology groups, that extensive topological

incongruence exists between genes physically linked on a single cluster, and that signal divergence causes ambivalence in assigning 1:1 orthology in concatenated *Hox* cluster analyses. Out of several possible explanations for this phenomenon I favour a model which keeps with the prevailing view of a single TGD prior to teleost radiation, but which also considers the timing of diploidization after duplication, relative to speciation events. I suggest that while the duplicated hoxa clusters diploidized prior to divergence of osteoglossomorphs, the duplicated hoxb, hoxc, and hoxd clusters concluded diploidization independently in osteoglossomorphs and other teleosts. I coin the term 'tetralogy' to describe the homology relationship which exists between duplicated sequences which originate via a shared WGD, but which diploidize into distinct paralogs from a common allelic pool independently in two lineages following speciation.

## Introduction

*Hox* genes encode transcription factors with many highly conserved roles during development including specification of anteroposterior positional identity across bilaterians (Mallo et al. 2010). Hox genes belong to a large class of homeodomain-containing transcription factors known as ANTP (*Antennapedia*-like) named after their sequence similarity to the *Drosophila* homeobox gene *Antp*. The homeodomain is a short, 60 amino acid DNA binding motif consisting of a stereotypical helix-loop-helix-turn-helix structure, encoded by the homeobox sequence in DNA (Bürglin 2011).

The expansion of the *Hox* gene repertoire by WGD in vertebrate evolution is perhaps the most widely cited example of developmental gene family expansion through duplication and has been credited with permitting a wide range of evolutionary innovations, including the establishment of fundamental components of the vertebrate body plan (Málaga-Trillo & Meyer 2001; Wada et al. 1999; Pick & Heffer 2012; Pascual-Anaya et al. 2013). The vertebrate whole genome duplication events (VGD1/VGD2) were responsible for generation of the four ancestral gnathostome *Hox* gene clusters (HoxA, HoxB, HoxC and HoxD) from the single cluster shared with the last common ancestor of vertebrates and the cephalochordate *Branchiostoma* (Garcia-Fernàndez & Holland 1994; Putnam et al. 2008). Due to their highly conserved, colinear genomic organization, the vertebrate *Hox* gene clusters can serve as useful markers of the duplication status of entire genomes. It was partially on this basis that an additional independent WGD event was found in the teleost fish lineage (Amores et al. 1998). This event known as the TGD (Teleost

Genome Duplication) (Figure 1.1; node 13), occurred in the common ancestor of all ~26,000 species of extant teleost fish around 350 million years ago and has been implicated by some authors in explosive speciation and morphological radiation of this group (which comprise ~50% of all vertebrate biodiversity) (Meyer & Van de Peer 2005; Christoffels et al. 2004; Jaillon et al. 2004; Ravi & Venkatesh 2008; Hoegg et al. 2004).

Seven clusters of *Hox* genes have been found in clupeocephalan teleost genomes, resulting from the duplication of the four gnathostome clusters to eight during the TGD, followed by the loss of either a single hoxc or a single hoxd cluster in the percomorph (i.e. pufferfishes, medaka and stickleback) and ostariophysiian (i.e. zebrafish) lineages respectively (Woltering & Durston 2006). Genomic studies of elopomorphs in the genus *Anguilla*, for which some genome scale data has recently been generated (Coppe et al. 2010; Henkel et al. 2012a), have suggested that eels have retained more *hox* genes (70) and hox clusters (8) than any other teleost, which has not experienced a 4$^{th}$ WGD beyond the TGD, observed to date (Henkel et al. 2012b; Guo et al. 2010). This suggests a remarkable conservation of an ancestral post-TGD *Hox* complement, but the authors noted difficulties in assigning 1:1 orthology relationships with clupeocephalan post-TGD 'a' and 'b' orthology groups (Henkel et al. 2012b). A PCR survey of *Hox* gene fragments in an osteoglossomorph, the Goldeye (*Hiodon alosoides*), has suggested that up to 8 *Hox* clusters may be present in this species, but relied on isolated fragments as short as 82bp to form hypotheses of homology (Chambers et al. 2009). Other studies have also found evidence for duplication of individual genes in osteoglossomorphs which

have been interpreted as originating with the TGD (Hoegg et al. 2004; Crow et al. 2006; Chambers et al. 2009), however in many of these cases accurate homology relationships between the osteoglossomorphs genes and other teleosts was hindered by the low phylogenetic support for inclusion of the osteoglossomorph sequence within either the 'a' or 'b' post-TGD orthology groups, and the absence of any potentially corroborating physical linkage data.

When reconstructing gene genealogies for multiple species the resolution of two reciprocally monophyletic orthology groups, conventionally labelled by the extensions 'a' and 'b' in teleosts, each containing one of a pair of duplicates is considered good evidence for a gene duplication event occurring prior to a speciation event. The frequency with which this phylogenetic tree topology is observed between duplicates located on chromosomal loci sharing synteny (paralogons) forms the backbone of the evidence for the TGD. The nomenclature of the eight clupeocephalan teleost *Hox* clusters was formed on this basis (hoxa*a*/hoxa*b*, hoxb*a*/hoxb*b*, hoxc*a*/hoxc*b*, hoxd*a*/hoxd*b*). The ability to reconstruct these orthology groups with phylogenetic methods relies on the shared derived mutations (fixed substitutions) unique to members of the 'a' and 'b' clades and it therefore stands to reason that by the time the major clupeocephalan lineages diverged, the majority of their duplicates were already segregating as distinct paralogous loci.

Following any polyploidization event whether ancient or recent, the complex and still largely enigmatic process of diploidization inevitably follows, whereby a tetraploid karyotype returns to a diploid state over several million years concomitant with the loss of most duplicated genes (Wolfe 2001; Comai 2005).

The aim of this chapter is to investigate the dynamics of the diploidization process following the TGD and to investigate the evolutionary changes to genome structure in the immediate wake of the TGD. To this end, I sought to characterize the *Hox* gene complement of the African butterflyfish *P. buchholzi*, a member of the early branching teleost subdivision Osteoglossomorpha. Using BAC 454 pyrosequencing and Illumina whole genome sequencing methods, I generated the first extensive genomic dataset from an osteoglossomorph fish. Using this dataset I was able to recover all *Hox* gene sequences and their complete gene cluster organization using a three-tiered strategy involving degenerate PCR amplification, BAC clone pyrosequencing, and target-restricted assembly of Illumina whole-genome sequencing data to exhaustively screen the *P. buchholzi* genome. I hypothesized that if *P. buchholzi* shared the TGD with other teleosts it might have diverged before genomic changes ubiquitous to clupeocephalan teleosts arose and provides an earlier comparative genomic reference for inferring the condition of the ancestral teleost genome.

## Materials and Methods

### Animal husbandry

Multiple adult specimens of *P. buchholzi* ranging in size from approximately 7cm – 13cm and a mixture of males and females were obtained from the online tropical fish retailer Fareham Tropicals (London, UK) or the local fish retailer The Goldfish Bowl (Oxford, UK). For the purpose of PCR amplification of Hox and other gene fragments multiple genomic DNA samples from several individuals were used. For the purpose of Illumina sequencing and TRAM analysis (see below) genomic DNA from a single adult male *P. buchholzi*specimen of unknown age measuring 10.3cm was used. All animals were maintained in freshwater aquaria at 26-28$^{o}$C between a pH of 6.5-7.0 with a total dissolved solid (TDS) between 50-350ppm of the University of Oxford and fed on live crickets and/or mealworms daily. Animals were sacrificed using UK Home Office approved Schedule 1 methods by first immersing in an overdose of the anaesthetic MS222 (Tricaine) at 400mg/L followed by decapitation with a razor blade.

### Genomic DNA isolation

Genomic DNA was prepared from approximately 25-50mg samples of axial skeletal muscle and skin tissue (fillet cut) using a combined proteinase K, phenol-chloroform-isoamyl alcohol (PCIA) extraction protocol. Tissue was dissected and digested in 800µl of proteinase K digestion buffer (50mM Tris pH 8.0, 100mM EDTA pH 8.0, 1% SDS, 100mM NaCl, 350mg proteinase K in milli-Q water) overnight at 50$^{o}$C in 2ml lock-cap Eppendorf tubes on a benchtop incubator/shaker rotating at

400rpm. Following digestion, extraction of genomic DNA was carried out in 3 successive treatments of the aqueous, DNA-containing phase, with an equal volume of a 25:24:1 ratio of equilibrated phenol:chloroform:isoamyl-alcohol (Sigma) in an emulsion formed by vigorous shaking for 15s followed by spinning in a benchtop centrifuge at >13,000 rpm for 10 minutes. Following each treatment, the aqueous (DNA-containing) phase was transferred to a new tube. Excess phenol was removed using 2 successive treatments as above except with only chloroform to form the emulsion. Genomic DNA was then precipitated by adding 1/10 volume of 3M sodium acetate and 3 volumes of ice-cold 100% ethanol and left to precipitate for 1-2 hours at -20$^{o}$C, washed in 70% ethanol, and finally resuspended in 100-200µl of milli-Q water. Genomic DNA concentration and purity was estimated using a NanoDrop 2000 spectrophotometer and dilutions for PCR amplification were prepared at ~100ng/µl.

**Primer design**

Using multi-sequence alignments of the full collection of chordate Hox homeodomain amino acid sequences contained in the HomeoDB database (Zhong et al. 2008) (consisting of human, mouse, chicken, frog, zebrafish and amphioxus sequences) produced by the ClustalW algorithm (Larkin et al. 2007) implemented in the Geneious software package (Biomatters Ltd), highly degenerate primers (SO1 5' garytngaraargartt 3' and SO2 5' cknckrttytgraacca 3') were designed in order to amplify a 116bp fragment of the homeobox of *Hox* and related ANTP class homeobox-containing genes in *P. buchholzi*. Additional primers were designed to target specific vertebrate *Hox* families PG1, PG5, PG9 and PG13: (Hox1SO1 5' acvgarytngaraargartt 3' and Hox1SO2 5' catnckncкrttytgraacca 3'), (Hox5F1 5'

gcntayacbcgytaycagac 3', used with SO2), (Hox9F1 5' tgyccytayacnaaryayca 3', used with SO2), (Hox13F1 5' aarmgnrtnccntayasnaa 3' and Hox13SO2 5' acnckbckrttytgraacca 3'). Because the Evx/Eve gene family flanks the Hox gene clusters I also designed primers to target these genes (EvxF 5' yaycgngcnttcacvmgnga 3' and EvxR 5' ckctgnckyttgtccttcat 3') specifically to recover BACs which may have few or no remaining Hox genes at their original locus.

**PCR conditions**

Standard PCR was carried out with the Promega GoTaq® kit in 50µl total volumes comprised of 10µl 5X reaction buffer, 1µl 10mM PCR nucleotide mix, 1µl each of a 20ng/µl dilution of forward primer and reverse primer, 0.25-0.5µl GoTaq® DNA polymerase (5U/µl) and 0.25-1µl of *P. buchholzi* genomic DNA and topped up to 50µl with double-distilled autoclaved PCR water. To capture as wide a diversity of amplicons as possible at a range of annealing temperatures from 45℃ to 55℃ and magnesium concentrations 0.5 – 5.0 mM $MgCl_2$ was used. In cases where the $MgCl_2$ concentration desired was different from the 1.25mM provided by the stock 5X reaction buffer, 'homebrew' PCR reactions were prepared using 5µl 10X $NH_4^+$ buffer, 0.66-4µl 50mM $MgCl_2$, 0.4µl dNTP mix, 1µl of each forward and reverse primer diluted to 20ng/µl, 0.5µl GoTaq® polymerase and milli-Q water up to a total volume of 50µl. PCR thermocycler conditions were as follows: 95$^{o}$C initial denaturation 10 minutes followed by 30 cycles of 95$^{o}$C denaturation for 1 minute, annealing between 45$^{o}$C - 55$^{o}$C for 1 minute; elongation at 72$^{o}$C for 1 minute and terminating with an additional elongation step of 72$^{o}$C for 10 minutes.

## Cloning

PCR products were run on a 1% agarose TBE gel containing ~0.5µg/ml ethidium bromide (EtBr) at 100-120V alongside both the 1kb and 100bp ladder from New England Biolabs for between 30-60 minutes before bands between in the 70bp-120bp range were gel extracted under a UV transilluminator using the GE Illustra™ DNA and Gel Band Purification Kit according to manufacturers instructions. Cloning was carried out using a T-A ligation approach with the Promega pGEM®-T Easy vector system in reactions with a final volume of 10µl which included 1µl T4 DNA Ligase (3U/µl), 1µl pGEM®-T Easy vector, 5µl 2X Rapid Ligation Buffer and 3µl of the extracted DNA band, left overnight at 4°C. Transformation of chemically competent (calcium chloride treated) DH5α strain *E. coli* (ThermoFisher) was carried out using the heat shock method by first pre-incubating 50µl of a cell suspension with the full 10µl ligation reaction on ice for 40 minutes followed by heat shock at 42°C for 30 seconds and a return to ice for 2 minutes. A 500µl volume of fresh SOC medium was added to each reaction tube and cells were allowed to recover at 37°C for 30 minutes before spinning and plating on LB-agar plates containing 50ng/µl ampicillin, and previously coated with 50µl of 100ng/ul X-gal and 10µl of 100ng/µl IPTG. Following overnight incubation at 37°C all white (insert-containing) colonies were picked and used to initiate 5ml liquid cultures of each clone in LB containing 50ng/µl ampicillin shaking at 225 rpm overnight at 37°C.

## Sequencing

Plasmid DNA was isolated using the Qiagen QIAprep® spin mini prep kit and concentration and purity was assessed with a NanoDrop 2000 spectrophotometer before sequencing. Sanger sequencing reactions were carried out in both directions with the M13F and M13R primers using the ThermoFisher BigDye$^{TM}$ v3.1 Terminator Cycle Sequencing Kit by using 4µl plasmid DNA, 2µl 5X Seq Buffer, 1µl of a 3.2µM stock of either M13F or M13R primer, 1µl BigDye, and 1µl milli-Q water. Thermocycling was carried out with the following conditions: Initial denaturation at 96$^{o}$C for 90 seconds, followed by 30 cycles of denaturation at 96$^{o}$C for 15 seconds, annealing at 50$^{o}$C for 10 seconds, and elongation at 60$^{o}$C for 4 minutes. Products were precipitated using 3µl of 3M sodium acetate, and 67µl of 70% EtOH and an additional wash with 100µl of 70% EtOH. Cleaned up sequencing products were run on an ABI 3730xl DNA Analyser by the staff of the University of Oxford Department of Zoology Sequencing Center. Primer and vector sequences were trimmed and sequencing errors were manually corrected using the software Geneious (Biomatters Ltd).

## Clone identification

In total 213 clones were sequenced and identified by similarity to other vertebrate Hox or other ANTP genes by BLAST (Altschul et al. 1990) against the NCBI nr and nt database online (http://blast.ncbi.nlm.nih.gov/Blast.cgi). In this way 35 unique Hox-like and 3 unique Evx-like clones were identified between 82bp-114bp in length depending on the primer combination.

**BAC library screening**

A ~6X BAC library for *P. buchholzi* containing 27,648 clones with an average insert size of 138kb was prepared by Amplicon Express (BAC Library Code ABF) using the pCC1 Epicenter vector with the BamH1 cloning site from genomic DNA extracted from muscle tissue from five mixed sex (3 male, 2 female) individuals. DIG-labelled probes were synthesized by PCR from cloned *Hox* and *Evx* gene fragments. Each DIG labeling PCR reaction mixture was made to a final volume of 50µl using 5µl 10X $NH_4^+$ buffer, 2.5µl each of forward and reverse primers (20ng/ul), 1.5µl 50mM $MgCl_2$, 5µl DIG nucleotide mix (Roche), 0.5µl GoTaq® DNA polymerase, 1µl of plasmid DNA (100-200ng/µl) and 32µl milli-Q water. Thermocycler conditions used were the same as those originally used to amplify the gene fragments with either the SO1/SO2 or Hox paralogy group specific primer pairs. The GE Illustra$^{TM}$ DNA and Gel Band Purification Kit was used to clean up the reaction products and remove unincorporated DIG-labelled nucleotides and probes were eluted in a final volume of 30µl of milli-Q water. In this way, five pools of *Hox* probes and 3 separate *Evx* probes were synthesized. Each BAC library filter was subsequently pre-hybridized with DIG Easy Hyb Buffer (Roche) for 2-4 hours at 35-42$^{o}$C. The full 30µl eluent containing the DIG-labelled probes were then denatured at 98$^{o}$C for 10 minutes, and cooled rapidly on ice for 2 minutes. The full volume of denatured probes were added to 30ml of warm Easy Hyb Buffer (1:1000 dilution factor) which was used to replace the prehybridization solution and hybridization was carried out under relaxed stringency conditions for 16-18 hours between 35-42$^{o}$C to encourage cross-hybridization between the Hox probes and sequences besides the 35 which had

previously been cloned but may be present in the BAC library. Filters were subsequently washed with 2XSSC, 0.1%SDS, then 0.2XSSC, 0.1%SDS (25-42℃), and finally in wash buffer (0.1M Maleic acid, 0.3% Tween, 25-42℃). Blocking was carried out with 100ml Roche Blocking Solution for 30 minutes at room temperature. Roche anti-DIG fab fragment antibody (5µl) was then added to the blocking solution and incubated with the filters for 30 minutes at room temperature. Two 15 minute washes in wash buffer was followed by rinsing in detection buffer (0.1M Tris-HCl, 0.1M NaCl pH 9.5) for 5 minutes. Filters were then removed from the hybridization tubes, laid flat and 2ml of CDP-Star (Roche) was then added to the filters before wrapping in cling film and exposure to X-Ray film for between 5-15 minutes. Two separate library screens, one with the 35 combined *Hox* probes and one with the 3 *Evx* probes, were performed and the results yielded 82 candidate hox cluster-containing BACs. Cells corresponding to each of these BACs were picked and cultured overnight in 200ml volumes of LB-chloramphenicol (25µg/ml) shaking at 225rpm overnight at 37$^{o}$C and DNA was extracted using a modified “dirty-midi” protocol by following the QIAGEN® Plasmid Midi Kit protocol with a precipitation step using a 1/10 volume 3M sodium-acetate and 2 volumes of 70% EtOH as described above in place of the expensive midi column purification step. DNA from each of these 82 BACs was subsequently used to PCR amplify Hox and Evx gene fragments which were cloned and sequenced exactly as described above for their original amplification and identification from the genome. In this way, 10 BACs (57D22, 23G6, 39B7, 1I16, 27G8, 7O2, 3P22, 31M4, 22E23 and 42K2) were identified which

comprised all 34 *Hox* and 3 *Evx* gene fragments previously identified from the genome. No new *Hox* or *Evx* genes were amplified from any of these 82 BACs.

**Isolation and sequencing of BAC clones**

High-quality DNA from these 10 BACs comprising the full diversity of amplified probes (57D22, 23G6, 39B7, 1I16, 27G8, 7O2, 3P22, 31M4, 22E23 and 42K2) was prepared using the NucleoBond® BAC 100 DNA preparation kit protocol with the standard column purification step. DNA was quantified using the NanoDrop 2000 spectrophotometer, pooled in equimolar quantities and used to prepare one single-end library for Roche 454 pyrosequencing (Library 'E') at the University of Liverpool Centre for Genomic Research. Sequencing of this library on one quarter plate yielded a sequence coverage of ~30X. DNA from an additional 48 of the original 82 candidate BACs isolated by hybridization, but which did not yield known *Hox* or other homeobox gene fragments by PCR, was prepared using the NucleoBond® BAC 100 DNA preparation kit as 8 pools of separate cultures of 6 BACs each. These BAC pools consisted of: pool A (1E1, 2H15, 3I22, 5P8, 10E24, 10M4), pool B (14D16, 16K5, 17M5, 19D4, 19N18, 21J10), pool C (22I13, 22A23, 26A20, 28C18, 29I4, 30I11), pool D (33B5, 33D19, 37H21, 39J24, 41C14, 41B1), pool E (42H17, 44N19, 44O19, 49E9, 49M13, 56F18), pool F (51H11, 51L3, 51N5, 52E9, 52H23, 53N20), pool G (57L10, 58N11, 59D23, 61N19, 64F9, 47B22), and pool H (65E11, 65I16, 66K2, 67O21, 67F12, 72O20). Equimolar quantities of DNA from each of these pools, plus an aliquot of library 'E' containing the 10 confirmed Hox-containing BACs, was mixed and used to prepare a 3kb paired-end library (library 'G') for Roche 454

sequencing at the University of Liverpool Centre for Genomic Research on one quarter plate and yielded an approximately ~8X coverage of each BAC.

**BAC library sequence assembly**

Quality control was carried out on all 454 reads by using the software prinseq-lite (Schmieder & Edwards 2011) to remove sequences with phred scores < 20, and *E. coli* sequences which contaminated the BAC library prep from the host strains were removed using BLAST2GO (Gotz et all, 2008 REF) using a blastn cut-off similarity threshold of $E<10^{-6}$ . Sequenced libraries 'E' and 'G' were subsequently assembled together with the Roche Newbler assembler version 2.6 using default parameters. Assembly resulted in 251 scaffolds. 5 scaffolds comprising the full *Hox* clusters of *P. buchholzi* for hoxax (200,509 bp), hoxbx (379,906 bp), hoxby (370,487 bp), hoxcx (247,835 bp), and hoxdx (254,542 bp) and corresponding to the 10 original BAC samples from library 'E' (each cluster was spanned by two BACs from the original 10 BACs sequenced in library 'E') were identified by tblastn searches of a custom blast database constructed from the 251 scaffolds with each of the full-length chordate homeobox sequences from HomeoDB. Each locus which had a hit ($E<10^{-3}$) to a homeobox gene from HomeoDB was reverse-blasted using blastx to the NCBI nr database online (http://blast.ncbi.nlm.nih.gov/Blast.cgi) to verify its identity. None of the other scaffolds contained any *Hox* or *Evx* sequences, but some other homeodomain-containing gene families were found which could potentially explain the hybridization of the *Hox* and *Evx* DIG labelled probes to these BACs. Genscan (Burge & Karlin 1997) (http://genes.mit.edu/GENSCANinfo.html ) was used to predict the coding sequences of all genes on each of these five scaffolds and intron-exon

boundaries, start and stop codons were manually refined using the results of the blastx searches.

**Whole genome sequencing and the target restricted assembly method (TRAM)**

5ug of DNA from a single adult male specimen of *P. buchholzi* was used for Illumina HiSeq sequencing at the Wellcome Trust Center for Human Genetics, Oxford, using a 100bp paired-end protocol and 180bp insert size library. 171 million reads, corresponding to ~22.9X coverage of the ~753Mb genome, were generated and used to build both local BLAST and HMMER search databases. To seed the searches against these databases all vertebrate homeodomain protein sequences were downloaded from the homeodb web server (http://homeodb.zoo.ox.ac.uk/). Extremely relaxed stringency tblastn search (E < 10) returned 14,297 sequences, which were extracted from the 171 million read fasta database using the grep command in linux, and were assembled using cap3 to produce 378 consensus sequences with a mean size of 227bp and an N50 of 233bp, corresponding to the ~180bp homeobox encoding exon. HMMER searches were performed using the PFAM homeobox.hmm profile and the hmmsearch algorithm to search the 6 frame translation of the reads. 9,881 reads were extracted and assembled with cap3 (Huang 1999) into 338 consensus sequences. Cross referencing the hmmsearch and tblastn results revealed 456 unique consensus sequences which represent putative homeobox sequences from the *P. buchholzi* genome (See Chapters 3 and 4 for further analysis using this dataset). These were annotated using tblastn to the homeodb database and to a manually curated set of vertebrate *Hox* genes. This

revealed that no additional *Hox* or *Evx* sequences beyond those previously annotated in the five 454-generated scaffolds were present in the genome.

**Annotation of *Hox* cluster scaffolds**

Initial *ab initio* predictions of all coding genes in the BAC assembly scaffolds was performed with genscan, and verified by blastx searches against NCBI nr database. Only predicted proteins with high similarity ($E<10^{-3}$) to known proteins were retained for further annotation. Predicted *Hox* gene sequences were mapped to the scaffolds and predicted coding sequences of these scaffolds were refined by blastn searches to the NCBI nr database and clustalW multiple sequence alignment of predicted genes with the *Hox* gene coding sequences of *Danio rerio, Takifugu rubripes, Gasterosteus aculeatus, Tetraodon nigroviridis, Homo sapiens, Xenopus tropicalis,* and *Gallus gallus* to manually refine predicted start codons, stop codons and splice sites. The five *Hox* cluster scaffold sequences of *P. buchholzi*are deposited in GenBank under accession numbers KM102157–KM102161.

**Phylogenetic analyses**

Vertebrate taxa chosen for phylogenetic analysis were the clupeocephalan teleosts *Danio rerio, Salmo salar*, *Oryzias latipes, Gasterosteus aculeatus, Takifugu rubripes, Tetraodon nigroviridis*, and *Astatotilapia burtoni*, the elopomorph teleost *Anguilla anguilla* and sarcopterygian outgroups *Latimeria menadoensis, Xenopus tropicalis*, *Anolis caroliensis*, *Mus musculus* and *Homo sapiens* as well as the chondrichthyan *Callorhinchus milii.* Full length protein coding *Hox* gene sequences from five complete teleost genomes *D. rerio* (zv9)*, O. latipes* (MEDAKA1)*, G.*

*aculeatus* (BROADS1)*, T. rubripes* (FUGU4) and *T. nigroviridis* (TETRAODON8) as well as the sarcopterygian outgroups *X. tropicalis* (JGI_4.2), *A. caroliensis* (AnoCar2.0)*, M. musculus* (GRCm38)*,* and *H. sapiens* (GRCh37) were downloaded using the export function of the Ensembl genome browser (http://www.ensembl.org). *S. salar* sequences were obtained from NCBI accession numbers (EF695248-EF695353; EU02568-EU025719; EU221640-EU221655), *C. milli* from NCBI accession numbers (FJ824598-FJ824601), *L. menadoensis* from NCBI accession numbers (FJ497005-FJ497008), *A. burtoni* from NCBI accession numbers (EF594310-EF594316), and *A. anguilla* from NCBI accession numbers (JF891390-JF891400). Alignments for all individual *Hox* gene families were performed on amino acid sequences using ClustalW (Larkin et al. 2007) and trimmed with trimAl (Capella-Gutiérrez et al. 2009) using a gap threshold of 25%**.** These individual gene family alignments were both used for the AU likelihood ratio tests for individual families, and also used to build the whole *Hox* cluster concatenations after separate alignment. Maximum-Likelihood trees for both individual gene families and whole cluster concatenations were built using RAxML (Stamatakis 2006) using the LG amino acid substitution model with estimation of the gamma distribution parameter and the proportion of invariant sites, and 1000 bootstrap replicates. Bayesian trees were built using mrbayes (Huelsenbeck & Ronquist 2001) using the WAG amino acid substitution model with estimation of the gamma distribution parameter and the proportion of invariant sites across 4 chains for 500000 generations and sampling every 100 generations and checked for convergence by confirming that the Log likelihood values achieved by each of the chains achieved stationarity at high and

similar plateaus. A burnin value of 100000 generations was used and the consensus tree using a 50 percent majority rule was produced using the sumt function. Alternative topologies for likelihood ratio tests were constructed by building sets of 3 constrained trees containing 3 branches each – the clupeocephalan post-TGD 'a' constrained branch, the clupeocephalan post-TGD 'b' constrained branch and an unduplicated outgroup branch – and forcing the individual *P. buchholzi* genes, or the whole concatenated cluster, to segregate alternatively within the 'a' (topology A), 'b' (topology B), or outgroup clades (topology O). *A. anguilla* sequences were left unconstrained. The total tree and site-wise likelihood of each constrained topology was calculated using RAxML and the AU test (Shimodaira 2002) was performed with the program consel (Shimodaira & Hasegawa 2001). In each case the unconstrained best topology was also built with RAxML and compared with the 3 constrained topologies. 3D Site-Likelihood ratio plots (SLRPs) were constructed using R.

**CNE analysis**

The mVISTA application of VISTA-tools (http://genome.lbl.gov/vista/index.shtml) using the sLAGAN algorithm was used to form VISTA plots of *P. buchholzi* and other vertebrate *hox* gene clusters. CNEs using the Elephant shark *C. millii* sequence as the baseline were predicted using the cut-off threshold 70% identity over 100bp. CNEs were downloaded, and the total number, length, and average size were counted using thePERL scripts:

(1) Modify the default VISTA output file to create tab delineated columns

```
perl -ne 's/\s+/\t/g; print "$_\n";' abc.test | grep -v exon filename.txt | cut -f8 | tr -d [:alpha:] | awk '{sum += $1} END {print sum}' > filename2.txt
```

(2) Count the number of non-exonic CNEs

```
perl -ne 's/\s+/\t/g; print "$_\n";' abc.test | grep -v exon filename.txt | wc -l
```

**Selection analysis**

For selection analyses individual *Hox* gene trees were constrained by assuming *P. buchholzi* shared the TGD and cluster diploidization with other teleosts and forced all *P. buchholzi* sequences in a single cluster within either the post-TGD 'a' or 'b' orthology groups as follows (hoxax + hoxaa, hoxbx + hoxba, hoxby +hoxbb, hoxcx + hoxca, hoxdx + hoxda) and built these constrained trees using RAxML with the same parameters as the unconstrained trees. I estimated selection on the most proximate post-TGD branches ('a' + 'x', 'b' + 'y') using local branch models and branch-site models fit to the resultant ML trees using the software package HYPHY (Kosakovsky Pond et al. 2005). First, the MG94 codon substitution model (Muse & Gaut 1994) was fit locally to each individual branch to estimate dN, dS and ω using the AnalyzeCodonData.bf function in HYPHY (Pond et al. 2005). Next, I used the BranchSiteREL.bf function of HYPHY (Kosakovsky Pond et al. 2011) to fit the MG94 model allowing 3 independent estimates of ω ($\omega^-$, $\omega^0$, $\omega^+$) along each branch to representing negative, neutral, and positive selection respectively, and quantified the proportion of sites ($p^+$) which appeared to evolve under positive selection.

## Results

### Extensive *Hox* cluster loss in *P. buchholzi*

PCR amplification from *P. buchholzi* genomic DNA with degenerate primers targeted to homeobox sequences of *Hox* and the physically linked *Evx/Eve* gene families resulted in the amplification of 34 unique *Hox* gene and 3 *Evx/Eve* gene homeobox fragments. These fragments were pooled and used to screen a custom-made *P. buchholzi* ~6X coverage BAC library with an average clone insert size of 138kb. 10 *Hox*-positive BACs confirmed to contain all 37 PCR fragments were pooled to build one single-end library which was fully sequenced by 454 to an estimated average depth of 30X (Table 2.1). A further 48 BACs isolated in my initial screens were pooled with the previous 10 *Hox*-positive BACS to construct an additional 3kb paired-end library which was sequenced by 454 to an additional depth of approximately 8X (Table 2.1). Sequencing of both libraries yielded a sufficient amount of data for *de novo* BAC assembly which was performed by combining both data sets using Roche's newbler v2.6 software (Table 2.2). Careful manual annotation of the resulting scaffolds revealed that *P. buchholzi* possesses 45 intact full-length *Hox* gene coding sequences, interspersed with three mir-196 and four mir-10 family microRNAs in conserved positions (upstream *hox9* and *hox4* family genes respectively), present on five separate scaffolds confirmed to encompass the 10 original *Hox* BAC sequences by Sanger sequencing of the insert ends using the T7 and SP6 primer sites from the pCC1BAC backbone(Figure 2.1). The five *Hox* gene clusters could be readily identified as one HoxA, two HoxB, one HoxC and one HoxD cluster according to protein sequence similarity to the four ancestral gnathostome

*Hox* gene clusters. On this basis, the five *Hox* clusters of *P. buchholzi* were tentatively named hoxa*x*, hoxb*x*, hoxb*y*, hoxc*x* and hoxd*x* pending the phylogenetic evaluation of their homology with clupeocephalan teleost post-TGD 'a' and 'b' orthology groups.

Measured from the coding exons of the most 5' and 3' *hox* genes the approximate size of each *P. buchholzi* hox cluster is as follows: hoxax 55.2kb, hoxbx 66.1kb, hoxby 70kb, hoxcx 93.2kb, and hoxdx 45.4kb. The size of the *P. buchholzi* genome has been estimated at 753Mb (Hinegardner & Rosen 1972), intermediate between the genome sizes of zebrafish (1.5Gb) and *Tetraodon* (342Mb), the smallest vertebrate genome sequenced to date. The sizes of the hoxc and hoxd clusters of teleosts correlates well with the genome size and accordingly I found that the *P. buchholzi* hoxcx (93.2kb) and hoxdx (45.4kb) clusters are smaller than zebrafish hoxca (130.5kb) and hoxda (52.6kb) and larger than *Tetraodon* hoxca (69.3kb) and hoxda (36kb). The hoxcb cluster of zebrafish (72.2kb) and the hoxdb cluster of *Tetraodon* (24.5kb) are both reduced in size due to extensive gene loss not observed in *P. buchholzi* and are thus less relevant for comparison. The size of the hoxa and hoxb clusters of teleosts does not correlate as well with genome size and some *Tetraodon* clusters are larger than zebrafish. The *P. buchholzi* hoxax cluster (55.2kb) is smaller than the hoxaa clusters of both zebrafish (58.2kb) and *Tetraodon* (66.6kb), but larger than their hoxab clusters (33.3kb and 21.4kb respectively) which have experienced many gene losses. Both the *P. buchholzi* hoxbx (66.1kb) and hoxby (70kb) clusters are also smaller than the hoxba clusters of zebrafish (116.9kb) and *Tetraodon* (148.8kb) but larger than their hoxbb clusters (25.2kb and 15.2kb

respectively). Overall, the sizes of the hox clusters in *P. buchholzi* correlate with genome size in the case of hoxc and hoxd clusters, but not hoxa or hoxb clusters as observed in other teleosts.

Because five clusters is fewer than what has been found in any other teleost, I sought to guard against potential biases introduced by PCR amplification or BAC library construction and therefore sequenced the full genome of an individual male *P. buchholzi* specimen on one lane of Illumina using HiSeq2000 technology. This yielded 17.32Gb of 100bp paired-end (180bp insert) reads corresponding to an estimated depth of 22.9X based on a genome size of 753Mb (Hinegardner & Rosen 1972). I searched for *Hox* homeodomain sequences using the Target Restricted Assembly Method (TRAM) which allows targeted recovery and assembly of specific genomic loci from short-read sequencing data independent of whole genome assembly (Johnson et al. 2013). The TRAM method is carried out by using the raw reads to build databases formatted for BLAST (Altschul et al. 1990) and HMMER (Söding 2005) searches which were queried using the collection of chordate homeodomain protein sequences contained in HomeoDB (Zhong et al. 2008) or the PFAM homeodomain hidden markov model profile (PF00046) using tblastn and hmmsearch algorithms respectively, and reads were assembled using cap3 (Huang 1999). This method was specific enough to be able to independently assemble all 45 previously discovered *Hox* homeodomain-containing exon sequences at a read coverage level from 14X to 43X, and sensitive enough to recover homeodomain encoding exons corresponding to members of all 11 currently recognized homeodomain classes. Two of these TRAM assembled exons appear to be

fragments of *Hox* pseudogenes. A short fragment of a *hoxd2* gene, ψ*hoxd2x*, was subsequently found within a hoxdx BAC (7O2) sequence, while the homeodomain-containing exon of a second *hoxa13* gene, ψ*hoxa13y*, could not be recovered from my BAC assemblies. No sequences with similarity to an additional *hoxa13* exon 1 sequence could be identified by TRAM. I infer that this is a remnant of a deleted cluster, Δhoxay. No further *Hox* homeobox sequences could be detected. Searches for additional members of the mir-10 and mir-196 families by TRAM also failed to recover any sequences in addition to those present within the five previously characterized clusters. The results from the *P. buchholzi* genome searches therefore strongly corroborate the conclusion that I had characterized the complete *Hox* complement of *P. buchholzi*. Hence, according to the prevailing pan-teleost TGD model, *P. buchholzi* has lost three complete gene clusters (one each of HoxA, HoxC and HoxD) following the TGD. This is the most extensive case of *Hox* cluster loss following the TGD described to date, tripling the number of whole *Hox* cluster losses previously observed in any other teleost. It therefore seems that whatever constraint might exist to limit the number of *Hox* cluster losses in other teleosts does not apply to the *P. buchholzi* lineage.

Interestingly, despite the unprecedented number of whole *Hox* cluster losses, *P. buchholzi* has retained a total *Hox* gene complement (45 genes) similar to other teleosts (e.g. 49 in zebrafish; 46 in medaka). This is a consequence of the high proportion of duplicated genes retained in the hoxbx and hoxby clusters. Two members of the *hoxb13* family are retained in *P. buchholzi*, whereas at least one copy has been lost from the genomes of all other examined teleosts. Two members

of the *hoxb2*, *hoxb4*, and *hoxb9* families, which have been reduced to only one member in all clupeocephalans examined but are retained in duplicate in the elopomorph *Anguilla anguilla* (Henkel et al. 2012b), are also kept in *P. buchholzi*. These additional genes, notable for their absence in the genomes of zebrafish, stickleback, medaka and pufferfish were generally assumed to have been lost simultaneously with the TGD. I infer they were lost specifically in clupeocephans following their divergence from osteoglossomorphs and elopomorphs, indicating a significant time lag between the TGD and many duplicated gene losses. Individual *hox* gene losses have been documented even several hundred million years following the TGD (Hoegg et al. 2007). The conservation of both the duplicated *hoxb2*, *hoxb4*, *hoxb6* and *hoxb13* gene family in osteoglossomorphs suggests that the ancestral teleost hoxb clusters were far more complete than previously thought from examining only clupeocephalans.

In addition to the three whole-cluster losses, *P. buchholzi* has experienced several individual *Hox* gene losses since diverging from the last teleost common ancestor, including *hoxb1x*, *hoxb7x*, *hoxb10x*, *hoxb10y*, *hoxc1x*, *hoxd8x* and *hoxd1x*. I sought to evaluate to what extent these losses were shared with clupeocephalans suggesting rapid gene loss following the TGD, or independent suggesting a time delay between TGD and major changes to the teleost *Hox* gene complement. In order to ensure the same genes are compared in *P. buchholzi* and other teleosts it is necessary to establish 1:1 orthology between the *P. buchholzi* genes and the clupeocephalan counterparts belonging to either the post-TGD ‘a’ or ‘b’ orthology groups.

**Discordant *Hox* gene genealogies in *P. buchholzi***

I used phylogenetic methods to reconstruct unconstrained gene trees for each of the individual full-length *Hox* gene sequences of *P. buchholzi*to determine their orthology to clupeocephalan teleost *Hox* genes in the 'a' and 'b' orthology groups which formed following the TGD. The existence of more than four *Hox* gene clusters in *P. buchholzi* is consistent with a shared TGD model followed by the independent loss of three whole clusters in the *P. buchholzi* lineage. I therefore initially hypothesized that *Hox* genes in *P. buchholzi* would segregate with the canonical 'a' or 'b' post-TGD orthology groups recovered in clupeocephalan teleosts, whereas an unduplicated or independently duplicated sequence in *P. buchholzi* would segregate strictly outside of these clades. Because of their early divergence from other teleosts, osteoglossomorph sequences would by default be expected to segregate at the root of any teleost gene tree, potentially causing problems due to hidden paralogy where clupeocephalan teleosts retained only a single duplicate (Kuraku 2010). I therefore limited analyses to 26 orthology-informative *P. buchholzi* sequences (comprising 22 gene families) for which at least one clupeocephalan teleost with publicly available complete *Hox* gene complements retained both 'a' and 'b' paralogs. Two different phylogenetic reconstruction methods, Maximum-Likelihood (ML) and Bayesian Inference (BI), were used to evaluate the robustness of each gene genealogy and in order to minimize the potential impact of saturation due to the large amount of time since the TGD, only amino acid sequences were used.

Contrary to my initial hypothesis, I found that many individual *P. buchholzi* gene tree topologies were inconsistent with the expectations under a shared TGD

event (Figure 2.2; Figure 2.3). Rather than a single 1:1 orthology for each *P. buchholzi* gene with clupeocephalan genes belonging to either the TGD 'a' or 'b' orthology groups, several *P. buchholzi* genes segregated as an outgroup to, or in a polytomy with these clades (Figure 2.2; Figure 2.3). Furthermore, while one would expect genes which segregate within the 'a' orthology group to be physically clustered together on a separate chromosome from those which segregate with the 'b' orthology group, as found in other teleosts, I found that this pattern is disrupted in *P. buchholzi*. Even where strong phylogenetic support for 1:1 orthology of an individual *P. buchholzi* gene with members of either the clupeocephalan post-TGD 'a' or 'b' clades existed, adjacent genes yielded incongruent topologies such that within a single *P. buchholzi* cluster, genes grouping with the 'a' and 'b' orthology groups were interleaved, along with those which resolve best outside the combined TGD 'a' and 'b' clades entirely (Figure 2.2).

Examining the five orthology-informative genes on the *P. buchholzi* hoxax cluster revealed that four genes segregate with high support and good concordance between both ML and BI methods (bootstrap/posterior probability) with the post-TGD 'a' orthology group: *hoxa13x* (76/1.0), *hoxa10x* (72/0.98), *hoxa9x* (98/0.58), and *hoxa2x* (99/1.0) (Figure 2.3, A, A', B, B', C, C', E, E'). Only *hoxa11x* segregated as an outgroup (64/0.82) to both the post-TGD 'a' and 'b' clades with the caveat that zebrafish *hoxa11b* broke the expected monophyly of the clupeocephalan 'b' clade (Figure 2.3, D, D'). I used the likelihood-based approximately unbiased (AU) test for alternative tree topologies (Shimodaira 2002) to evaluate three constrained topologies in which the *P. buchholzi* sequences were grouped within either the

clupeocephalan ‘a’ (topology A) or ‘b’ (topology B) post-TGD orthology groups or positioned as an outgroup to the combined TGD ‘a’ and ‘b’ clade (topology O) (see Figure 2.4 A). In the case of genes within the duplicated clusters hoxbx and hoxby, the constraint imposed for topology O was exclusion from the ‘a’ or ‘b’ post-TGD orthology groups but the sequences were allowed freedom to adopt the either a nested or a paired configuration. All three constrained topologies enforced the respective monophyly of the post-TGD ‘a’ and ‘b’ orthology groups of clupeocephalans, as well as the monophyly of the sequences used as unduplicated outgroups. Using this test, I was able to confidently ($p < 0.05$) rule out alternatives to orthology with the post-TGD ‘a’ group (topology A) for three out of five genes on the hoxax cluster: *hoxa13x, hoxa9x* and *hoxa2x.* I was also able to rule out orthology of *hoxa11x* with the post-TGD ‘b’ group (topology B), but not the possibility of an outgroup relationship (topology O). Only for *hoxa10x* was this test not able to rule out any of the three alternative hypotheses. Taking both the unconstrained phylogenies and constrained topology tests into account I suggest that the hoxax cluster is a true 1:1 ortholog of the clupeocephalan hoxaa clusters. This is in line with the default expectations for simple 1:1 orthology between *P. buchholzi* sequences and other teleosts due to a shared TGD event. The results from the remaining *Hox* genes, however, did not meet with expectations under the prevailing model of the TGD.

Out of four orthology informative genes on the *P. buchholzi* hoxbx cluster two genes, *hoxb8x* (67/1.0) and *hoxb6x* (76/1.0) segregated with the post-TGD ‘a’ orthology group, while *hoxb5x* grouped either weakly within a subset of clupeocephalan sequences belonging to an atypically paraphyletic ‘a’ clade (ML) or

formed a polytomy with ‘a’ and ‘b’ clades (BI), and *hoxb3x* segregated best in an contradictory manner with the post-TGD ‘b’ orthology group with both tree reconstruction methods, albeit with low support (24/0.52). Most alternative constrained topologies evaluated by the AU test were not significantly different from each other and only topology O for *hoxb8x* could be confidently rejected ($p<0.05$).

Considering the hoxby cluster, with five informative genes, *hoxb6y* grouped reliably with the post-TGD ‘a’ orthology group according to both methods (76/1.0), while both *hoxb8y* and *hoxb3y* segregated either as outgroups to or formed polytomies with the inclusive TGD ‘a’ and ‘b’ clades, and *hoxb1y* segregated with the post-TGD ‘b’ orthology group (81/1.0). Curiously, *hoxb5y* grouped closely with *hoxb5x* amongst a clade of clupeocephalan sequences belonging to an atypically paraphyletic clade comprising the non-percomorph members of the post-TGD ‘a’ orthology group (ML), or formed a polytomy with both ‘a’ and ‘b’ clades (BI). I also observed that both *hoxb6x* and *hoxb6y* also segregate best within the same clade comprising members of the clupeocephalan ‘a’ orthology group. These irregularities are contrary to the default expectation under a shared TGD event which predicts that each *P. buchholzi* duplicate would belong to a single one of either the ‘a’ or ‘b’ orthology groups rather than both segregating together with only one. When I tested the statistical support for alternative constrained topologies I was able to eliminate topology O for *hoxb8x*, topology A for *hoxb8y*, and both topology A and O for *hoxb3y* which had appeared to segregate best as an outgroup to the inclusive TGD ‘a’ and ‘b’ clades in the unconstrained ML trees.

The hoxcx cluster has nine informative genes, five of which group with the post-TGD ‘a’ orthology group: *hoxc11x* (58/0.87)*, hoxc8x* (57/0.85), *hoxc6x* (65/0.84), *hoxc5x* (53/0.76), and *hoxc4x* (66/0.83). Three genes segregate as outgroups to the TGD ‘a’ and ‘b’ clades: *hoxc13x* (70/0.98), *hoxc10x* (78/0.90), and *hoxc9x* (49/0.99). Finally one gene segregated best with the post-TGD ‘b’ clade: *hoxc12x* (35/0.85). The AU test was unable to rule out more than one of the three competing topologies for any gene of the hoxcx cluster, eliminating only topology B in the case of *hoxc10x, hoxc9x* and *hoxc5x.*

Finally, within the hoxdx cluster, none of the three informative genes present segregated with high support in either the TGD ‘a’ or ‘b’ orthology groups. While *hoxd4x* grouped best with clupeocephalan ‘b’ genes in the ML analysis, bootstrap support was so exceedingly low (13) as to be considered unsupported. Of the other two informative genes, *hoxd11x* segregated as an outgroup to the TGD clade (27/0.76), while *hoxd9x* segregated either in an outgroup or in a polytomy inclusive of both TGD ‘a’ and ‘b’ clades. When the competing constrained topologies were evaluated via the AU test none of the three topologies for any genes in the hoxdx could be rejected at a confidence level of $p < 0.05$.

The topological incongruences I observed between the individual *P. buchholzi* gene trees could be strictly interpreted as a contradiction of the prevailing hypothesis of a TGD event inclusive of osteoglossomorphs. However, both the presence of a duplicated cluster (hoxbx and hoxby), and the strong support for the inclusion of multiple individual *P. buchholzi* genes within the clupeocephalan teleost post-TGD ‘a’

or ‘b’ orthology groups (particularly those of the hoxax cluster), supports the idea of a TGD event inclusive of *P. buchholzi*.

**Detecting selection in Hox genes following the TGD**

I sought an answer to the question of why the hoxbx, hoxby, hoxcx and hoxdx clusters behaved atypically compared with the hoxax cluster in thephylogenetic analyses. It has been proposed that rapid sequence divergence, potentially as a result of either positive selection or relaxed purifying selection occurred in the immediate wake of the TGD between some gene duplicates, driving them apart, but not others (Crow et al. 2006). If this occurred between the genes in the duplicated hoxa clusters, but not in the hoxb, hoxc or hoxd clusters this might cause stronger signal to accumulate between these duplicates and facilitate faithful reconstruction of their orthology. I tested for the presence of diversifying selection acting on the immediate post-TGD branches of each of the 22 orthology-informative *Hox* gene trees to compare the selective pressures on each of the duplicated *Hox* clusters under the assumption of an osteoglossomorph-inclusive TGD. Using partially constrained trees which enforced monophyly of the post-TGD ‘a’ and ‘b’ orthology groups of clupeocephalans but which left the *P. buchholzi* sequences unconstrained and free to adopt the most likely fit topology the MG94 codon substitution model (Muse & Gaut 1994) was used to estimate dN, dS and ω independently for each branch of the tree using a local model fit with the HyPhy software package (Pond et al. 2005). Overall, I found no evidence for different selective pressures acting immediately after the TGD between the hoxa gene trees and the hoxb, hoxc, or hoxd

gene trees, which could account for the difference in phylogenetic resolution (Table 2.3). Only two branches arising directly from the TGD node (hoxa9a/hoxa9x and hoxc4b) had ω > 1 (ω=0.183/0.128=1.43 and ω =0.014/0.002=6.68 respectively), consistent with positive selection. There were 15 additional proximate post-TGD branches in gene trees belonging to each of the hoxa, hoxb, hoxc, and hoxd clusters which returned infinite estimates for ω because dS was estimated at 0 under the topological constraints I imposed. In each of these cases dN was estimated at < 0.02 making any inference of positive selection based on these results weak. Even if I were to take these results as acceptable evidence for positive selection, the prevalence of selection after the TGD node is no higher in the duplicated hoxa than in the hoxb, hoxc or hoxd clusters.

Finally, because diversifying selection may be both episodic and affect only a small proportion of sites in an alignment, I tested for selection acting on only a small number of sites which might otherwise be lost in an overwhelmingly constrained sequence using the branch-site REL method implemented in HyPhy (Kosakovsky Pond et al. 2011). This method allows for three independent categories of ω estimates along each branch to accommodate site-wise variation in selective pressures, and does not require constraining the rest of the tree to a single dN and dS rate while testing for selection along one branch. Only one additional statistically significant instance of positive selection (Bonferroni corrected $p < 0.05$) could be found at the root of a post-TGD clade: a small proportion of sites ($q^+ = 4.44\%$) under positive selection could be found in the *hoxa2a/x* branch. I therefore believe that

differences in selective pressure alone are insufficient to explain the incongruent gene tree topologies in the *P. buchholzi* clusters.

**Heterogeneous phylogenetic signal within concatenated Hox clusters**

I sought to explore additional explanations for the failure of the *Hox* gene genealogies of *P. buchholzi* to match the predictions of a shared TGD event. Given the relatively short sequence length of individual *Hox* genes and their high levels of conservation, I asked whether the incongruent topologies I recovered were the result of stochastic variation due to low signal intensity. The concatenation method has been useful in resolving phylogenetic problems where weak signal due to limited data has obscured underlying evolutionary relationships (Rokas et al. 2003). I built concatenated superalignments comprising all full-length protein coding genes from each physically linked *Hox* cluster in *P. buchholzi* and a collection of other vertebrates. For each *Hox* cluster superalignment, I explicitly tested the collective orthology of the sequences within the *P. buchholzi* clusters to those of other teleosts by constructing trees for three competing constrained topologies and testing the statistical support for each using the approximately unbiased (AU) test (Shimodaira 2002). As for individual genes, I constrained the *P. buchholzi* clusters explicitly within either the post-TGD ‘a’ cluster (topology A) or ‘b’ cluster (topology B) as would be expected under the canonical shared TGD, or outside the inclusive TGD ‘a’ and ‘b’ clades (topology O). In the case of the duplicated hoxbx and hoxby clusters each was alternatively constrained with either the post-TGD ‘a’ or ‘b’ clusters for alternate versions of topology A and topology B, while for topology O both hoxbx and hoxby were constrained outside of the post-TGD node but given freedom to adopt their most

likely configuration. The results of this test indicate that there is strong evidence that the *P. buchholzi* hoxax cluster is a 1:1 ortholog of the post-TGD 'a' clusters of clupeocephalan teleosts, as expected for a shared TGD. Topologies B ($p= 2 \times 10^{-4}$) and O ($p = 2 \times 10^{-4}$), could be confidently rejected consistent with the results from individual gene genealogies. However, even when the signal of all informative *Hox* genes on a single cluster were concatenated, the results with the *P. buchholzi* hoxbx, hoxby, hoxcx and hoxdx clusters still broke the assumptions of 1:1 orthology with other teleosts, as in the individual gene genealogies. For the *P. buchholzi* hoxby cluster and the hoxdx cluster none of the three competing topologies could be rejected, while for the hoxbx and the hoxcx clusters only topology O could be rejected ($p = 0.02$, $p = 0.05$ respectively).

I employed a novel graphical method to visualize signal intensity and direction under three competing constrained tree topologies: 3D site-likelihood ratio plotting (3D-SLRP). I plotted the per-site likelihood ratios ($\delta_s$= 2Ln((Site-likelihood Topology 1) – (Site-likelihood Topology 2)) which were calculated during the AU test between each pair of competing constrained topologies (A vs. B, A vs. O, O vs. B) to obtain a 3D plot of the site-wise distribution of support for each of the three test topologies (Figure 2.4). Except for *P. buchholzi* hoxax, which shows a clear majority of sites uniquely supporting topology A (Figure 2.4, B), the plots for the four remaining clusters show signal trifurcations in site-wise support for topologies A (red dots), B (blue dots), and O (green dots) (Figure 2.4, C-F). While the intensity of the trifurcating signals in hoxbx, hoxby, and hoxcx are as strong as the unidirectional signal in hoxax, in the case of hoxdx the signal is slightly weaker and this may have

contributed to the difficulty in assigning orthology to this cluster. This heterogeneity in the signal for the full *Hox* cluster concatenations indicates that the failure of at least the *P. buchholzi* hoxbx, hoxby, hoxcx cluster concatenations to form clear 1:1 orthologous relationships with either the 'a' or 'b' post-TGD orthology groups in the AU tests is the result of conflicting signal between different sites in the *P. buchholzi* clusters, rather than an absence of signal. This outcome corroborates the results from the individual gene genealogies and supports the idea that the phylogenetic histories of individual sites/genes within the *Hox* clusters of *P. buchholzi* are heterogeneous.

**Ancient CNE retention patterns in *P. buchholzi* Hox clusters**

Identifying conserved non-coding elements (CNEs) is a useful method of identifying functional genomic elements, and many vertebrate CNEs found near developmental genes have been shown to play a role in gene regulation (Pennacchio et al. 2006; Woolfe et al. 2005). Genome-wide, ancient gnathostome CNEs (conserved between chondrichthyans and osteichthyans) appear to be evolving more quickly in teleosts than in other vertebrate lineages (Lee et al. 2011; Venkatesh et al. 2006). Earlier comparison of *Takifugu* and human Hox clusters with elephant shark *Callorhinchus milii* revealed that the duplicated *Takifugu* clusters retain fewer ancient CNEs than human and that all of the *Takifugu* post-TGD 'b' clusters are depleted of ancient CNEs compared with the 'a' clusters (Ravi et al. 2009). I performed VISTA alignments using the SLAGAN algorithm to identify ancient CNEs between elephant shark and the Hox clusters of *P. buchholzi* and five other teleosts (zebrafish, medaka, stickleback, *Takifugu* and *Tetraodon*) (Figure 2.5). My analysis reveals that while the

post-TGD ‘b’ clusters of all five clupeocephalan teleosts show a distinctive asymmetry in the retention of CNEs, the duplicated *P. buchholzi* hoxb clusters retain nearly the same number of CNEs (hoxbx:14, hoxby:15) (Figure 2.5, B). In clupeocephalans, the hoxbb cluster appears to have diverged more than the hoxba cluster having lost more CNEs (e.g. zebrafish hoxba: 26; hoxbb: 4, *Tetraodon* hoxba: 8, hoxbb: 3). The gene content of the clupeocephalan hoxbb clusters is also significantly lower than hoxba (e.g. zebrafish hoxba: 11; hoxbb: 4, *Tetraodon* hoxba: 9; hoxbb: 4). The symmetric pattern of CNE conservation in both *P. buchholzi* hoxb clusters is consistent with the higher degree of *Hox* gene retention (hoxbx: 8; hoxby: 10) and suggests a functional role for these CNEs, potentially in the regulation of *Hox* gene duplicates conserved in *P. buchholzi* but lost in other teleosts. This also suggests that the rapid structural evolution of the clupeocephalan hoxbb cluster only got underway after the speciation event separating *P. buchholzi* and other teleosts.

Ancient CNE retention patterns in the *P. buchholzi* hoxax, hoxcx and hoxdx clusters are more similar to clupeocephalan hoxaa, hoxca and hoxda clusters respectively than to hoxab, hoxcb or hoxdb which like hoxbb have experienced extensive *Hox* gene and CNE loss (Figure 2.5, A, C, D). However, without corresponding duplicated clusters in *P. buchholzi* to compare with it is impossible to conclude whether the rapid structural evolution of the clupeocephalan hoxab, hoxcb and hoxdb clusters occurred before or after separation of the lineage leading to *P. buchholzi*. If asymmetric structural evolution only occurred in the clupeocephalan lineage after this speciation, as seems likely in the case of hoxbb, then presence or absence of CNEs is not phylogenetically informative.

## Discussion

I suggest a biological explanation for the unusual pattern of molecular evolution I observe that fits all of these data and is in line with the current understanding of genome diploidization dynamics in polyploids, but which has been largely overlooked in the study of ancient WGDs. I propose that the four alleles of the duplicated hoxb, hoxc and hoxd clusters did not complete diploidization into two pairs of alleles on separate paralogous gene clusters until after the very early divergence of osteoglossomorphs and clupeocephalans. In contrast, after the TGD the hoxa clusters resolved faster, prior to divergence of these lineages. This model implies that the last teleost common ancestor was a pseudotetraploid, like present-day species which more recently underwent WGD such as salmonids, and had a composite genome consisting of both diploidized and tetraploid segments. Prolonged segregation as four alleles in a pseudotetraploid lineage would potentially subject the duplicated hoxb, hoxc, and hoxd clusters to concerted evolution between duplicated loci, violating the assumptions of immediate genetic independence and simple 1:1 orthology of all teleost duplicates following the TGD.

One alternative model is that osteoglossomorphs experienced an independent WGD around the same time as the TGD, after diverging from other teleosts. Recently, an independent WGD event has been described in the American paddlefish (*Polyodon spathula*), a non-teleost actinopterygian fish, underlining the potential for undiscovered independent WGDs in other vertebrate lineages (Crow et al. 2012). However, this model does not explain the strong phylogenetic affinity for *P. buchholzi* hoxax with clupeocephalan hoxaa and, depending on the relative position of

osteoglossomorphs and elopomorphs in the teleost tree, it might require yet another independent WGD in the elopomorph lineage. Likewise, independent duplication of the hoxb cluster in *P. buchholzi* is unlikely to have occurred as this would implicate additional independent duplications of the other clusters in the *H. alosoides* lineage, which retains eight Hox clusters (Chambers et al. 2009), and does not fit my phylogenetic results. Another model might be that while osteoglossomorphs shared the TGD, the unprecedented loss of three whole Hox clusters in *P. buchholzi* constrained the evolution of the remaining clusters causing them to retain their similarity to their pre-TGD counterparts. While plausible, under this model one would expect that singular clusters in *P. buchholzi* (hoxax, hoxcx, hoxdx) would be more constrained and more difficult to homologize than clusters retained in duplicate (hoxbx/hoxby). However, what I observe is that the singular *P. buchholzi* hoxax cluster segregates well with the hoxaa clusters of other teleosts (which have never lost either of their hoxaa or hoxab clusters). Conversely the pair of *P. buchholzi* hoxbx/hoxby clusters do not form clear 1:1 orthologies with other teleosts. Additionally, in clupeocephalan *Hox* genes I do not note any consistent difference in the ability of orphaned duplicates compared with replete pairs of genes in for their ability to segregate in a single post-TGD orthology group. Finally, while my analyses suggest that selection on protein-coding genes was not significantly different between Hox clusters in the immediate wake of the TGD, and that signal trifurcation was responsible for ambiguity in assigning 1:1 orthologies to the most of the clusters, it remains possible that forces operating at some as yet unidentified level are responsible for the phenomenon I observe here. The possible effect of long-branch

attraction which can result in the erroneous grouping of quickly-evolving branches in a phylogenetic tree by misinterpreting homoplastic characters as homologous changes should not be discounted. However, the generally high sequence conservation of *P. buchholzi* genes and the absence of topologies which support the grouping the *P. buchholzi* sequences with quickly evolving branches such as the percomorph *hoxb3b*, *hoxb5a, hoxc5a, hoxd4b*, *hoxd9b* and *hoxd11b* (Figure 2.3) suggests that long-branch attraction is not conspicuous in this case. The most plausible explanation which can be given at this time remains the segmental and independent diploidization of the *P. buchholzi* genome following a shared TGD.

**Considering diploidization following the TGD**

The finding that *P. buchholzi* has five *Hox* clusters including two HoxB loci is in agreement with the prevailing view that the TGD predates the divergence of the Osteoglossomorpha and hence is shared by all teleosts (Hoegg et al. 2004; Crow et al. 2006; Hurley et al. 2007). Based on observations in other teleosts, it is thought that the last teleost common ancestor possessed eight *Hox* clusters immediately after the TGD. This implies that three *Hox* clusters were lost in the *P. buchholzi* lineage (one HoxA, one HoxC and one HoxD cluster). Teleost *Hox* cluster evolution appears to be significantly more dynamic than other gnathostome vertebrates which almost invariably retain the four ancestral *Hox* clusters with relatively few independent gene losses in different lineages (Liang et al. 2011). The only known exception to this stability are the elasmobranchs *Scyliorhinus canicula* and *Leucoraja erinacea* which appear to have lost their HoxC cluster after diverging from chimaerids > 400 Myr ago

(Oulion et al. 2010; King et al. 2011; Ravi et al. 2009). However, even amongst teleosts which have seen the ostariophysiian, and salmonid lineages independently lose their hoxdb clusters (Woltering & Durston 2006; Mungpakdee et al. 2008), and percomorphs including medaka, pufferfish, and cichlids lose their hoxcb clusters (Hoegg et al. 2007; Amores et al. 2004; Tümpel et al. 2006), the loss of the hoxay, hoxcy and hoxdy clusters in *P. buchholzi* stands out as the most extensive case of Hox cluster loss currently known.

The presence of the duplicated HoxB cluster alone, however, is not strictly sufficient to conclude that *P. buchholzi* shared the TGD and lost three clusters. My phylogenetic analyses revealed that most individual genes within the *P. buchholzi* hoxax cluster, and the whole cluster considered together in a concatenation of its component genes, groups with very good support with the hoxaa clusters of clupeocephalan teleosts exactly as one would expect if *P. buchholzi* shared the TGD with clupeocephalans and subsequently lost the hoxay cluster. This also implies that by the time *P. buchholzi* diverged from other teleosts, the duplicated HoxA clusters were already segregating as fully diploidized paralogous loci. However, the interleaving phylogenetic affinities of individual *Hox* genes within the *P. buchholzi* hoxbx, hoxby, hoxcx and hoxdx clusters for incongruent topologies, considered together with the trifurcation in the support of individual sites within these clusters for the post-TGD ‘a’ or ‘b’ orthology groups, or outgroup topologies, contradicts the predictions for a shared TGD event if one assumes that all Hox clusters concluded diploidization simultaneously prior to the radiation of major teleost subdivisions.

These apparently paradoxical findings can be reconciled by postulating a protracted pseudo-tetraploid period in teleost evolution, following the ancient TGD polyploidization event and preceding the establishment of the fully diploid (palaeopolyploid) karyotype of modern teleosts. If *P. buchholzi* diverged from other teleosts during this pseudo-tetraploid period, prior to the conclusion of the diploidization process, it is likely that it shared the diploidization of some genomic loci prior to speciation, but not others. The effects of homeologous recombination and gene conversion between the highly similar tetraploid alleles of duplicated but undiploidized loci, as well as the potential for independent sorting of alleles during the speciation process, could have led to the divergent phylogenetic pattern which was observed if osteoglossomorphs diverged early enough that diploidization was concluded independently from other teleosts. If these genetic homogenizing forces were strong enough they would have the effect of slowing down or even stopping the molecular clock, preventing differences between duplicates from becoming fixed and effectively delaying their divergence until after diploidization had concluded. If this continued until after the deep divergence between osteoglossomorphs and clupeocephalans, duplicates in *P. buchholzi* would have separate coalescents from other teleosts marking the independent cessation of recombination, and would fall as outgroups to the clupeocephalan teleost ‘a’ and ‘b’ clades, as I observed in many cases (e.g. *hoxb8y*, *hoxc13x*). Likewise, a weaker degree of genetic exchange between homeologous loci might allow some mutations between duplicates to become fixed, but subsequently cause them to swap their physical position between homeologous chromosomes, resulting in the interleaving pattern I observe (e.g.

*hoxb8x*, *hoxb3x*). Non-reciprocal homeologous recombination (gene conversion) may also occur resulting in the segregation of two duplicates within a single orthology group (e.g. *hoxb6x*, *hoxb6y*). It may be impossible to mark a definitive point in time when genetic exchange between duplicated homeologous loci ceases completely. It is likely that homeologous loci diverge progressively, undergoing genetic exchange less and less frequently with time until finally it becomes so infrequent as to be effectively non-existent. Using lineages which speciated close to the TGD event, I was able to use phylogenetic data to date the relative order that duplicated genomic segments achieved complete independence. With this method one can infer that the duplicated hoxa clusters diploidized first, prior to the speciation event separating osteoglossomorphs and clupeocephalans, while the hoxb, hoxc and hoxd clusters most likely concluded diploidization independently after divergence of these two lineages. A new model of teleost *Hox* cluster evolution which gives an account of the relative timing of diploidization of the different clusters in addition to detailing individual gene and whole cluster loss patterns is outlined in (Figure 2.6).

**Hox diploidization dynamics following the TGD**

The results presented here are consistent with the ancient TGD being either an autopolyploidy event where the duplicated HoxA clusters diploidized before the HoxB, HoxC, and HoxD clusters, or a segmental allopolyploidy event in which the HoxA clusters of the parental species were effectively pre-diploidized immediately upon hybrid formation, but strongly exclude genomic allopolyploidy as an option.

If the divergence of osteoglossomorphs and clupeocephalans occurred rapidly following the TGD, it would therefore be conceivable that the diploidization

process was only partially complete, and concluded independently following their divergence. The earliest recognizable fossils of the major teleost subdivisions (Osteoglossmorpha, Elopomorpha, Clupeocephala) only appear near the Jurassic-Cretaceous boundary ~150 Mya (Arratia 1997). Molecular estimates for the divergence time of crown teleosts vary widely according to whether mitochondrial (268–326 Mya) or nuclear (181–265 Mya) data are used (Hurley et al. 2007). Estimates for the timing of the TGD based on dating the divergence of paralogous genes in clupeocephalan teleosts range between 320-350 Mya (Taylor et al. 2001; Christoffels et al. 2004; Vandepoele et al. 2004). Using the median divergence time estimates for the teleost radiation and the TGD respectively a very rough approximation of 38 million years (mitochondrial) to 112 million years (nuclear) between the TGD and the divergence of the major teleost subdivisions was estimated. Because of the strong dependence of the molecular clock used to make these estimates on the diploidization process itself, the relative dating of the TGD and teleost diversification may become somewhat confounded. However, while keeping this caveat in mind, it is still striking that the estimated time lag between the TGD and the speciation of the major teleost subdivisions is comparable to the time lag estimated recently estimated between the salmonid 4R WGD and the radiation of the major salmonid subfamilies (40-50Ma) (Macqueen & Johnston 2014). This could suggest that an initial lag phase, rather than an acceleration, in speciation following WGD is not exceptional and may even be commonplace. In the case of salmonids, Macqueen & Johnston concluded that climate-linked ecophysiological factors (namely glaciation cycles and the evolution of anadromy) were major players in the

diversification of the group rather than the 4R WGD *per se.* Whether or not the 4R WGD was necessary for the adaptation and diversification of salmonids in response to climactic changes (i.e. would the group have diversified in the same way had 4R WGD not occurred) is nigh impossible to test but it seems increasingly unlikely that WGD was sufficient as a driver. In the case of the TGD it is interesting to speculate whether or not the End-Permian extinction event ~252 My ago (Shen et al. 2011) played a role in the subsequent diversification of teleosts which took place over a similar geological time-scale, perhaps by opening new ecological niches into which surviving members of the recently polyploidized teleost lineage could expand.

**Tetralogy: a new homology subtype**

I suggest that the classical concepts of orthology and paralogy are not adequate to describe the homology relationship between gene duplicates in two species which originated from a single duplication event, yet do not share a coalescent as a result of independent diploidization following speciation. I therefore recommend the use of novel terminology for this subtype of molecular homology which will serve to quickly and accurately describe it: tetralogy.

Under the classical model, relationships between genes in two species which experienced a shared WGD (and diploidization) event prior to speciation resolve unambiguously into either 1:1 orthologs which descend from the same gene present in their last common ancestor, or paralogs which are descendants of two different duplicates in their last common ancestor (Figure 2.7 A). If two separate WGD events occurred independently following speciation, 1:1 orthologs in the descendant species could not exist because the duplicates originated from two events and two distinct

ancestral genes, and only paralogous relationships would exist (Figure 2.7 B). Correspondingly, according to the model proposed here, where two species share a WGD event but where duplicates diploidize into genetically independent loci from a shared allelic pool independently in the two lineages (Figure 2.7 C), none of the resultant genes can be considered 1:1 orthologs because they do not share a single common ancestral sequence. However they share an important property with the orthologous duplicates generated under the model depicted in Figure 2.7 A in that they fundamentally originate from the exact same duplication event. This distinguishes them from the paralogous duplicates generated in the model depicted in Figure 2.7 B. I use the term tetralogy to describe this unique type of 2:2 orthology relationship. Because the topology of individual gene trees resulting from independent diploidization are indistinguishable from those produced by independent duplication, only by examining multiple loci and syntenic genes can independent duplication be conclusively ruled out and tetralogy as a result of independent diploidization be inferred. When I compared phylogenetic trees of multiple genes across the *Hox* clusters of *P. buchholzi* and other teleosts, what I found was a combination of topologies. Some genes segregate in ways consistent with both the independent diploidization and independent duplication scenarios, while others segregate exactly as one would expect for a shared WGD and diploidization scenario. I can therefore infer that *P. buchholzi* did not experience a separate WGD event but rather that independent diploidization is likely to have occurred.

Because consistent *Hox* gene nomenclature which reflects both homology and physical position is needed, I advocate retaining the suffixes '*x'* and '*y'* for all *Hox*

genes in *P. buchholzi*, including those individual genes which segregate within clupeocephalan ‘a’ or ‘b’ orthology groups but are part of a cluster which does not otherwise share unanimous 1:1 orthology with other teleost clusters. I also retain the naming system for members of the hoxax cluster mainly for consistency.

## Conclusion

I have demonstrated that the *Hox* gene clusters of *P. buchholzi* represent a unique departure from the situation seen in other teleost fishes. *P. buchholzi* has fewer *Hox* gene clusters than any other teleost (five), but a similar total *Hox* gene number (45). I also observe that phylogenetic analysis of the Hox clusters of *P. buchholzi* fails to support a shared TGD hypothesis unless a new model accounting for diploidization after duplication is considered. Under this model I propose that after a shared TGD, there was gradual and segmental diploidization of the genome which concluded independently in osteoglossomorphs and clupeocephalans, and that the last teleost common ancestor was therefore likely a pseudotetraploid. I suggest that *P. buchholzi* shared TGD with other teleost fish, but that most of its *Hox* gene clusters (hoxbx/hoxby, hoxcx, hoxdx) loci violate assumptions of 1:1 orthology with clupeocephalan teleost *Hox* clusters as a result of independent diploidization. Therefore, instead of all duplicates which originated at the TGD showing clear 1:1 orthology across teleosts, some may share a novel type of 2:2 orthology relationship as a result of independent diploidization: tetralogy. The prevalence of this phylogenetic pattern in other gene families across the genome of *P. buchholzi* and related osteoglossomorphs will be of significant interest and can be tested once additional genomic resources become available.

**Table 2.1.** 454 sequencing of candidate Hox cluster-containing BACs.

Candidate *Hox* cluster-containing BACs identified by screening the *P. buchholzi* BAC library were pooled in two ways to sequence on two separate ¼ plate runs. Library 'E' comprised 10 BACs verified to contain *Hox* genes by PCR, cloning and Sanger sequencing and was sequenced in the first ¼ plate run as a single-end library. Library 'G' comprised 48 candidate *Hox* cluster-containing BACs which were identified by hybridization but *Hox* negative by PCR, pooled in equimolar quantities with the 10 *Hox* positive BACs from library 'E' and sequenced on a separate ¼ plate run as a 3kb paired-end library.

| Library | n BACs | n sequences | total length (bp) | min length (bp) | max length (bp) | mean length (bp) | %GC | Estimated BAC Coverage (138kb average) | % Aligned Reads |
|---|---|---|---|---|---|---|---|---|---|
| E (Single-End) | 10 | 111,427 | 41,626,828 | 30 | 1,007 | 373.58 | 44.30 | 30.16 X | 98.22 |
| G (3kb Paired-End) | 58 | 261,570 | 66,092,114 | 40 | 1,038 | 252.67 | 41.54 | 8.26 X | 91.45 |

**Table 2.2.** Assembly of 454 sequencing results from candidate Hox cluster-containing BACs.

Library 'E' and library 'F' (see Table 2.1) were assembled together using Roche's Newbler v 2.6 software using default parameters. Summary statistics for total number of large contigs (>500bp), scaffolds, and contigs contained within these scaffolds is shown.

| Newbler Assembly Output | n | N50 (bp) | mean (bp) | total length (bp) |
|---|---|---|---|---|
| Large Contigs (>500bp) | 5,250 | 2,477 | 1,753 | 9,204,834 |
| Scaffolded Contigs | 2,794 | 3,282 | 2,616 | 7,310,703 |
| Scaffolds | 251 | 81,936 | 36,435 | 9,145,368 |

**Table 2.3.** Test for positive selection on the stem of the teleost 'a' and 'b' paralogy groups.

For each gene I constrained the *P. buchholzi* sequences of uncertain orthology 'x' and 'y' within either the teleost paralogy groups 'a' or 'b' in order to test if there was any evidence for diversifying selection acting on particular genes

| | | | Local MG94 (free branch ratio, single site ratio) | | | Branch-site REL MG94 (free branch ratio, 3 site ratio categories) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Gene | Node | ω | dN | dS | average ω | ω+ | q+ | LRT | p (corr) |
| HOXA | *hoxa2* | Tel a/x | 0.21 | 0.059 | 0.277 | 0.223 | 1188.050 | 0.044 | 20.0216 | **1.57E-04** |
| | | Tel b | 0.06 | 0.004 | 0.069 | 0.049 | 1.216 | 0.000 | 0 | 1 |
| | *hoxa9* | Tel a/x | **1.43** | 0.184 | 0.128 | 0.555 | 58.011 | 0.141 | 6.53446 | 2.17E-01 |
| | | Tel b | 0.20 | 0.132 | 0.664 | 0.331 | 3.315 | 0.175 | 1.84881 | 1 |
| | *hoxa10* | Tel a/x | 0.10 | 0.038 | 0.371 | 0.148 | 8444.520 | 0.014 | 4.00484 | 8.62E-01 |
| | | Tel b | 0.19 | 0.009 | 0.047 | 0.130 | 1.315 | 0.000 | 0 | 1 |
| | *hoxa11* | Tel a/x | 0.28 | 0.047 | 0.164 | 0.231 | 10000.000 | 0.030 | 8.46542 | 7.60E-02 |
| | | Tel b | ∞ | 0.034 | 0.000 | 10.000 | 10000.000 | 0.092 | 7.38577 | 1.28E-01 |
| | *hoxa13* | Tel a/x | 0.03 | 0.029 | 0.857 | 0.037 | 0.416 | 0.000 | 0 | 1 |
| | | Tel b | ∞ | 0.019 | 0.000 | 10.000 | 10000.000 | 1.000 | 0.328994 | 1 |
| HOXB | *hoxb1* | Tel a | 0.04 | 0.008 | 0.187 | 0.026 | 2.398 | 0.000 | 0 | 1 |
| | | Tel b/y | ∞ | 0.044 | 0.000 | 10.000 | 10000.000 | 0.093 | 7.11315 | 1.45E-01 |
| | *hoxb3* | Tel a/x | ∞ | 0.015 | 0.000 | 0.265 | 33.542 | 0.020 | 7.37897 | 1.22E-01 |
| | | Tel b/y | 0.10 | 0.027 | 0.257 | 0.211 | 0.797 | 0.000 | 0 | 1 |
| | *hoxb5* | Tel a/x | 0.18 | 0.014 | 0.080 | 0.149 | 1.129 | 0.000 | 0 | 1 |
| | | Tel b/y | 0.40 | 0.019 | 0.046 | 0.481 | 0.719 | 1.000 | 0 | 1 |
| | *hoxb6* | Tel a/x | ∞ | 0.031 | 0.000 | 0.269 | 0.718 | 0.000 | 0 | 1 |
| | | Tel b/y | 0.17 | 0.016 | 0.094 | 0.603 | 159.365 | 0.011 | 2.886 | 1 |
| | *hoxb8* | Tel a/x | 0.38 | 0.031 | 0.081 | 0.088 | 1806.220 | 0.008 | 3.92361 | 7.62E-01 |
| | | Tel b/y | 0.00 | 0.000 | 0.010 | 10.000 | 4.565 | 0.982 | -0.003186 | 1 |
| HOXC | *hoxc4* | Tel a/x | 0.23 | 0.016 | 0.067 | 0.223 | 1.511 | 0.000 | 0 | 1 |
| | | Tel b | **6.68** | 0.014 | 0.002 | 10.000 | 10000.000 | 1.000 | 0.020273 | 1 |
| | *hoxc5* | Tel a/x | ∞ | 0.034 | 0.000 | 0.763 | 1.286 | 0.000 | 0 | 1 |
| | | Tel b | ∞ | 0.022 | 0.000 | 10.000 | 10000.000 | 1.000 | -0.026997 | 1 |
| | *hoxc6* | Tel a/x | ∞ | 0.018 | 0.000 | 10.000 | 10000.000 | 1.000 | 0.136076 | 1 |
| | | Tel b | ∞ | 0.014 | 0.000 | 10.000 | 10000.000 | 1.000 | 0.28772 | 1 |
| | *hoxc8* | Tel a/x | 0.02 | 0.001 | 0.066 | 0.023 | 1.291 | 0.000 | 0 | 1 |
| | | Tel b | 0.13 | 0.016 | 0.127 | 8.319 | 8.294 | 1.000 | 0.009644 | 1 |
| | *hoxc9* | Tel a/x | ∞ | 0.006 | 0.000 | 10.000 | 10000.000 | 1.000 | 0.150502 | 1 |
| | | Tel b | 0.30 | 0.022 | 0.071 | 0.338 | 1.345 | 0.000 | 0 | 1 |
| | *hoxc10* | Tel a/x | 0.21 | 0.019 | 0.089 | 0.245 | 4.425 | 0.034 | 0.011136 | 1 |
| | | Tel b | 0.52 | 0.115 | 0.222 | 0.835 | 1166.530 | 0.050 | 6.52905 | 1.54E-01 |
| | *hoxc11* | Tel a/x | 0.15 | 0.010 | 0.069 | 0.251 | 0.321 | 0.417 | 0 | 1 |
| | | Tel b | 0.65 | 0.012 | 0.019 | 0.180 | 0.096 | 0.067 | 0 | 1 |
| | *hoxc12* | Tel a/x | 0.05 | 0.012 | 0.256 | 0.043 | 0.067 | 1.000 | 0 | 1 |
| | | Tel b | ∞ | 0.013 | 0.000 | 10.000 | 10000.000 | 0.037 | 3.63108 | 8.51E-01 |
| | *hoxc13* | Tel a/x | 0.03 | 0.004 | 0.133 | 0.038 | 0.240 | 0.000 | 0 | 1 |
| | | Tel b | ∞ | 0.011 | 0.000 | 10.000 | 10000.000 | 1.000 | 0.281265 | 1 |
| HOXD | *hoxd4* | Tel a/x | ∞ | 0.005 | 0.000 | 10.000 | 67.714 | 0.999 | -0.013905 | 1 |
| | | Tel b | ∞ | 0.052 | 0.000 | 10.000 | 10000.000 | 0.106 | 0.968843 | 1 |
| | *hoxd9* | Tel a/x | 0.13 | 0.024 | 0.183 | 0.143 | 0.215 | 0.000 | 0 | 1 |
| | | Tel b | ∞ | 0.020 | 0.000 | 10.000 | 10000.000 | 0.077 | 1.18551 | 1 |
| | *hoxd11* | Tel a/x | 0.10 | 0.018 | 0.179 | 0.104 | 0.132 | 0.000 | 0 | 1 |
| | | Tel b | 0.02 | 0.009 | 0.454 | 0.043 | 0.955 | 0.000 | 0 | 1 |

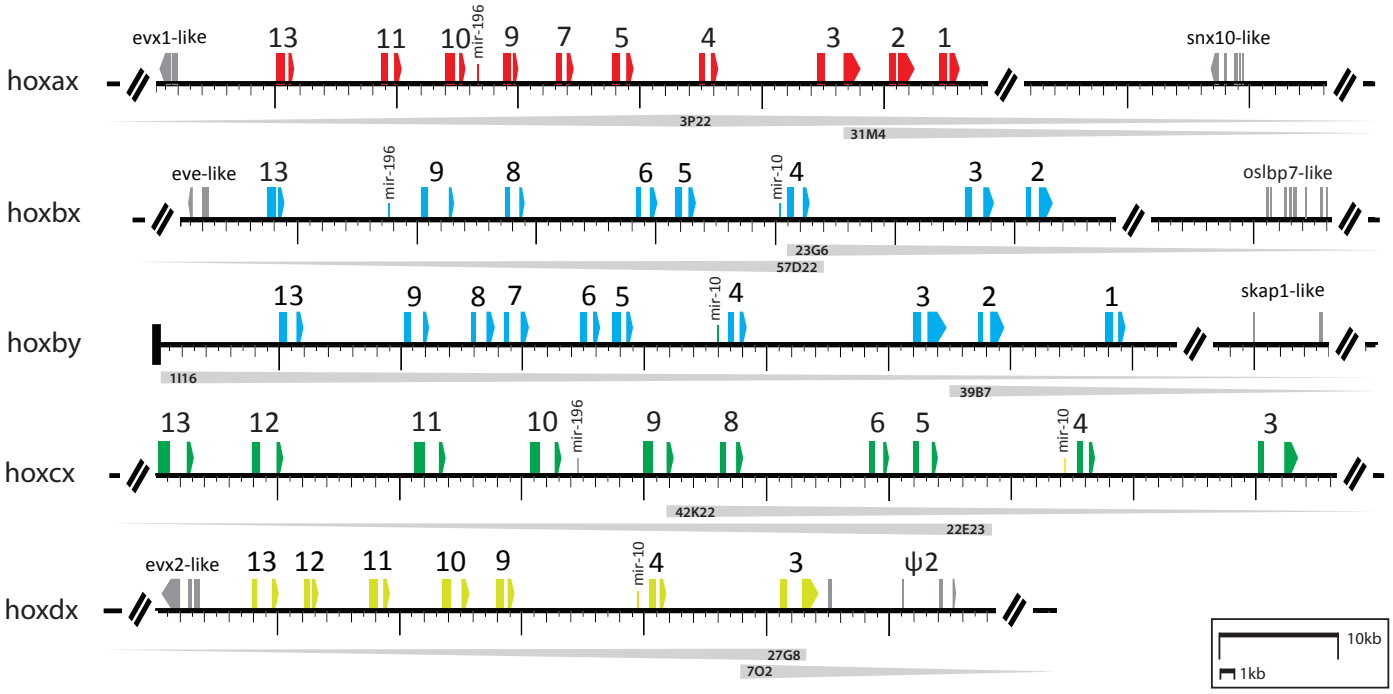



**Figure 2.1.** The Hox gene complement of *Pantodon buchholzi*

With only five clusters (hoxax, hoxbx, hoxby, hoxcx and hoxdx) containing a total of 45 full-length protein-coding Hox genes, *P. buccholzi* possess fewer *Hox* gene clusters than any other teleosts known, but retain a similar total number of individual genes. Exons of predicted Hox genes are depicted to scale and colour coded according to their cluster of origin(red = hoxa, blue = hoxb, green = hoxc, yellow = hoxd) alongside conserved microRNA genes in the mir-196 and mir-10 families, flanking Evx family genes and the most proximate non-homeobox flanking genes or pseudogenes (grey boxes) which could be identifed in the region which was sequenced. The IDs of BAC clones sequenced to assemble each of these scaffolds noted underlying each cluster. The total sizes of the assembled loci, and size of the region limited by the most 5' and 3' Hox gene coding exon for each cluster respectively are: hoxax (199,116 bp/55,231 bp), hoxbx (379,217 bp/66,149 bp), hoxby (268,422 bp/69,795 bp), hoxcx (246,971 bp/93,161 bp), and hoxdx (254,542 bp/45,374 bp).

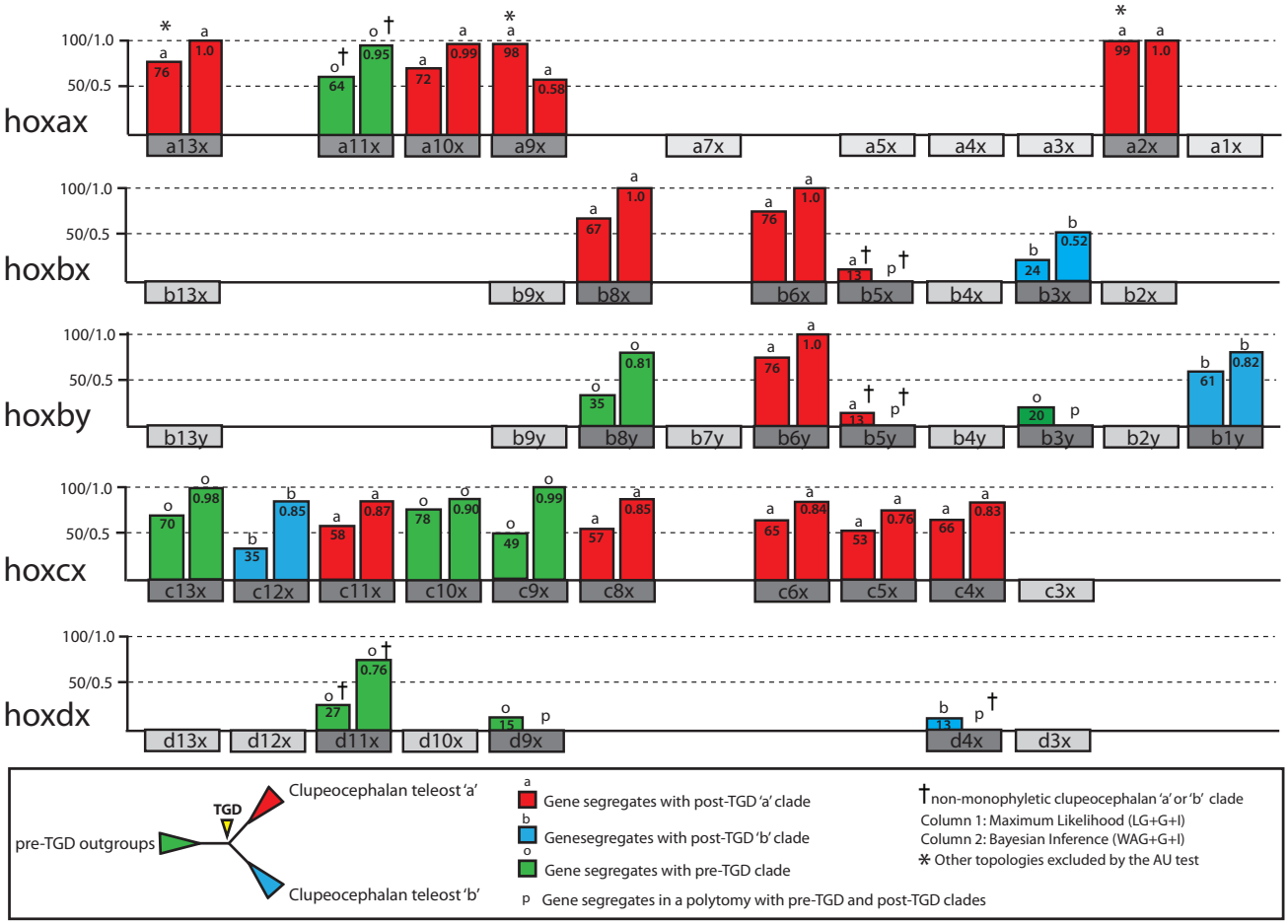
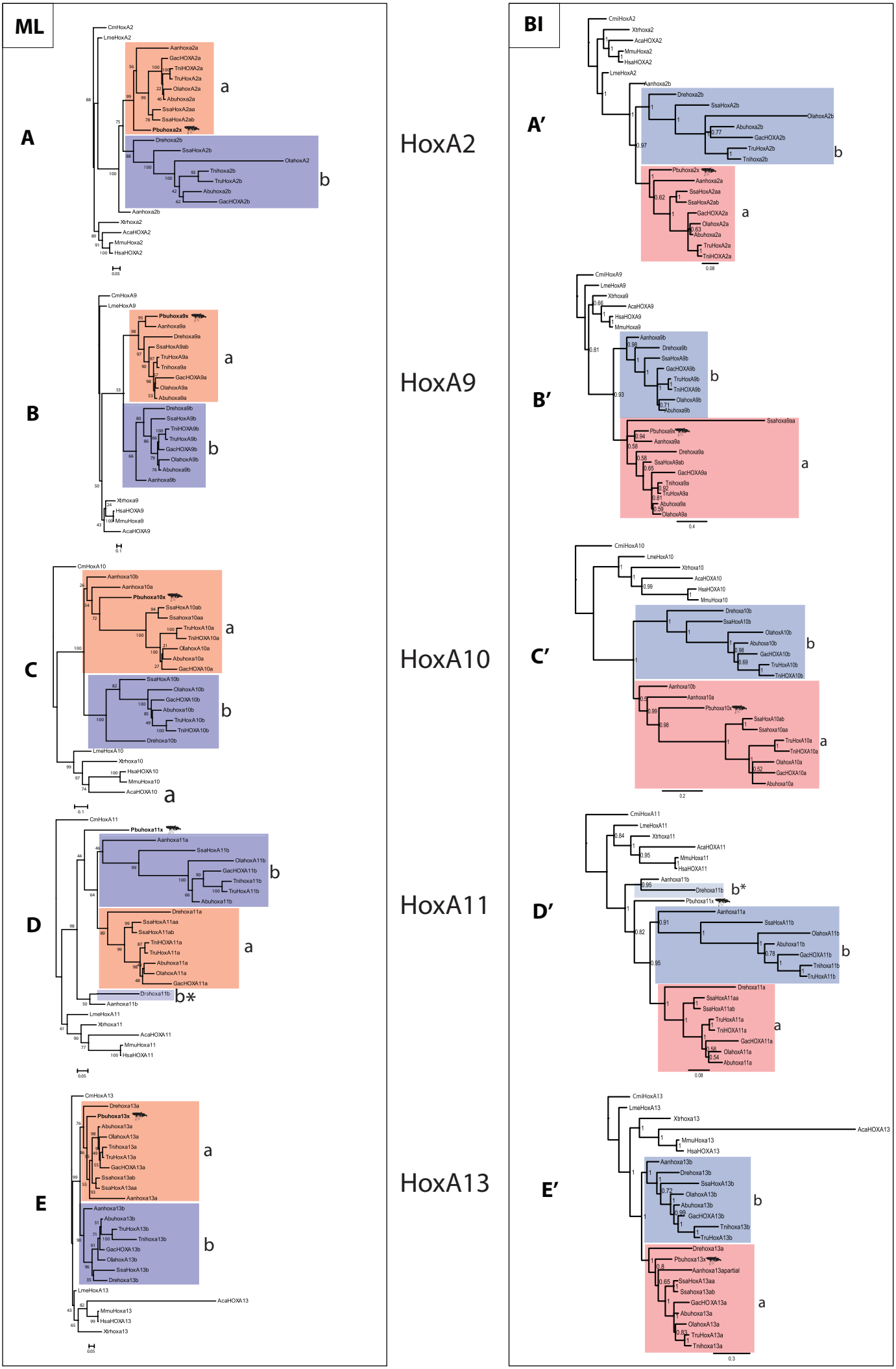
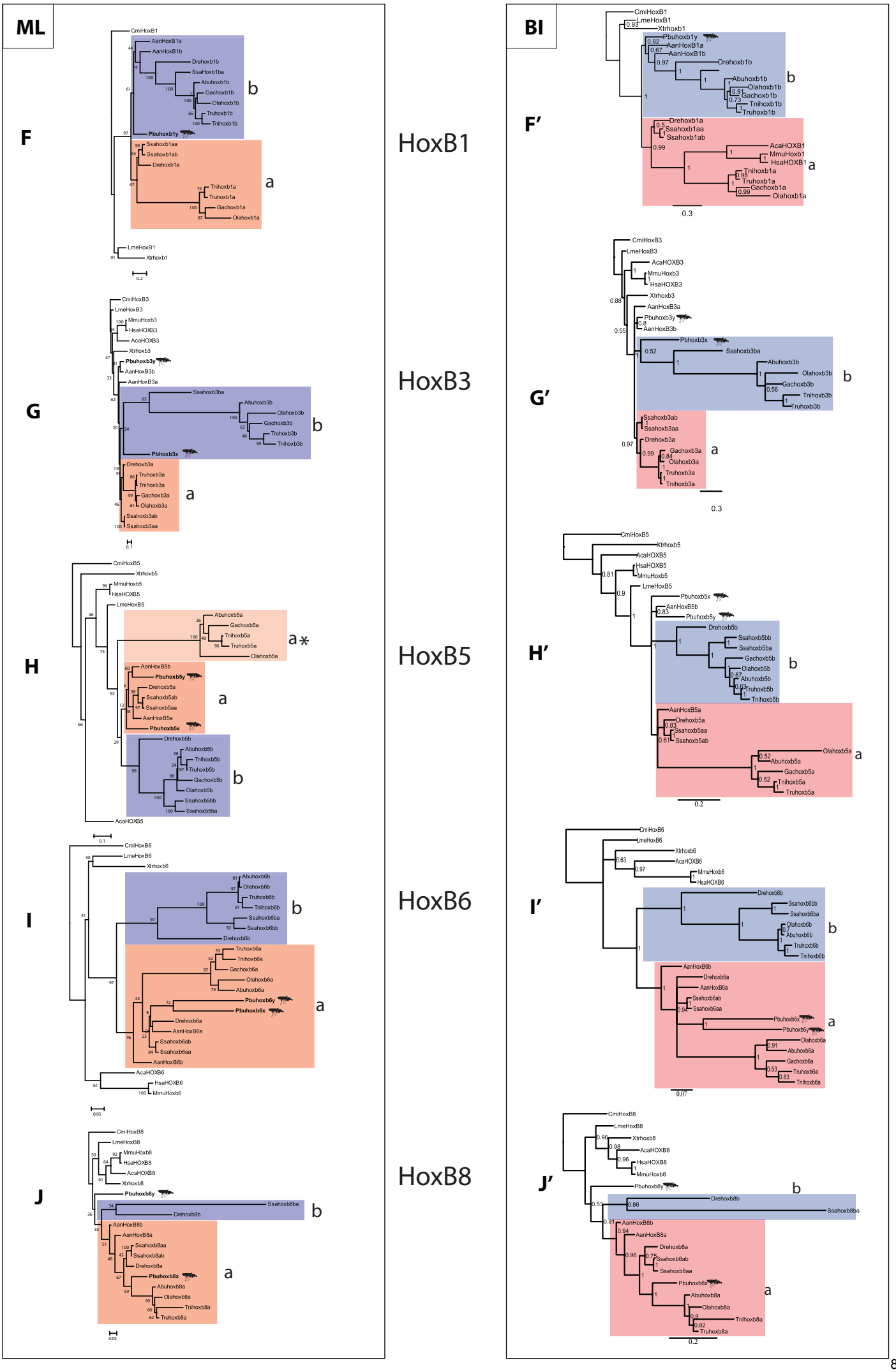
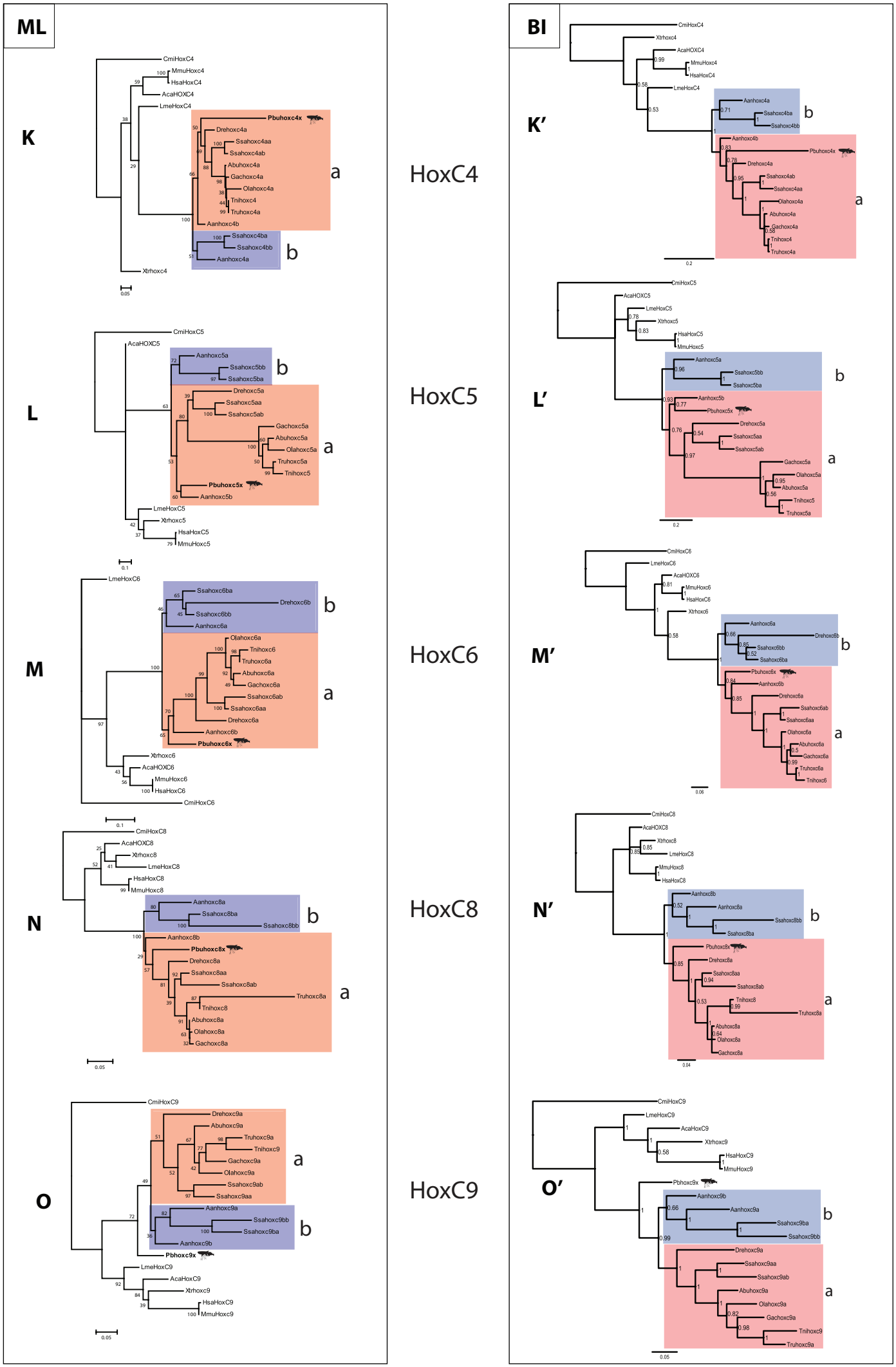


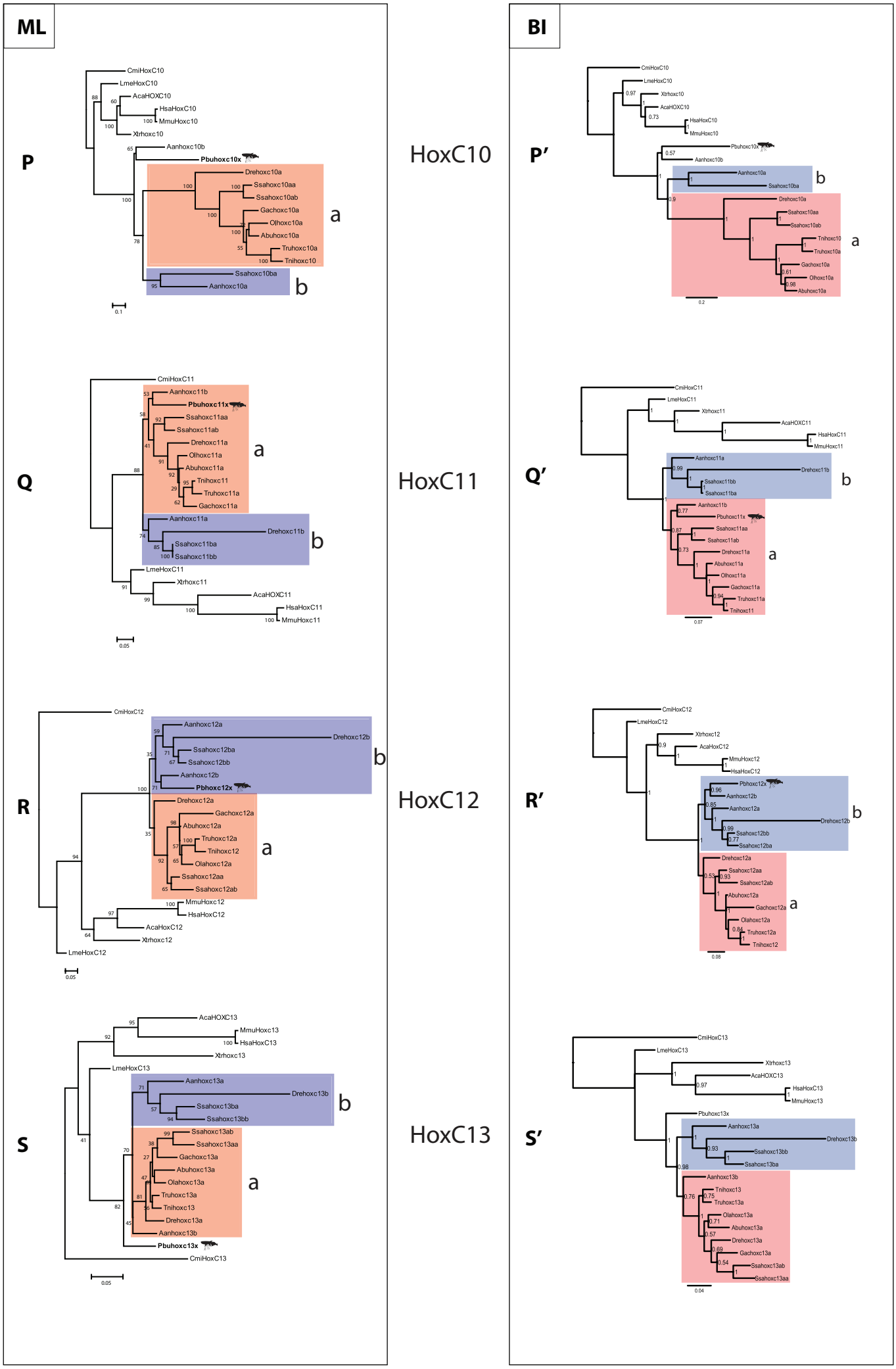

**Figure 2.2.** The Hox genes of *Pantodon buchholzi* do not segregate consistently with the post-TGD 'a' and 'b' orthology groups of other teleosts.

Phylogenetic trees were computed using maximum-likelihood and Bayesian methods with the *Hox* gene coding sequences of *P. buchholzi* and other vertebrates. Support values corresponding to clades containing *P. buchholzi* sequences are plotted as two columns above each individual orthology-informative gene in the Hox cluster schematic. The height of each column corresponds to either the bootstrapsupport value (left column) or the posterior probability (right column). The column colour corresponds to the clade containing the *P. buchholzi* sequence. We observe that across a single cluster, Hox genes where the *P. buchholzi* sequence clusters best with clupeocephalan post-TGD 'a' orthologs (red), are interleaved with those which cluster best with the post-TGD 'b' orthologs (blue), or outside the combined post-TGD 'a' and 'b' clades with non-teleost outgroups (green). Polytomies between the *P. buchholzi* sequence, the pre-TGD outgroups and post-TGD 'a' and 'b' clades occurring in the Bayesian trees are indicated with a letter 'p'. Under a classic pan-teleost TGD model we would expect all Pantodon genes on each Hox cluster to segregate with either the 'a' or the 'b' orthology groups rather than a mixture, and no sequences which segregate best as an outgroup to the TGD node.


ML
BI
A
A'
HoxA2
B
B'
HoxA9
C
C'
HoxA10
D
D'
HoxA11
E
E'
HoxA13
a
b
b*


ML
BI
F
F'
HoxB1
G
G'
HoxB3
H
H'
HoxB5
I
I'
HoxB6
J
J'
HoxB8
a
b
CmiHoxB1
LmeHoxB1
Xtrhoxb1
AanHoxB1a
AanHoxB1b
Drehoxb1b
SsaHoxb1ba
Abuhoxb1b
Gachoxb1b
Olahoxb1b
Truhoxb1b
Tnihoxb1b
Pbuhoxb1y
Ssahoxb1aa
Ssahoxb1ab
Drehoxb1a
Tnihoxb1a
Truhoxb1a
Gachoxb1a
Olahoxb1a
AcaHOXB1
MmuHoxb1
HsaHOXB1
CmiHoxB3
LmeHoxB3
MmuHoxb3
HsaHOXB3
AcaHOXB3
Xtrhoxb3
Pbuhoxb3y
AanHoxB3b
AanHoxB3a
Ssahoxb3ba
Abuhoxb3b
Olahoxb3b
Gachoxb3b
Truhoxb3b
Tnihoxb3b
Pbhoxb3x
Drehoxb3a
Truhoxb3a
Tnihoxb3a
Gachoxb3a
Olahoxb3a
Ssahoxb3ab
Ssahoxb3aa
CmiHoxB5
Xtrhoxb5
AcaHOXB5
HsaHOXB5
MmuHoxb5
LmeHoxB5
Pbuhoxb5x
AanHoxB5b
Pbuhoxb5y
Drehoxb5b
Ssahoxb5bb
Ssahoxb5ba
Gachoxb5b
Olahoxb5b
Abuhoxb5b
Truhoxb5b
Tnihoxb5b
AanHoxB5a
Drehoxb5a
Ssahoxb5aa
Ssahoxb5ab
Olahoxb5a
Abuhoxb5a
Gachoxb5a
Tnihoxb5a
Truhoxb5a
CmiHoxB6
LmeHoxB6
Xtrhoxb6
AcaHOXB6
MmuHoxb6
HsaHOXB6
Drehoxb6b
Ssahoxb6bb
Ssahoxb6ba
Olahoxb6b
Abuhoxb6b
Truhoxb6b
Tnihoxb6b
AanHoxB6b
Drehoxb6a
AanHoxB6a
Ssahoxb6ab
Ssahoxb6aa
Pbuhoxb6x
Pbuhoxb6y
Olahoxb6a
Abuhoxb6a
Gachoxb6a
Truhoxb6a
Tnihoxb6a
CmiHoxB8
LmeHoxB8
Xtrhoxb8
AcaHOXB8
HsaHOXB8
MmuHoxb8
Pbuhoxb8y
Drehoxb8b
Ssahoxb8ba
AanHoxB8b
AanHoxB8a
Drehoxb8a
Ssahoxb8ab
Ssahoxb8aa
Pbuhoxb8x
Abuhoxb8a
Olahoxb8a
Tnihoxb8a
Truhoxb8a


ML
BI
K
K'
HoxC4
L
L'
HoxC5
M
M'
HoxC6
N
N'
HoxC8
O
O'
HoxC9
a
b
CmiHoxC4
MmuHoxc4
HsaHoxC4
AcaHOXC4
LmeHoxC4
Pbuhoxc4x
Drehoxc4a
Ssahoxc4aa
Ssahoxc4ab
Abuhoxc4a
Gachoxc4a
Olahoxc4a
Tnihoxc4
Truhoxc4a
Aanhoxc4b
Ssahoxc4ba
Ssahoxc4bb
Aanhoxc4a
Xtrhoxc4
CmiHoxC5
AcaHOXC5
Aanhoxc5a
Ssahoxc5bb
Ssahoxc5ba
Drehoxc5a
Ssahoxc5aa
Ssahoxc5ab
Gachoxc5a
Abuhoxc5a
Olahoxc5a
Truhoxc5a
Tnihoxc5
Pbuhoxc5x
Aanhoxc5b
LmeHoxC5
Xtrhoxc5
HsaHoxC5
MmuHoxc5
LmeHoxC6
Ssahoxc6ba
Drehoxc6b
Ssahoxc6bb
Aanhoxc6a
Olahoxc6a
Tnihoxc6
Truhoxc6a
Abuhoxc6a
Gachoxc6a
Ssahoxc6ab
Ssahoxc6aa
Drehoxc6a
Aanhoxc6b
Pbuhoxc6x
Xtrhoxc6
AcaHOXC6
MmuHoxc6
HsaHoxC6
CmiHoxC6
CmiHoxC8
AcaHOXC8
Xtrhoxc8
LmeHoxC8
HsaHoxC8
MmuHoxc8
Aanhoxc8a
Ssahoxc8ba
Ssahoxc8bb
Aanhoxc8b
Pbuhoxc8x
Drehoxc8a
Ssahoxc8aa
Ssahoxc8ab
Truhoxc8a
Tnihoxc8
Abuhoxc8a
Olahoxc8a
Gachoxc8a
CmiHoxC9
Drehoxc9a
Abuhoxc9a
Truhoxc9a
Tnihoxc9
Gachoxc9a
Olahoxc9a
Ssahoxc9ab
Ssahoxc9aa
Aanhoxc9a
Ssahoxc9bb
Ssahoxc9ba
Aanhoxc9b
Pbhoxc9x
LmeHoxC9
AcaHoxC9
Xtrhoxc9
HsaHoxC9
MmuHoxc9


ML
BI
P
HoxC10
P'
Q
HoxC11
Q'
R
HoxC12
R'
S
HoxC13
S'
a
b

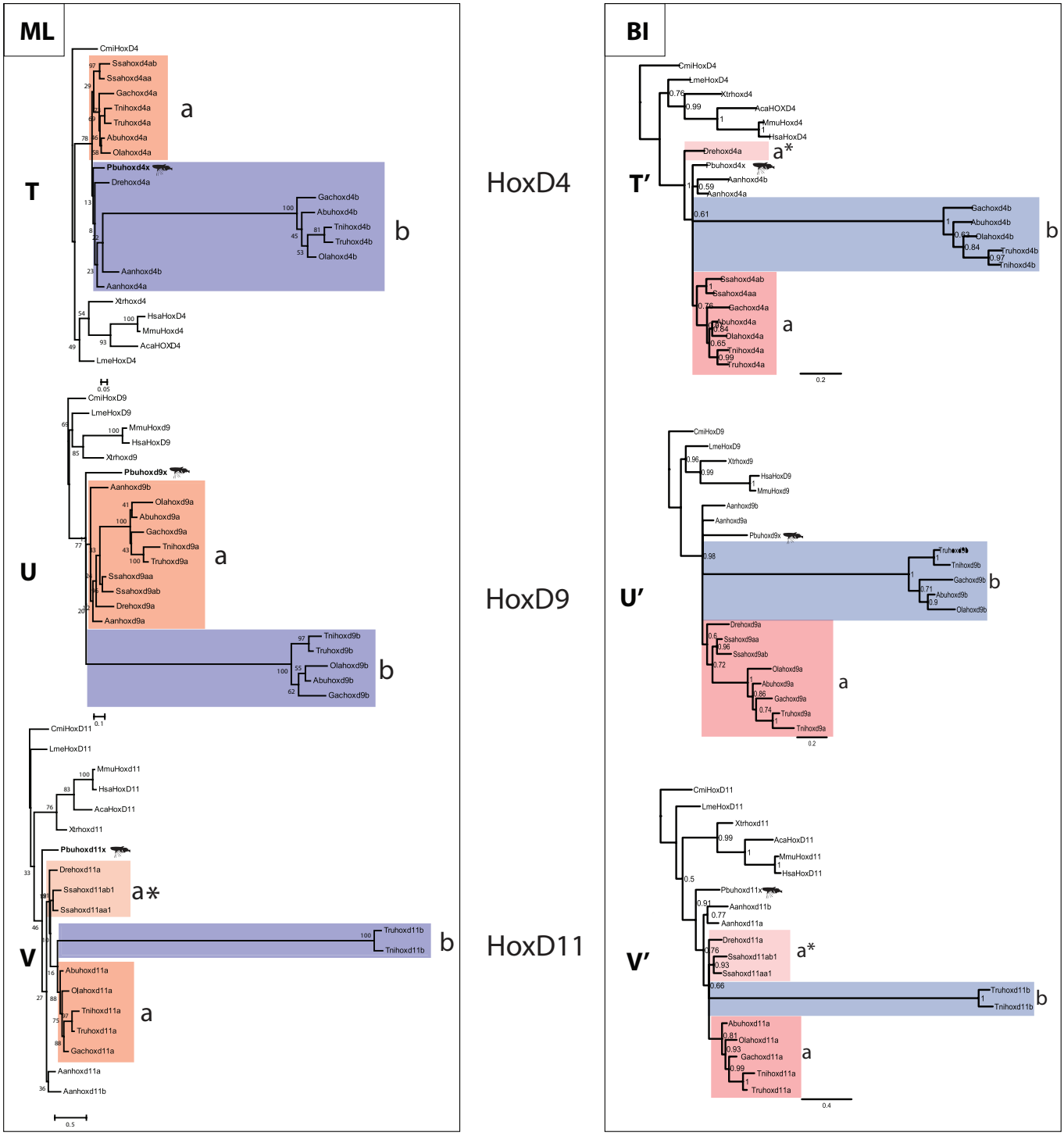

ML
BI
T
T'
U
U'
V
V'
HoxD4
HoxD9
HoxD11
a
a*
b
CmiHoxD4
LmeHoxD4
Xtrhoxd4
AcaHOXD4
MmuHoxd4
HsaHoxD4
Drehoxd4a
Pbuhoxd4x
Aanhoxd4b
Aanhoxd4a
Gachoxd4b
Abuhoxd4b
Olahoxd4b
Truhoxd4b
Tnihoxd4b
Ssahoxd4ab
Ssahoxd4aa
Gachoxd4a
Abuhoxd4a
Olahoxd4a
Tnihoxd4a
Truhoxd4a
CmiHoxD9
LmeHoxD9
Xtrhoxd9
HsaHoxD9
MmuHoxd9
Pbuhoxd9x
Aanhoxd9b
Aanhoxd9a
Olahoxd9a
Abuhoxd9a
Gachoxd9a
Tnihoxd9a
Truhoxd9a
Ssahoxd9aa
Ssahoxd9ab
Drehoxd9a
Tnihoxd9b
Truhoxd9b
Olahoxd9b
Abuhoxd9b
Gachoxd9b
CmiHoxD11
LmeHoxD11
Xtrhoxd11
AcaHoxD11
MmuHoxd11
HsaHoxD11
Pbuhoxd11x
Drehoxd11a
Ssahoxd11ab1
Ssahoxd11aa1
Truhoxd11b
Tnihoxd11b
Abuhoxd11a
Olahoxd11a
Tnihoxd11a
Truhoxd11a
Gachoxd11a
Aanhoxd11a
Aanhoxd11b

**Figure 2.3.** Indivdual *Hox* gene trees between *Pantodon buchholzi* and other vertebrates.
Both Maximum-Likelihood (ML) **(A-V)** and Bayesian (BI) **(A'-V')** trees were reconstructed from individual gene family alignments using all phylogenetically informative Hox genes from *P. buchholzi* (those retaining both paralogs in at least one teleost) and a collection of other vertebrates. ML trees were reconstructed with RAxML using the LG+gamma+I parameters and confidence was assessed by performing 1000 bootstrap replicates. BI trees were reconstructed using mrbayes (WAG+G+I parameters, 4 chains, 500000 generations, sampling every 100 generations and a 100000 generation burnin) and summarized as a 50% majority rule consensus tree. In the HoxA cluster five phylogenetically informative gene trees were reconstructed using ML and BI methods respectively: HoxA2 **(A, A')**, HoxA9 **(B, B')**, HoxA10 **(C, C')**, HoxA11 **(D, D')** and HoxA13 **(E, E')**. In the HoxB cluster five gene trees were reconstructed: HoxB1 **(F, F')**, HoxB3 **(G, G')**, HoxB5 **(H, H')**, HoxB6 **(I, I')**, and HoxB8 **(J, J')**. For the HoxC cluster 9 gene trees were reconstructed: HoxC4 **(K, K')**, HoxC5 **(J, J')**, HoxC6 **(M, M')**, HoxC8 **(N, N')**, HoxC9 **(O, O')**, HoxC10 **(P, P')**, HoxC11 **(Q, Q')**, HoxC12 **(R, R')** and HoxC13 **(S, S')**. For the HoxD cluster 3 gene trees were reconstructed: HoxD4 **(T, T')**, HoxD9 **(U, U')** and HoxD11 **(V, V')**. Red boxes enveloping clades demarcate the post-TGD 'a' orthology group, while blue boxes mark the post-TGD 'b' orthology group. Each clade is delineated at the most highly supported node inclusive of all clupeocephalan teleost 'a' or 'b' genes and may or may not include the *P. buchholzi* or *A. anguilla* sequences. Where the expected clupeocephalan post-TGD 'a' and 'b' orthology groups are atypically paraphyletic even in clupeocephala the clade is demarcated with an asterisk (*). Scale bars denote substitutions per site. Bootstrap values (ML) or posterior probability values (BI) are shown adjacent to each node. *Pantodon* genes are denoted by the silhouette ( ). Abbreviated taxa names correspond as follows: Cmi, *Callorhinchus milii*; Lme, *Latimeria menadoensis*; Xtr, *Xenopus tropicalis*; Aca, *Anolis caroliensis*; Mmu, *Mus musculus*; Hsa, *Homo sapiens*; Aan, *Anguilla anguilla*; Pbu, *Pantodon buchholzi*; Dre, *Danio rerio*; Ssa, *Salmo salar*; Ola, *Oryzias latipes*; Gac, *Gasterosteus aculeatus*; Abu, *Astatotilapia burtoni*; Tru, *Takifugu rubripes*; Tni, *Tetraodon nigroviridis*.

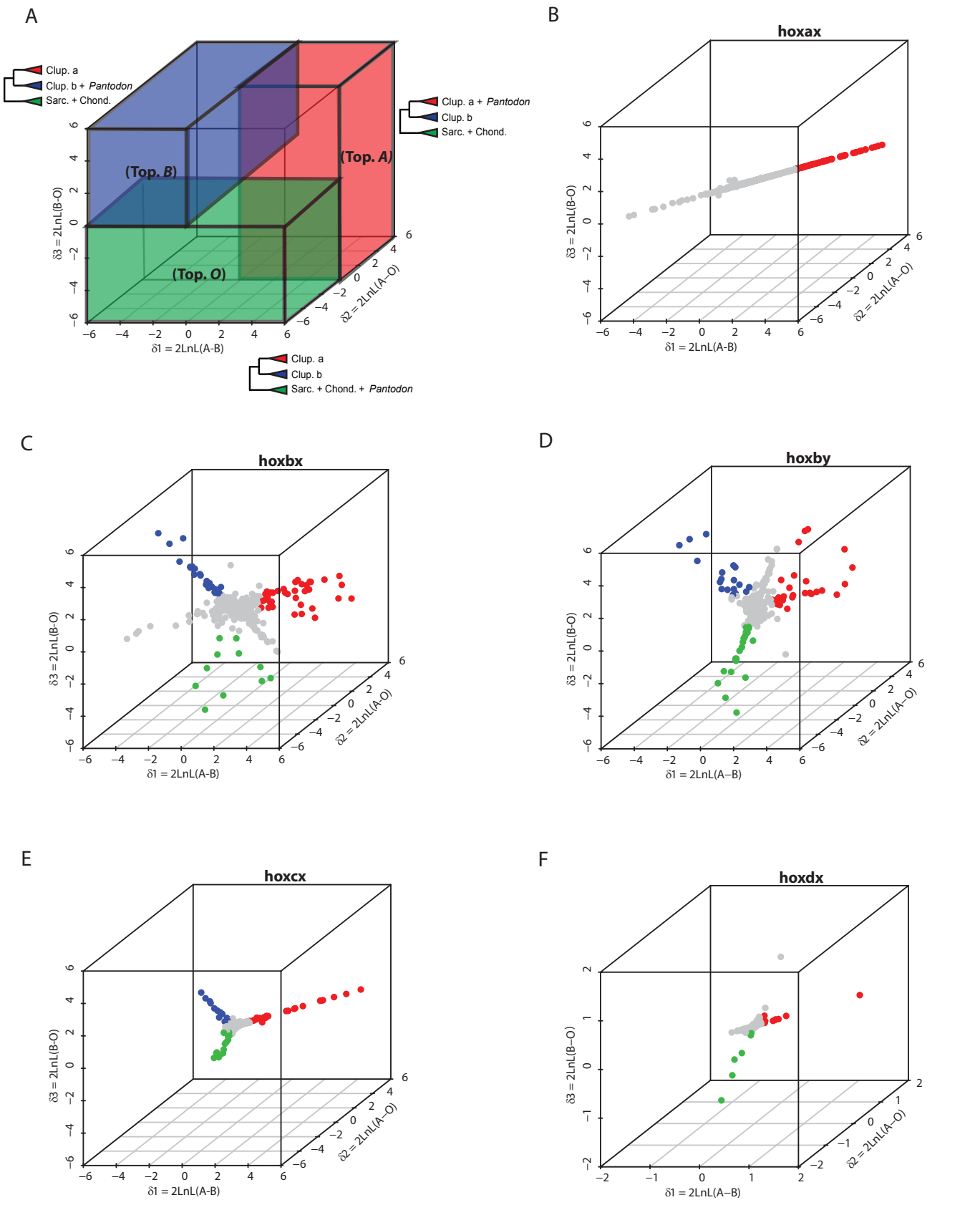



**Figure 2.4.** 3D site-likelihood ratio plots of concatenated Hox clusters reveal strong cluster-wide conflict between alternative constrained hypotheses of *Pantodon buchholzi* Hox cluster homology.

Three different constrained tree topologies representing different hypotheses of *P. buchholzi* Hox cluster homology were compared in a pairwise fashion under a maximum-likelihood framework to model the support for each hypothesis of homology across each site in whole Hox cluster concatenated alignments. Schematized outline of a site-likelihood ratio plot showing regions in the graph which support each topology depicted in (A). Actual site-likelihood ratio plots for each *P. buchholzi* cluster are shown for hoxax (B), hoxbx (C), hoxby (D), hoxcx (E) and hoxdx(F). Each axis plots the site-wise likelihood ratio difference between one pair of competing topologies. The x axis plots the likelihood ratio between topology A and topology B (δ1), the y axis plots the likelihood ratio between topology A and topology O (δ2), and on the z axis the likelihood ratio between topology B and topology O (δ3) is plotted. Each point represents a single amino acid site. Sites are coloured if the absolute magnitude of the corresponding site-likelihood ratio is more than 2 standard deviations greater than the mean. Sites which support topology A are coloured in red, sites which support topology B are blue are, while sites which support topology O are green.

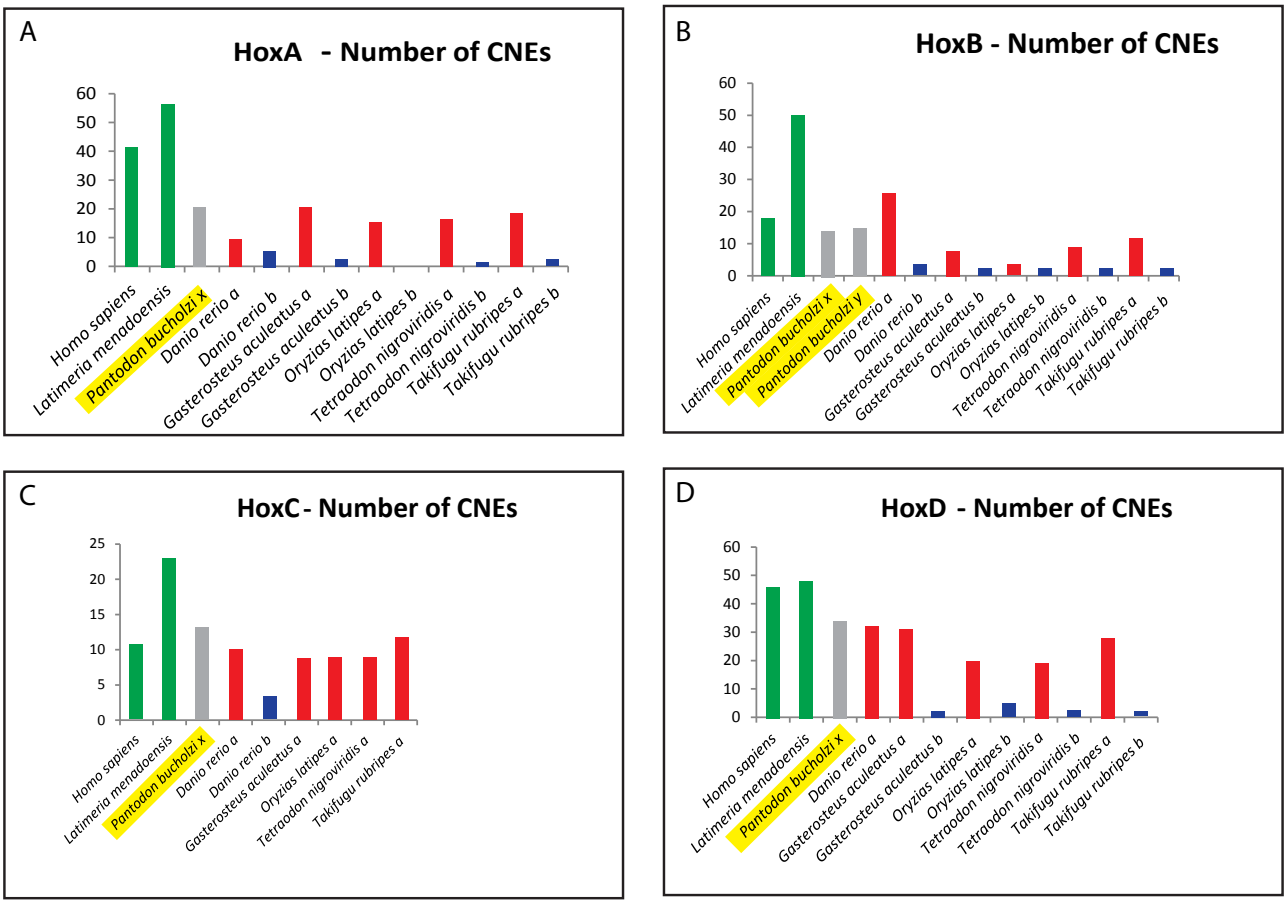



**Figure 2.5.** Ancient CNE conservation within Hox clusters.
Histograms showing the number of CNEs identified by multi VISTA alignment using *C. milii* as the baseline sequence. The sLAGAN algorithm and a cutoff of 70% identity over 100bp was used. Columns corresponding to pre-TGD sarcopterygian Hox cluster CNE counts are represented in green, the post-TGD clupeocephalan 'a' cluster counts are in red, and the post-TGD clupeocephalan 'b' clusters are in blue. *P. buchholzi* Hox cluster CNE counts are coloured in grey. While extreme asymmetry in the retention of ancient CNEs between the post-TGD 'a' and 'b' clusters of other teleosts is evident, the hoxbx and hoxby clusters of *P. buchholzi* retain nearly equal numbers of ancient CNEs (B).

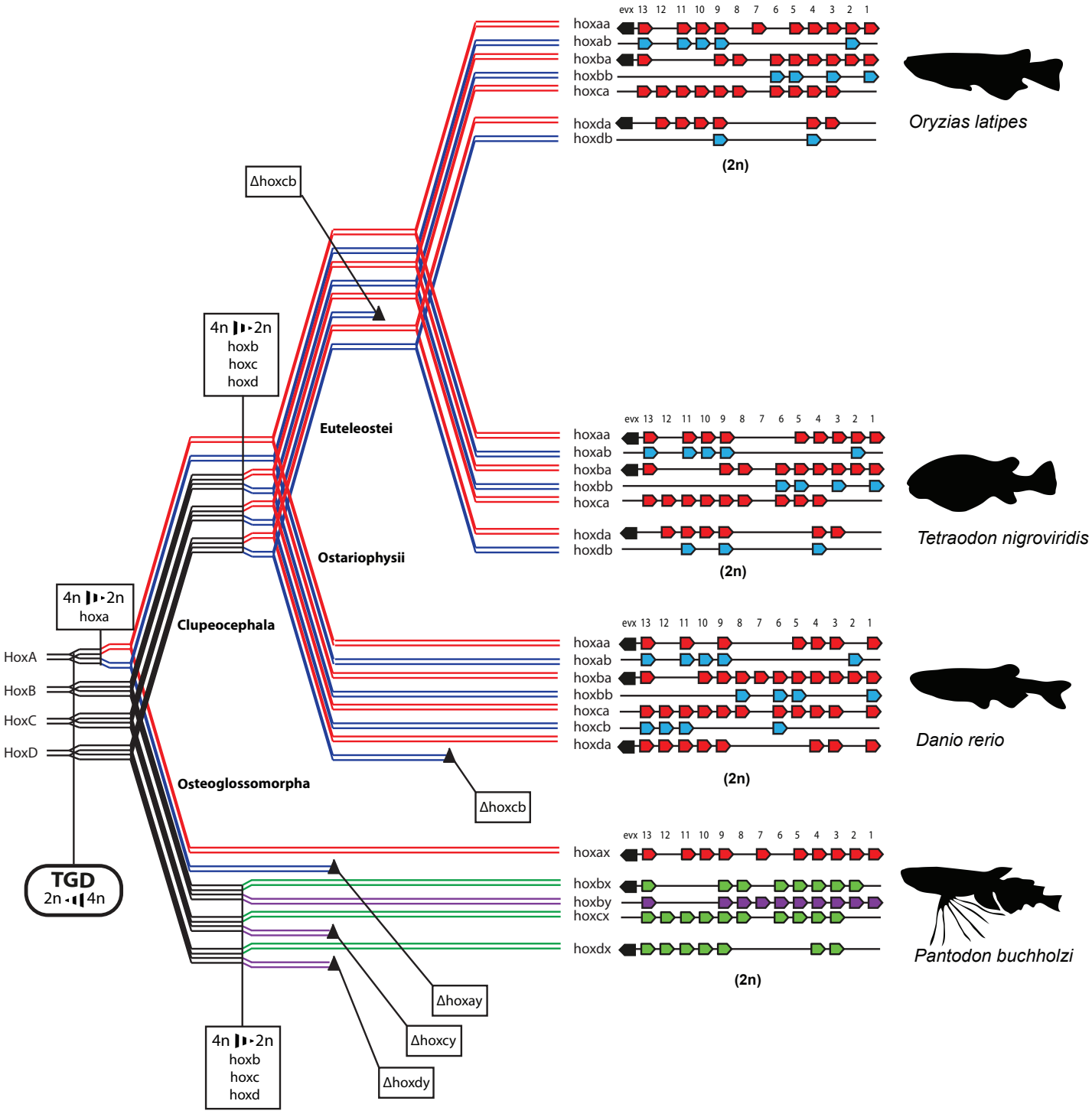



**Figure 2.6.** The evolution of teleost Hox gene clusters following the TGD outlining the relative timing of diploidization events and the speciation of major teleost subdivisions.

This model of Hox cluster evolution in teleosts illustrates the independent diploidization of Hox clusters following the TGD and the relative timing of cluster diploidization and speciation events. Each line in the phylogram represents an allele and the separation of pairs of lines accompanied with a change in colour represents the completion of the diploidization of this locus. In this model, the duplicated hoxa clusters diploidize first, before the last common ancestor of all teleosts. The remaining clusters diploidize later, and independently in Clupeocephala and Osteoglossomorpha. Whole Hox cluster losses (black triangles) are also mapped highlighting the massive Hox cluster losses in the *P. buchholzi* lineage.

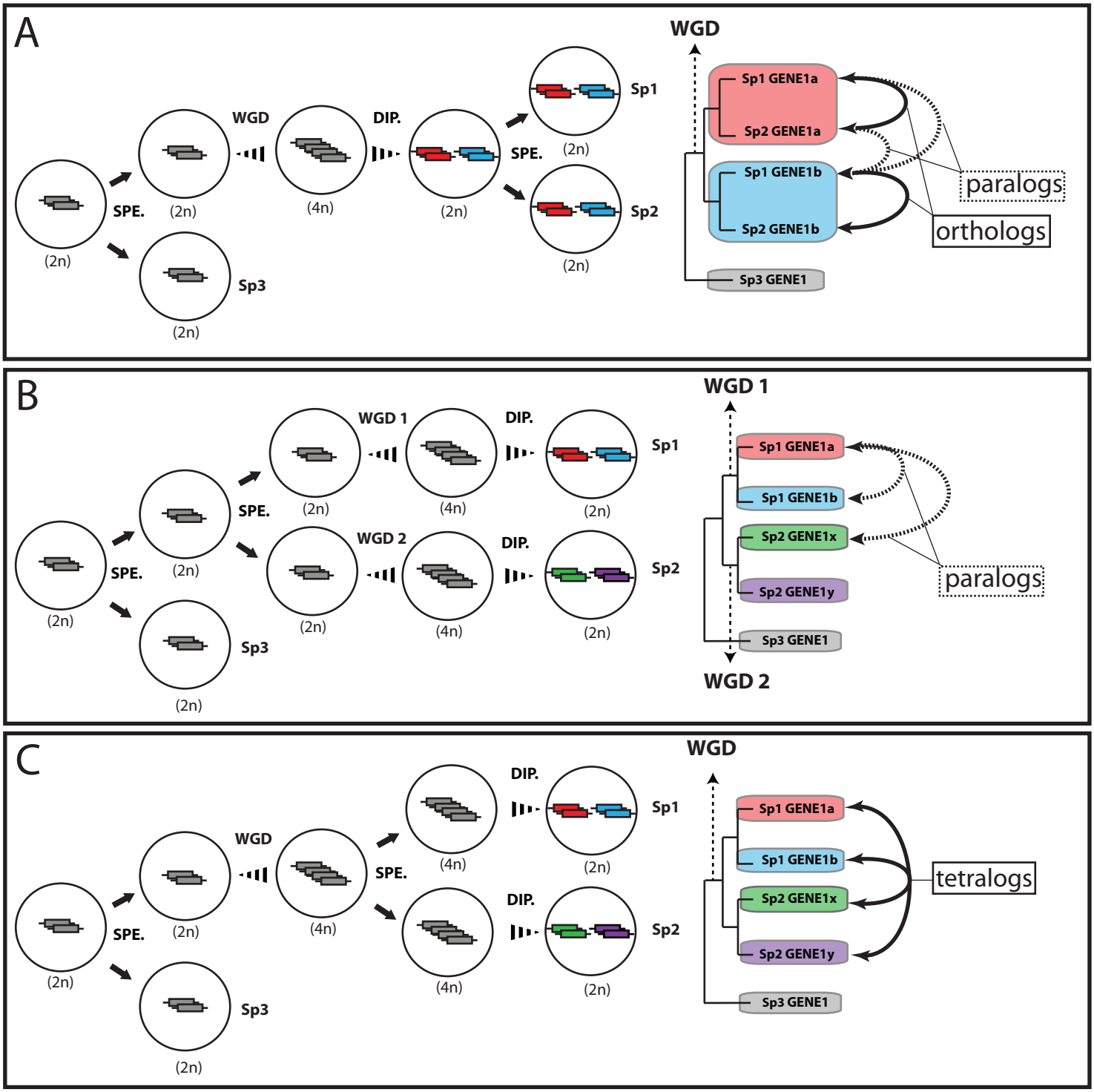

A
WGD
DIP.
SPE.
SPE.
(2n)
(2n)
(4n)
(2n)
(2n)
(2n)
(2n)
Sp1
Sp2
Sp3
Sp1 GENE1a
Sp2 GENE1a
Sp1 GENE1b
Sp2 GENE1b
Sp3 GENE1
paralogs
orthologs
B
WGD 1
WGD 2
DIP.
DIP.
SPE.
SPE.
(2n)
(2n)
(4n)
(2n)
(2n)
(2n)
(4n)
(2n)
Sp1
Sp2
Sp3
Sp1 GENE1a
Sp1 GENE1b
Sp2 GENE1x
Sp2 GENE1y
Sp3 GENE1
paralogs
WGD 1
WGD 2
C
WGD
DIP.
DIP.
SPE.
SPE.
(2n)
(2n)
(4n)
(4n)
(2n)
(2n)
(4n)
(2n)
Sp1
Sp2
Sp3
Sp1 GENE1a
Sp1 GENE1b
Sp2 GENE1x
Sp2 GENE1y
Sp3 GENE1
tetralogs

**Figure 2.7.** Types of homology relationships following WGD.

Schematic outlining the types of homology relationships which can exist between gene duplicates as a result of WGD, taking into account the relative timing of diploidization. (A) The classical model where both whole genome duplication (WGD) and full diploidization (DIP.) occur beforespeciation (SPE.) results in two types of homology relationships: orthology (e.g. Sp1 GENE1b and Sp2GENE1b) and paralogy (e.g. Sp1 GENE1a and Sp2 GENE1b). (B) The classical alternative scenario to the model presented in (A), where speciation occurs before two independent WGDs and subsequent diploidization, only results in the formation of paralogs (e.g. Sp1 GENE1a and Sp2 GENE1x). (C) A new scenario where a single whole genome duplication occurs prior to speciation but where diploidization occurs independently in each lineage can result in a novel type of homology relationship: tetralogy (e.g. Sp1 GENE 1a and Sp1 GENE 1b are tetralogous with Sp2 GENE 1x and Sp2GENE 1y). Following duplication in a diploid (2n) ancestor, and the resultant tetraploid (4n) speciated prior to full diploidization, recombination will cease independently and loci will make the transition from 4n alleles to 2n paralogs separately in each lineage. Shared derived mutations between them will then be able to accumulate independently between each pair of duplicates in each lineage, and no single duplicate in one lineage can be considered orthologous with a single duplicate in the other.

## Chapter 3

## Characterization of a *de novo* draft assembly the genome of the osteoglossomorph fish *Pantodon buchholzi*


### Abstract

The teleost superorder Osteoglossomorpha is a basal lineage which diverged from other teleosts soon after the teleost genome duplication (TGD). Genome scale data from this species-poor but ancient superorder will increase the range of comparative genomic analyses by providing a reference point closer to the TGD, and help resolve outstanding questions in the early evolution of teleosts. I have produced a draft assembly (comprising 445.4 Mb, 136,181 contigs, N50=6.9kb) of the nuclear genome of the osteoglossomorph fish *Pantodon buchholzi.* Approximately 11% of the assembly consists of transposable elements, and while DNA transposons are the most abundant (3.46%) and also most diverse (22 families) as observed in most clupeocephalan teleosts, a SINE element (SINE-V) is the single most abundant family (1.77%) and both the Hox-cluster TE invasion profile, and the genome-wide TE profile of *P. buchholzi* are notably different from most other teleosts. A *de novo* transcriptome assembly using the Trinity pipeline was also produced from adult brain and muscle tissue in order to facilitate the process of genome annotation. I annotated 47,894 partial protein-coding gene models *de novo* using the MAKER pipeline using both the *P. buchholzi* transcriptome and sequence similarity-based approaches. Overall, the draft genome of *P. buchholzi* produced here reveals several shared and derived features

compared with other teleosts and will provide a useful resource for further context-specific comparative analyses of the TGD on a genome-wide scale.

## Introduction

Teleost fishes are attractive models for investigating genome evolution due to their astonishing diversity and morphological complexity, boasting > 26,000 extant species and representing more than 50% of extant vertebrate biodiversity (Nelson 2006), and relatively small genome sizes compared with many other vertebrates (Gregory et al. 2007). The sequencing of the first teleost whole genomes was motivated largely by their utility for annotating the human genome and included two pufferfishes *Takifugu rubripes* (Aparicio et al. 2002) and *Tetraodon nigroviridis* (Jaillon et al. 2004). However, it was also due in large part to the comparative analysis of the *T. nigroviridis*, *T. rubripes* and human genomes that the occurrence of TGD could be confirmed through large scale conserved synteny analyses (Jaillon et al. 2004; Christoffels et al. 2004). The zebrafish, *Danio rerio*, is a frequently used developmental genetic and biomedical model (Lieschke & Currie 2007) and a genome project which had been practically useful but incomplete for almost a decade was recently completed and published (Hoi et al. 2013). Around the same time the genome projects for the Japanese rice fish (medaka) *Oryzias latipes* (Kasahara et al. 2007) and the stickleback *Gasterosteus aculeatus* (Jones et al. 2012) were underway and many subsequent analyses using each contributed largely to our understanding of adaptive diversification and large scale genomic changes which occurred following the TGD in the teleost lineage (Ravi & Venkatesh 2008). Each of these genomes were sequenced largely or entirely with costly and time-consuming Sanger sequencing methods (Sanger et al. 1977) which required collaborative efforts between multiple groups and funding agencies worldwide. Recently, thanks to the development of 'next-generation'

sequencing (NGS) technologies including Roche 454, Illumina HiSeq, and PacBio (Shendure & Ji 2008; Metzker 2009; Goodwin et al. 2016) the costs and time necessary to generate *de novo* genome assemblies from even species with relatively large and complex genomes such as many vertebrates has come into the reach of individual laboratories. The genomes of several further clupeocephalan teleosts including the cod *Gadus morhua* (Star et al. 2011), the platyfish *Xiphophorus maculeatus* (Schartl et al. 2013), common carp *Cyprinius carpio* (Xu et al. 2014), Mexican cave tetra *Astyanax mexicanus* (McGaugh et al. 2014), rainbow trout *Oncorhyncus mykiss* (Berthelot et al. 2014), and Atlantic salmon *Salmo salar* (Lien et al. 2016) to name just a few have all recently been sequenced using either Illumina HiSeq technology alone or in combination with Roche 454, PacBio, and traditional Sanger methods and resulted in draft genome assemblies of high enough quality for many types of large scale genome analyses. Notably, without high quality genetic or physical maps to anchor NGS-generated assemblies to chromosomes large, megabase-scale, linkage information is often lacking from many recent NGS-driven vertebrate genomes with notable exceptions including the Atlantic salmon (Lien et al. 2016; Phillips et al. 2009) and the gar (Braasch et al. 2016; Amores et al. 2011). Although the majority of the efforts towards sequencing of teleosts has concentrated heavily in clupeocephalans, the draft genome of the eel *Anguilla anguilla* was recently released using Illumina data and although lacking in contiguity compared to fishes with high-quality reference genomes will provide a valuable starting point for comparative genomics in the Elopomorpha (Coppe et al. 2010). The purpose of the work carried out in this chapter is to provide insights into the condition of the genome of the last of the three major teleostean

subdivisions which has currently not been explored, the Osteoglossomorpha (bony-tongues), with the aim of establishing a comparative genomic reference point at the deep divergence of the three major subdivisions, closer to the time of the TGD than currently available when comparing clupeocephalan teleosts alone. To this end, I carried out whole genome sequencing and produced a *de novo* draft assembly of the African Butterfly Fish *P. buchholzi*, a small freshwater species with a genome size estimated at 753Mb (Hinegardner & Rosen 1972). Using this *de novo* assembly I carried out comparative analyses of the total gene content, transposable element (TE) quantity and diversity, and single nucleotide variant (SNV) content compared with other teleosts.

Differences in transposable element (TE) content is a major contributor to the differences in genome sizes between species, and while a majority of accumulated TEs can be considered 'junk DNA' without a functional role, there has been a growing tide of interest in their potential as drivers of important functional evolutionary changes through a process known as exaptation, particularly through recruitment into *cis*-regulatory DNA (Feschotte 2008; Lol & Haussler 2012; de Souza et al. 2013). There is also evidence of TE exaptation into exons of protein coding genes (Burns & Boeke 2008). In teleosts, one example is the exaptation of a medaka-specific P-element type DNA transposon, *Izanagi*, into the regulatory circuitry of one paralog of the TGD-duplicated *dmrt1* genes (*dmrt1bY*), a master regulatory gene of male development in medaka (Herpin et al. 2010). In this case, the TE insertion was specific to only one paralog and led to the acquisition of a novel regulatory feedback mechanism. Some of the most poignant further examples of TE exaptation come from SINEs, including AmnSINE1 which is an

amniote specific element which has been exapted into cis-regulatory DNA in multiple instances and has shown to play a role in the regulation of the developmental gene *fgf8* (Sasaki et al. 2008; Nishihara et al. 2006). While the potential of TEs as drivers of evolutionary change is high, and several careful case studies of particular elements have shown this clearly in diverse vertebrates, many genome-wide studies lack functional evidence to support their claims for large-scale TE exaptation into cis-regulatory DNA and more functional studies are needed (de Souza et al. 2013). Novel genome-wide data on transposon content and diversity in *P. buchholzi* will shed new light from an important lineage which may help resolve some of these outstanding questions.

Despite their potential for long term utility, the vast majority of TE insertion events are neutral at best, and can be detrimental if the insertion interrupts a functional genomic element. The exclusion of repetitive elements is therefore predicted in *cis*-regulatory DNA rich regions of the genome where complex gene regulatory interactions necessitate the conservation of specific gene order and intergenic distances. Genes within vertebrate Hox clusters exhibit complex position dependent regulation of transcription which tightly controls their expression patterns in space and time during embryogenesis, leading to the canonical collinearity observed between their genomic position and their expression pattern along the anterior-posterior axis (Tschopp et al. 2009; Ravi et al. 2009; Deschamps 2007; Mallo & Alonso 2013). Correspondingly, vertebrate Hox clusters across a broad taxonomic range including tetrapods, lobe-finned fish, and chondrichthyans have been found largely to exclude repetitive elements from the intergenic regions between *hox* genes in the cluster and typical values range

from 1-10% interspersed repetitive elements within the Hox cluster (Fried et al. 2004; Ravi et al. 2009; Koh et al. 2003). Exclusion of repetitive elements is also observed from the single Hox cluster of the cephalochordate *Branchiostoma floridae*, suggesting that the tight functional regulatory constraints on *hox* genes predates the origin of vertebrates (Amemiya et al. 2008). There are, however, notable exceptions to this pattern including in the urochordate *Ciona intestinalis* in which the genes of the ancestral *hox* cluster have become dispersed in the genome (Ikuta et al. 2004), and the Japanese lamprey *Lethenteron japonicum* where Hox clusters consist of ~23-33% interspersed repetitive elements (Mehta et al. 2013). An additional outlier are the squamate reptiles whose expanded, TE rich Hox clusters have been implicated in morphological diversification of their anterior-posterior axis (Di-Poï et al. 2010). Nevertheless, while the repeat content of their Hox clusters might be high (e.g. 6-10% in *Anolis caroliensis*), they are still many times lower than the genome-wide average (34%) (Di-Poï et al. 2009).

Systematic study of TE content in the Hox clusters of teleosts and basal actinopterygians lags far behind tetrapods, with scattered reports of partial cluster repeat content using different methods in different lineages. It has been suggested that the TGD resulted in a relaxation of functional constraints on duplicated Hox clusters which was permissive to invasion of TEs, especially DNA class transposons. However, zebrafish are estimated to have only 1.39-3.33% TE (Di-Poï et al. 2009) or 3.1-8.5% TE (Fried et al. 2004) in their Hox clusters depending on methodology used, which are still far below the genome wide average of ~55%. The pre-TGD actinopterygian *Polypterus senegalus* HoxA cluster was also found to generally exclude TEs (Chiu et al. 2004)

while another study of partially sequenced *P. senegalus* Hox clusters concluded that they have been minimally invaded by TEs including some DNA, LINE, SINE, and LTR types (Raincrow et al. 2011).

## Materials and Methods

### Whole genome sequencing and assembly

DNA from 25mg of axial skeletal muscle and skin of one adult male specimen of *P. buchholzi* was digested for 8hrs at 50$^{o}$C in 50µg/ml proteinase K, and then extracted by PCIA followed by ethanol precipitation as described in detail in Chapter 2. Extracted DNA was used to prepare one paired-end library with a target insert size of 180bp at the Wellcome Trust Center for Human Genomics, Oxford (http://www.well.ox.ac.uk). The library was sequenced using the 2 x 100bp paired end Illumina HiSeq2000 protocol and produced 171,515,994 paired reads corresponding to 17.15 Gb of raw data. Quality metrics for the reads were determined using the program fastqc (Schmieder & Edwards 2011) (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). After initial evaluation, reads were quality filtered and trimmed using the program prinseq-lite (Schmieder & Edwards 2011). The quality cutoffs parameters used involved trimming of all N's from the 5' and 3' ends of the sequence, trimming all bases with a phred score < 20 from the 5' and 3' ends to a minimum length of 50bp, and finally discarding all reads which had an average phred score < 20. Exact duplicates from both the 5' and 3' direction were removed using the –derep parameter. The final filtered dataset comprised 164,278,554 million reads 50bp-100bp in length (16.14Gb), and was re-assessed with [fastqc] confirming reads no strong representation biases. The velvet assembler v1.2.10 (Zerbino & Birney 2008) was used to assemble the reads which passed quality control supplying a range of k-mers (33,37,47,57) as the velveth hash lengths and velvetg was called using the "–cov_cutoff auto" and "-min_contig_lgth 200" parameters in each case. As an initial measure of protein coding coverage in each

assembly, the software CEGMA (Parra et al. 2007) was used to search for a set of 248 conserved “core” eukaryotic genes using default parameters. Genome size estimations were made by counting kmers with the program jellyfish (Marçais & Kingsford 2011). Read mapping was performed with the software bowtie2 (Langmead & Salzberg 2012) and tophat2 (Kim et al. 2013). SNV and indels were called using the bowtie2 genome read re-mapping data using the samtools suite (Li et al. 2009) and vcftools (Danecek et al. 2011) using a stringent cutoff criteria (--min-alleles 2, --max-alleles 2, --minQ 30, --min-depth 4).

**Transcriptome sequencing and assembly**

Using 5-10mg of muscle tissue, or total brain tissue from a single individual of indeterminate sex, total cellular RNA was extracted by homogenization in TRIzol® (Invitrogen) using a TissueLyzer (Qiagen) using 2 x 2minute cycles at 20 Hz with a 50mm steel bead at room temperature followed by chloroform extraction and ethanol precipitation according to the TRIzol® protocol’s instructions. Total RNA quality was assessed using an Agilent 2100 Bioanalyzer (Agilent Technologies) and confirmed to have an RIN value > 8. Two 100bp paired-end libraries were prepared according to Illumina’s HiSeq2000 paired-end methodology at the Wellcome Trust Center for Human Genomics in Oxford, UK. Reads from the brain and muscle transcriptomes were quality checked with fastqc and filtered and trimmed with prinseq (Schmieder & Edwards 2011) using a quality cutoff threshold of Q>20, and end-trimmed to remove any Ns and bases with a Q-score <20. Assembly was carried out with the program Trinity version r20140717 using default parameters (Grabherr et al. 2011). The raw transcriptome assembly output (303,353 contigs) was clustered using cd-hit-est (Fu et al. 2012) with

an alignment threshold of 95% over at least 50% of the shortest putative EST length and then screened for putative homologs to vertebrate protein coding genes by blastx against the UniProtKB-SwissProt protein database ($E<10^{-6}$) which resulted in a final set of 23,636 putative protein-coding transcripts with some degree of recognizable sequence homology to other manually annotated vertebrate sequences.

**Repeat analysis**

The program RepeatMasker (Smit et al. 1996) (www.repeatmasker.org) was run on both the Pbuc_v37 genome assembly, as well as separately using the *P. buchholzi* Hox scaffolds (KM102157–KM102161), using a custom repeat library generated by merging all vertebrate interspersed sequences extracted from the Repbase version 20140131 repeat library (Bao et al. 2015) with a *P. buchholzi* specific library of *ab initio* repeat predictions prepared using the RepeatModeler pipeline on the Pbuc_v57 assembly of the *P. buchholzi* genome with the expectation that the higher kmer size assembly would resolve more unique repeats and improve classification and annotation. The RepeatMasker accessory script buildSummary.pl was used to extract transposable element family classification data from the output of the RepeatMasker runs and classify them.

**MAKER genome annotation**

The genome annotation pipeline MAKER v2.31.8 (Cantarel et al. 2008) was used to annotate the *P. buchholzi* genome. Using the set of 148 full-length “core” eukaryotic genes detected by CEGMA in the Pb_v37 assembly, combined with the 45 complete *hox* gene sequences previously annotated (Chapter 2; Martin & Holland 2014) I

performed an initial training run to calibrate the hidden markov model which would eventually be used by MAKER to define protein coding genes by using the *ab initio* gene predictions produced using AUGUSTUS (Stanke et al. 2004) and SNAP (Korf 2004). The combined Repbase plus *P. buchholzi*-specific Pb_v57 derived RepeatModeler library was used to repeat-mask the genome. As protein evidence, all vertebrate proteins in the Refseq database (Pruitt et al. 2014) plus the predicted proteins from the gar genome (Braasch et al. 2016) were used. As EST evidence, the Pb_BM_Trinity Uniprotkb-Swissprot positive transcriptome assembly was supplied. Following successful completion of the first round of predictions, all MAKER models supported by evidence were used to refine markov-models used by the SNAP and AUGUSTUS *de novo* algorithms, and an additional iteration of the software was run.

## Results

### *P. buchholzi* genome assembly and quality assessment

The only available genome size estimate for *P. buchholzi* is based on flow cytometry and suggested its haploid genome size is 0.77pg = 753Mb (Hinegardner & Rosen 1972). I estimated genome size using Waterman's kmer counting method (Li & Waterman 2003; Liu et al. 2013) frequently employed to estimate genome sizes from next generation whole genome sequencing projects in vertebrates, including several teleosts (Li et al. 2010; Wang et al. 2015; Xu et al. 2014). Briefly, the formulas used were:

$$(1)\ C_{genome} = (C_{kmer} \times L)/(L - k + 1)$$

$$(2)\ G_{bp} = C_{genome}/D_{bp}$$

Where $C_{genome}$ = estimated genome coverage, $C_{kmer}$ = peak kmer coverage, L = mean read length, k = kmer length, $G_{bp}$ = estimated genome size in bp and $D_{bp}$ = total dataset size in bp.

Using this method with a range of kmers (17, 27, 37, 47) I estimated the *P. buchholzi* genome size at ~ 550.2Mb (522-586Mb) and that the actual genome coverage depth at ~29X (30.9X-27.5X) (Figure 3.1). This is significantly smaller than the estimate from flow cytometry but unsurprising as use of flow cytometry to estimate teleost genome size has often produced higher estimates than either kmer-counting or assembly based genome size estimates (e.g. zebrafish flow cytometry estimate = 1.8Gb, GRCz10 assembly estimate = 1.46Gb).

In an effort to optimize the assembly, I tested a range of k-mers (33, 37, 47, 57) using the velvet assembler (Zerbino 2010) and compared general metrics for each resultant assembly (all contigs > 200bp) including total number, N50, N90, largest contig size, and total assembly size between them (Table 3.1). Using k=57 produced the largest total assembly size (472.40 Mb), but also had the highest number of contigs (215,510) and the lowest N50/N90 scores (5.9kb/1.2kb). In contrast, using k=37 produced the both the highest contig N50/N90 scores (6.9kb/1.4kb) and the lowest number of contigs (136,818) but produced a total assembly (445.40Mb) which was 27Mb smaller than using k=57. Using kmer values of k=33 and k=47 produced metrics intermediate in value between k=37 and k=57.

Because the primary objective of this chapter was to assemble a draft of the *P. buchholzi* genome of sufficient quality for mining protein-coding genes for use in comparative evolutionary analyses with other teleosts, I sought to assess how well the different assemblies performed in covering the protein-coding portion of the *P. buchholzi* genome. The first approach I took was to scan each of the assemblies for a set of 248 highly conserved so-called 'Core Eukaryotic Genes' or CEGs using the CEGMA pipeline (Parra et al. 2007; Parra et al. 2009). This approach is frequently used in order to evaluate how representative of the protein-space a particular assembly is by searching for set of 248 CEGs considered highly unlikely to have been lost from the genome of any eukaryote. This approach has been successfully implemented in several recent vertebrate whole genome projects spanning diverse taxa from across the vertebrate tree of life including coelacanth (Amemiya et al. 2013), lamprey (Smith et al. 2013a) and the ostariophysian teleost *A. mexicanus* (cavefish) (McGaugh et al. 2014). In each of

the four *P. buchholzi* genome assemblies I detected between 88-91% of the 248 CEG set, of which 66-68% were considered full length (aligned over 75% of the reference length with > 90% identity). These results suggest a good protein recovery rate was achieved for all of the different assemblies, which are comparable to results obtained from the recently sequenced genome of the cavefish which captured 69% full and 95% partial CEGs (McGaugh et al. 2014). Because each of the assemblies performed comparably well in the CEGMA analyses, I selected the k=37 (henceforth 'Pbuc_v37') assembly for all subsequent analyses because of its overall higher N50/N90 and lower number of contigs, at the expense of the single additional CEG which was detected in the k=57 assembly.

The 445.4Mb Pbuc_v37 assembly therefore corresponds to 76-85% of the estimated genome size measured by kmer counting methods, or 59% of the genome size estimated based on flow cytometry. The N50 and N90 contig sizes of the Pbuc_v37 mean that 50% and 90% of the assembly is included in contigs larger than 6.9kb, and 1.4kb respectively. Although the N50 and N90 values are reasonable, in a frequency distribution of contig lengths I can see that there are a large number of small contigs, and the mean contig size is 3.3kb (Figure 3.2 A). All of the assemblies including Pbuc_v37 returned similar total GC content values of ~42% GC, which is very close to the reported value of 43.3% from CsCl ultracentrifugation (Bucciarelli et al. 2002), and in the range observed in other teleosts (e.g. medaka 39.41% GC, *T. nigroviridis* 45.38% GC). The distribution of per-contig %GC in the Pbuc_v37 assembly resembled a normal distribution with a mean of 42% GC (Figure 3.2 B).

By mapping the original genomic reads used in the assembly process back to the Pbuc_v37 assembly I obtained an overall mapping rate of 83.12%, and a concordant paired-end mapping rate of 76.67% (Table 3.2). This suggests that Pbuc_v37 assembly represents approximately 83% of the raw read data in the genomic library, and that accuracy of the assembly is reasonably good. In order to assess how representative of the transcriptome space the Pbuc_v37 assembly was, I used reads from two tissue specific 100bp paired-end Illumina RNAseq libraries (one from brain, and one from muscle from the same unsexed adult individual), and aligned them to the Pbuc_v37 assembly using tophat2 (Kim et al. 2013). This yielded an overall read mapping rate of 75.3% and 72.1% for brain and muscle respectively, and a concordant paired-end read mapping rate of 66.2% and 63.2% (Table 3.2). This indicates that a relatively high proportion of the transcriptomes of these two tissues are represented in the draft genome assembly. The smaller value of the concordant alignment rate is possibly a result of exons being broken on separate contigs and hence causing reads to align discordantly. Overall, the results of the read mapping were consistent with the kmer based genome size estimates and CEGMA protein coding content analysis, and support the idea that the true genome size is closer to the value reported here than the flow-cytometry estimate.

In summary, a *de novo* draft assembly of the *P. buchholzi* genome which corresponds to 76-85% of the estimated true genome size, and has a protein-coding gene representation of 67-91% and a partial transcriptome representation of 72-75%. These metrics suggest that this draft genome is of sufficient quality to carry out a range of useful downstream genome-wide analyses.

## Single nucleotide variants in the *P. buchholzi* genome

The highest rate of heterozygosity identified in a vertebrate was found in a teleost fish, the medaka *O. latipes* (0.034) (Kasahara et al. 2007). In general, it appears that teleost fish have a relatively high single nucleotide variant (SNV) rate. Using the bowtie2 realigned read data, and keeping only high quality variants filtered using the samtools-vcftools suite (alignment Q>30, minimum alignment depth 4, exactly 2 alleles) I identified 718,198 putative single nucleotide variants (SNVs) in *P. buchholzi*. I further identified 199,195 putative short indels (<10bp). Based on the assembly size of 445.4Mb calculated using the kmer method and assuming an even distribution, this corresponds to an estimated total rate of heterozygosity of ~2.06 x $10^{-3}$ or a SNV heterozygosity rate of ~1.61 x $10^{-3}$. This SNV heterozygosity rate is comparable the rate of SNVs identified in other teleosts including stickleback *G. aculeatus* (1.43 x $10^{-3}$) (Jones et al. 2012) and a wild-caught male grass carp *Ctenopharyngodon idellus* (1.36 x $10^{-3}$) while it is notably higher than a cultured gynogenetic female grass carp (7.16 x $10^{-4}$), and slightly lower than the large yellow croaker *Larimichthys crocea* (3.58 x $10^{-3}$) (Wu et al. 2014). Overall this suggests that the heterozygosity rate in the wild-caught male specimen of *P. buchholzi* is within the range observed in other wild populations of teleosts and offers a valuable resource for future population genetic studies.

## Transposable element content in the *P. buchholzi* genome

In light of the potential for TE exaptation as an important evolutionary mechanism particularly following WGD events, and the observation that teleosts have a more diverse representation of TE families than other vertebrates (Chalopin et al. 2015a),

some have speculated that TE invasion and subsequent exaptation could have been facilitated by the TGD through a relaxation of selective constraints afforded by the redundancy of newly duplicated genomic loci (Volff 2005). This, in turn, could be yet another driver of morphological diversification in teleosts, and the asymmetric exaptation of TEs between duplicated loci associated with sex-determining regions (as seen in the case of *dmrt1*) yet another driver of speciation.

In this section of Chapter 3, I sought to catalog the transposable element spectrum of the *de novo* draft assembly of *P. buchholzi* to verify the generality of the assumption that TE diversity is higher in all teleosts (and not just clupeocephalan teleosts) and could therefore be directly correlated with the TGD. I also aimed to find cases where the presence or absence of specific TE families might hold some phylogenetic information pertinent to the question of whether Osteoglossomorpha, Elopomorpha, or a clade containing both are the sister lineage to Clupeocephala. Furthermore, the development of a reliable species-specific catalog of repeats is a necessary prerequisite for downstream protein-coding gene predictions carried out in this chapter and used to address specific questions related to the retention of post-TGD paralogs in Chapter 4.

In order to survey the transposable element diversity in *P. buchholzi* I examined the Pbuc_v37 assembly using the repeat-annotation pipeline RepeatMasker, which offers a good combination of sensitivity, specificity, and ease of use amongst a wide variety of repeat detection software now available to the bioinformatics community (Lerat 2010). RepeatMasker is a similarity-search type repeat-finder which works by leveraging the accumulated diversity of previously catalogued transposable elements

deposited in the Repbase database and searches for similar elements in a target genome by running sequential similarity searches (RMBlast) across the whole genome chunk by chunk. I made a first-round analysis of the Pbuc_v37 assembly using default parameters and using the Repbase database as input and in this way I initially found 4.03% of the *P. buchholzi* genome could be classified as TEs, and that including low complexity DNA (0.11%) and simple repeats (2.22%) a combined 6.31% of the genome was classified as repetitive. Because this TE content was low for a vertebrate, and I reasoned that the most closely related species to *P. buchholzi* present in the Repbase database (zebrafish and other clupeocephalans) diverged ~300 Ma, and thus might make recognizing divergent TE sequences in *P. buchholzi* based on similarity alone difficult, I next set out to build a *P. buchholzi* specific repeat library based on *ab initio* repeat predictions. I ran the RepeatModeler pipeline (Smit et al. 1996) which leverages three *ab initio* repeat predictors (RECON, RepeatScout, and TRF) to predict repeats in a *de novo* genome without a species-specific repeat library, and then uses RepeatMasker and the Repbase database to attempt to classify these *de novo* repeat predictions. I ran RepeatModeler on the Pbuc_v57 assembly because it contained the most total sequence (472.4Mb) and I reasoned that the higher kmer-size would promote the resolution of more repetitive elements during the assembly process. By re-running RepeatMasker with the *P. buchholzi* specific library I saw a significant difference with the proportion of the Pbuc_v37 assembly classified as TEs jumping from 4.03% (Repbase database) to 6.72% (Pbuc_v57 database), while the low complexity (0.11%) and simple repeat (2.51%) content remained relatively constant resulting in a total repeat content of 9.33%. Finally, I merged the new *P. buchholzi*-specific repeat library

with the vertebrate Repbase repeat library to produce the most complete set of repeat family types I could produce as one library (Pbuc_v57_Repbase_MERGE) and ran RepeatMasker a third and final time with this combined library. Yet again, I noticed a significant increase in the proportion of the Pbuc_v37 assembly classified as TEs (10.99%) and almost no change in the proportion of low complexity (0.11%) or simple repeat (2.36%) elements, yielding a final repeat content of 13.46%. All of the results further described here are based on using RepeatMasker with the combined *P. buchholzi* specific + Repbase repeat libraries.

I compared the ratio of TE content to genome size between *P. buchholzi* and other vertebrates, including several teleosts. Both the genome size (445Mb) and the TE content (11%) are intermediate between the compact genomes of pufferfishes and the large, TE rich genome of the zebrafish (Figure 3.3). These observations in *P. buchholzi* do not change the prevailing view that TE content is exceptionally variable between teleost fish genomes, with as much as a ten-fold difference in TE content between the genomes of pufferfishes and zebrafish. Overall, I note that the genomes of actinopterygians, including gar, have a relatively higher proportion of TE content for a given genome size than sarcopterygians. There is also an apparent demarcation between the genomes of the ostariophysiians sampled (*D. rerio*, *C. idella*, and *A. mexicanus*) compared with the euteleosts sampled in that ostariophysiians appear to have a constantly higher genome size and TE content than euteleosts. Interestingly, I noted that the repeat content in the elopomorph *A. anguilla* is lower than expected from its genome size. There are, however, some important caveats to keep in mind when interpreting these results. In particular, the differences in the sequencing methods and

various levels of completeness of the genomes of the different species could have large effect on both the relative % of TE in the genome, and the diversity of TE families identified. Relatively complete genomes which have been sequenced with long read (Sanger) technology such as the zebrafish might be expected to resolve more individual TEs in the genome, which could be collapsed due to lack of information in the next-generation genomes using Illumina (e.g. *L. oculeatus*, *P. buchholzi, A. anguilla*) and 454 sequencing (*G. morhua*). In particular, the use of short insert size paired-end libraries which cannot span the full length of an insertion of a TE will fail to correctly assemble it and may leave it as a gap, or collapse it together with other similar TEs resulting in an underestimate of total TE quantity. Furthermore, in species where species-specific repeat libraries were used in their study (*D. rerio, X. maculatus, O. niloticus, G. aculeatus, T. nigroviridis, T. rubripes*) it is possible that both a greater diversity species-specific TE could be identified and this could contribute to overall higher estimation of % TE in the genomes. In support of this, the preliminary RepeatMasker runs on the *P. buchholzi* genome using only the collection of known vertebrate repeats present in Repbase resulted in only 4.03% of the genome being recognized as TE in origin, compared with the 10.99% found using the combined Repbase plus *P. buchholzi* specific *ab-initio* repeat predictions. Furthermore, entire families of TEs found with the *P. buchholzi* specific repeat libraries were not detected with the earlier Repbase-only run including the DNA transposon families EnSpm and MuDR, the LTR transposon family Copia/Ty1, and the LINE family Jockey. This highlights the importance of building species-specific *de novo* repeat libraries during the analysis of genomes from species without close evolutionary relatives already present in the Repbase database.

The draft genome of the eel *A. anguilla* has a contig N50 of only 1.6kb, and no species-specific *ab initio* repeat prediction was performed in the analysis of Chalopin et al. 2015. It is therefore probable that the % TE estimate from *A. anguilla* is an underestimate, and possible that entire families were missed. These same caveats, however, apply equally to other next-generation genomes, including *P. buchholzi*. With an N50 of 6.9kb, the genome is still highly fragmented and repeat regions are likely left unresolved and out of the assembly, and these results are therefore probably also an underestimate.

**Transposable element diversity in the *P. buchholzi* genome**

I next compared the overall genome coverage and relative proportions of the four main families of transposable elements (DNA, SINE, LINE, and LTR) between *P. buchholzi* and other vertebrates in order to see if I could detect any broad changes which could either be mapped to the teleost stem lineage, uniting osteoglossomorphs, elopomorphs, and clupeocephalans or which might be pertinent to outstanding questions of basal teleost phylogenetic relationships (Figure 3.4 A).

Overall, DNA transposons make up 3.46% of the *P. buchholzi* genome assembly, and in relative terms make up 31.48% of the total TE content making them the largest class in terms of genome coverage as well (Table 3.3, Figure 3.4 A). As with other teleost fish, I noted an overall high diversity of DNA transposons families in the genome of *P. buchholzi*. Using family classifications employed by RepeatMasker and considering only families which make up > 0.01% representation of TEs within their class, DNA transposons are still the most diverse class of elements in the *P. buchholzi* genome, with 22 families represented (Figure 3.4 B). In comparison, LINEs are the

second most diverse class with 10 families represented, followed by SINEs and LTRs with 8 families represented each (Table 3.3, Figure 3.4 A,B). Using the more inclusive classification of Chalopin et al. 2015, which collapses several RepeatMasker families into one (e.g. all Tc-Mariner type elements, and all hAT type elements are considered together as one family each), it is still evident that *P. buchholzi* has a high diversity of DNA type elements and in terms of strict presence/absence of families is tied with zebrafish for being the teleost with the greatest diversity of DNA transposons described with 18/20 types present in at least one copy in the genome. Considered together, all Tc1/Mariner family elements constitute the third largest family of TEs in the *P. buchholzi* genome.

I investigated whether the presence or absence of any DNA class transposons bore any phylogenetically informative signals which could help resolve the question of basal teleost interrelationships. The only DNA transposon families absent from the *de novo* assembly of the *P. buchholzi* genome produced here are Merlin and Zisupton elements. Merlin elements are ~2kb long DNA transposons which often encode endogenous transposases and are distributed sporadically across the tree of life including in several bacteria and eukaryotes (Feschotte 2004). Amongst chordates, Merlin elements are present in *C. intestinalis*, a limited range of vertebrates including human, *X. tropicalis*, *A. caroliensis*, and scattered representation of teleosts (zebrafish, cod, platyfish, stickleback) in varying copy numbers from a few dozen to a few thousand (Chalopin et al. 2015b). They have not been detected in the genome of the gar, eel, medaka, cichlids, or pufferfish and therefore their absence from *P. buchholzi* does not contain any useful phylogenetic data which might help resolve the relative placement of

Osteoglossomorpha and Elopomorpha. Rather, it seems this is an ancient TE family which has been lost multiple times in evolution. In contrast, Zisupton DNA transposons are large, low copy number elements, with only six copies identified in the platyfish, *X. maculatus,* where they were first identified (Böhne et al. 2012). The presence or absence of Zisupton element bears a strong phylogenetic signal as a rare genomic change (Rokas & Holland 2000) and has been detected in the genomes of all clupeocephalan teleosts examined, albeit in very low copy number (e.g. only one copy in *Tetraodon*). Teleost fish are the only vertebrates with the Zisupton element and it is therefore interesting to speculate whether this was an acquisition only in this lineage and whether it is associated somehow specifically with the TGD. Importantly, Zisupton elements were not detected in the gar, nor were they detected in *A. anguilla* or in *P. buchholzi*. This suggests that if Zisupton is a novel acquisition, it was only acquired in the clupeocephalan lineage, after the Osteoglossomorpha and Elopomorpha branched off and was not acquired simultaneously with the TGD. Unfortunately, while the absence of Zisupton from the gar, *P. buchholzi* and *A. anguilla* is consistent with the early branching of the osteoglossomorph and elopomorph lineages, it does not help answer the question of whether osteoglossomorphs, elopomorphs, or a group containing both, are the sister clade to clupeocephalan teleosts. However, because Zisupton are such low copy elements it is possible that they might simply have been missed in the assemblies of *P. buchholzi* and *A. anguilla*. I conducted low stringency (E=10) blastn searches of each of the *P. buchholzi* genome assembly versions produced with multiple kmers in case Pbuc_v37 was an exception, but did not find any significant hits, confirming the RepeatMasker analysis. It will therefore be important to

produce more genomes with higher contiguity in a broad taxonomic sampling from various other osteoglossomorphs and elopomorphs to determine if the Zisupton element can be found. I have already taken the first steps towards this goal by producing *de novo* draft assemblies of 3 additional Osteoglossomorphs as described in Chapter 5 and screening them with low stringency blastn searches in the same way. Preliminary results indicate that Zisupton is not present however this must be interpreted cautioiusly as these *de novo* assemblies were produced with low coverage and the assemblies were less contiguous than even the *P. buchholzi* assembly produced here.

The second most abundant class of TEs in the *P. buchholzi* genome are the SINE elements, representing 2.31% of the total genomic sequence and 20.02% of the total TE content respectively (Table 3.3, Figure 3.4 A). *P. buchholzi* therefore have a higher proportion of SINE elements in their genomes than any other teleost examined (with zebrafish in second place with 1.8% genome-wide SINE content). SINES are non-autonomous retrotransposons between 100-600bp in length and can be found in the genomes of most eukaryotes in copy numbers ranging from tens to hundreds of thousands (Kramerov & Vassetzky 2011). The most abundant family in the SINE class found in *P. buchholzi* are the SINE-V type elements which constitute 76.62% of all annotated SINE class elements, and are also the single most widespread TE family in the *P. buchholzi* genome across all classes with 70,172 annotated insertions covering 1.77% of the *P. buchholzi* genome with an average length of 107bp.

LTR (Long Terminal Repeat) type elements make up 2.06% of the *P. buchholzi* genome and 18.74% of the total TE diversity (Table 3.3, Figure 3.4 A). Chalopin et al. 2015 described 6 LTR type families and compared their presence and absence across

vertebrates. All actinopterygians including gar and eel were found to possess all 6 families. I also found the *P. buchholzi* genome possessed all of the same families, and therefore the presence/absence of LTR type TEs are not phylogenetically informative. Notable in *P. buchholzi* are the Ngaro family of LTR elements which are the second most abundant family of TEs after SINE-V elements and together comprise 1.75% of the *P. buchholzi* genome from 67,657 insertions with an average length of 110bp. Ngaro family LTR elements were also found to be present in high copy number in the gar genome (Braasch et al. 2016).

Finally, LINEs (Long Interspersed Nuclear Elements) constitute only 1.33% of the *P. buchholzi* genome and 12.10% of the total TE diversity and at this level are the least abundant class of TEs (Table 3.3, Figure 3.4A). Interestingly, however two families show a retention profile in actinopterygians which might hold relevant phylogenetic information. These are the CR1/LINE3 family which have been described in the genomes of all sarcopterygians, and have recently been found in the genomes of the gar (Braasch et al. 2016) and the eel (Chalopin et al. 2015b) but are apparently absent from the genomes of clupeocephalan teleosts and the Nimb family which appear to be clupeocephalan specific and is absent from the genomes of gar and eel (Chalopin et al. 2015b; Braasch et al. 2016). I found that CR1/LINE3 type elements were abundantly present in the *P. buchholzi* genome, as they are in the genomes of the gar and the eel. The presence of this element is therefore not phylogenetically informative. Likewise, I found no evidence for the Nimb family in *P. buchholzi*, supporting the idea that, like the DNA element Zisupton, Nimb is a clupeocephalan-specific family.

Overall, the results of this survey of TE diversity and content in the *P. buchholzi* genome was not able to resolve the question of whether osteoglossomorphs, elopomorphs, or a clade containing both are the sister lineage to clupeocephalan teleosts. Interestingly, with the addition of the first analysis of TEs in an osteoglossomorph genome more evidence has accumulated which suggests that clupeocephalans genomes are significantly different from other teleosts at least in terms of their TE family composition.

**Transposable element content in the *P. buchholzi* Hox clusters**

In Chapter 2, I characterized the Hox clusters of *P. buchholzi*, noting that their reduction in cluster number (from 8 to 5 following the TGD) was compensated for by an overall higher and largely symmetrical gene and ancient CNE retention rate between the duplicated HoxB clusters such that the total number of *hox* genes (45) is similar to clupeocephalan teleosts. However, the TE profiles of the *P. buchholzi* Hox clusters were not examined and it is currently unknown whether they show a more permissive invasion profile similar to zebrafish and other teleosts with multiple DNA transposon invasions, or more restrictive invasion profile similar to tetrapods and the bichir *P. senegalus* but with a broader diversity of transposable elements types. In order to answer these questions, I performed a separate repeat analysis on the previously sequenced *P. buchholzi* Hox clusters.

Scaffolds spanning the full length of each of the five *P. buchholzi* Hox clusters (hoxax, hoxbx, hoxby, hoxcx, and hoxdx) as defined by the first and last coding exons of the most 5' and 3' *hox* genes themselves, and including 884kb of flanking non-Hox

cluster sequences immediately 5' and 3' from the Hox clusters, were examined and the relative proportions of TEs inside the Hox clusters, and outside the Hox clusters were compared (Table 3.4). Strikingly, within the *P. buchholzi* Hox clusters there was an average TE content of only 0.86%, compared with an average flanking sequence TE content of 11.06%. This suggests that the flanking region surrounding the Hox clusters is almost exactly in line with the genome wide average (10.99%), while there is almost a complete exclusion of TEs from the Hox clusters. Interestingly, when examined on a cluster-by-cluster basis it is evident that there are significant differences in the TE content between different Hox clusters (Table 3.4, Figure 3.5). When broken down in this way it is evident that the majority (75% of all TEs present within the Hox clusters) are found within the hoxby cluster which has a total TE content of 3.06%, and a flanking sequence repeat content of 15.89%. Each of the other four clusters has a TE content < 1% and in some clusters (hoxbx, hoxcx, hoxdx) I found remnants of only a single TE insertion. The 3.06% repeat content of the *P. buchholzi* hoxby cluster is similar to estimates of TE content in the zebrafish Hox clusters. However, whereas most repeats which have invaded the zebrafish Hox clusters are DNA transposons, I found that in the *P. buchholzi* hoxby cluster a higher diversity of elements have invaded the cluster (Figure 3.5). It is also interesting to note that even within the largest cluster hoxcx there is only a single LINE element insertion and that therefore TE insertion is not strongly correlated with Hox cluster size in *P. buchholzi*, which seems to be the case in the expanded Hox clusters of squamates and newts (Di-Poï et al. 2009; Voss et al. 2013) as well as the expanded posterior HoxB cluster of humans and mouse (Liang et al. 2011).

These repeat structure and content analyses suggest that overall overwhelmingly strong selective constraints are operating on the *P. buchholzi* Hox clusters in an order of magnitude similar to those observed in the majority of tetrapods, and greater than observed in other teleosts. With a similar overall %TE in the Hox-flanking regions (11.06%) to the genome-wide predictions (10.99%), this independent data produced using a different sequencing technology (BACs sequenced by 454 producing a highly contiguous scaffolded sequence) also lends confidence to the genome-wide TE estimates.

**Genome-wide protein coding gene predictions**

As a first step in any large scale comparative genome analysis of protein-coding gene content and in order to gain a genome-wide view of the protein-coding gene representation it is necessary to first define a putative set of gene models. Using the *de novo* assembly of the *P. buchholzi* genome Pbuc_v37, I carried out a combined evidence-based and *ab initio* annotation of the full protein-coding complement using the MAKER pipeline (Cantarel et al. 2008). I used the combined Repbase+Pbuc_v57 custom library I built for TE analysis to mask the repetitive portions of the Pbuc_v37 assembly, and then performed two iterative runs of MAKER to predict protein coding gene models from the *P. buchholzi* genome. MAKER combines the use of both *ab initio* gene prediction algorithms as well as supplied protein and EST evidence in order to annotate novel genomes. It is a tool now widely in use in the *de novo* genome sequencing and annotation community and has been used successfully to predict gene models in vertebrates across a broad taxonomic range even when closely related species are not available including the recently sequenced genomes of the gar

(Braasch et al. 2016), the northern pike (Rondeau et al. 2014), and the sea lamprey (Smith et al. 2013b).

I performed an initial training of the *ab initio* prediction algorithms using the set of 167 full-length CEGs predicted previously predicted using CEGMA, as well as the 45 manually curated set of *P. buchholzi* Hox genes. As protein-coding evidence I supplied the entire vertebrate protein repository in Refseq (Pruitt et al. 2014) as well as the recently predicted proteins from the gar genome assembly version LepOcu1 (Braasch et al. 2016). I also supplied as EST evidence a *de novo* transcriptome assembly produced with the brain and muscle RNAseq data from *P. buchholzi* and assembled with the program Trinity (k=25) (Schulz et al. 2012), and further screened for homologs of known vertebrate putative protein-coding genes by blastx against the UniProtKB-SwissProt protein database ($E<10^{-6}$) which resulted in a high-confidence putative protein coding set of 23,636 transcripts with a mean length of 2.3kb and N50 of 3.2kb (Figure 3.6). The predicted gene models from the first round of MAKER, excluding those without supporting protein or EST evidence, were then used to re-train *ab initio* predictors SNAP and Augustus and a second round of MAKER was run.

The final annotation set included 41,236 MAKER models with protein/EST alignment evidence, and 40,701 SNAP/Augustus models which lacked protein/EST alignment evidence. Out of the 40,791 *ab initio* only models which lacked protein/EST alignment evidence I found that 5,971 produced a significant hit ($E=10^{-6}$) when blastx against the vertebrate subset of the UniProtKB-SwissProt protein database was carried out. Out of the MAKER models 35,978 models produced a significant hit ($E=10^{-6}$) against the vertebrate UniProtKB-SwissProt protein database. All resultant gene

models (MAKER evidence models and *ab initio* only models) were further subjected to a search for putative protein domains using InterproScan (Jones et al. 2014). I found that 5,629 *ab initio* only models could be annotated with at least one protein domain using InterproScan, while 35,497 MAKER models were annotated with InterproScan. In total, I found 6,658 gene models from the *ab initio* only set which produced either a significant vertebrate UniProtKB-SwissProt blastx hit or possessed an InterproScan protein-domain and therefore I merged these together with the 41,236 evidence-supported MAKER models to produce a final putative gene model set of 47,894. The translated models were subsequently annotated roughly with blastp ($E=10^{-6}$) against the vertebrate UniProtKB-SwissProt and assigned putative gene names corresponding to their top hits. A total of 41,607 of the models were assigned a gene name in this way.

I next sought to assess how representative of a typical vertebrate proteome the predicted *P. buchholzi* gene models were by conducting reciprocal blastp searches against four reference vertebrate proteomes. I used ENSEMBL predicted proteomes from Human (GRCh37), gar (LepOcu1), zebrafish (Zv9) and *T. nigroviridis* (TETRAODON8.0) alternately as query and input sequences against the 47,894 translated putative MAKER gene models. I determined the number of proteins which found significant ($E=10^{-6}$) hits, the percentage identity of those hits, and the length of the alignments in order to appreciate the overall proteome representation, mean similarity, and completeness of the gene models compared with the four chosen reference genomes. Results of these analyses are presented in (Figure 3.7). In human I found that 89% of the reference proteome aligned to the *P. buchholzi* MAKER models with a mean percent identity of 64.5% over an average of 35% of their length. In gar, I found

96% of the reference proteome could be aligned to the MAKER models with a mean identity of 70.9% over a mean of 35% of their length. Finally in the two teleost reference proteomes I found that in zebrafish 96% of the reference proteome aligned with a mean identity of 70.3% over a mean of 37% of their length, while in *T. nigroviridis* 97% of the proteome aligned to the *P. buchholzi* models with a mean identity of 67.8% over a mean of 37% of the reference proteome length. Of the MAKER models, I found between 86-89% of the models found significant hits in each of the reference proteomes. Overall, this suggests a high representation of the reference proteomes in the MAKER model gene set. However, with a mean alignment length of only of 35-37% against the reference proteomes, this suggests that the models are highly fragmented. In each case only ~1000 MAKER models could be aligned against 100% of the length of the reference protein. This is consistent with a genome assembly covering the majority of protein coding exons, but where exons are often separated on multiple contigs the result is separate gene models. The majority of the resultant 47,894 predicted gene models are therefore expected to be partial models only and that the real gene number in *P. buchholzi* is significantly lower.

## Discussion

Overall, despite a notably lower overall genome coverage (~30X) and library type diversity (a single 200bp paired-end library) than targeted in recent high-throughput next generation genomes of other vertebrates, the *de novo* assembly of the *P. buchholzi* genome presented here was of sufficient contiguity to yield results consistent with expectations in terms of CNV content, TE representation, and protein coding gene content. This suggests that preliminary low-budget sequencing and *de novo* assembly approaches in non-model organisms led by one individual can be a fruitful endeavour. If financial and informatics resources were not limited there would be no reason to expect that the methods employed here but with higher coverage (upwards of 100-200X), and multiple insert size libraries (200bp, 500bp, and 3kb) could not result in a much greater contiguity genome if assembled with programs which can readily handle multiple insert-size libraries such as ALLPATHS-LG (Gnerre et al. 2010) as done in the case of the gar genome (Braasch et al. 2016). While the 47,894 predicted protein coding gene models is clearly an overrepresentation based on fragmentation of single genes across multiple contigs, they nevertheless represent a significant resource for future analyses of gene family evolution such as those carried out in Chapter 4. Likewise, although far from being complete, the assembly presented here revealed several interesting features especially pertaining to TE evolution.

From my analyses of TE diversity in *P. buchholzi* and other chordates it appears that Zisupton and Nimb family transposons were acquired specifically in the clupeocephalan lineage, while the CR1/LINE3 family was lost. It would be interesting to examine whether there is any evidence that the gain and losses of these families has

had any functional consequences specific to the clupeocephalan lineage, for example through exaptation, which could have facilitated the more widespread diversification of clupeocephalans compared with osteoglossomorphs and elopomorphs.

While the Zisupton element was not recovered from this *de novo* assembly of the *P. buchholzi* genome, nor any of the other low-coverage osteoglossomorph assemblies produced in Chapter 5 of this thesis, future higher-contiguity assemblies may reveal its absence was an artefact. If Zisupton is eventually found in either one of the osteoglossomorpha or elopomorpha but not in the other it could suggest that the lineage in which it is found is more closely related to clupeocephalans and that the Zisupton insertion occurred in the common ancestor of elopomorphs and clupeocephalans (Elopocephala), or in the common ancestor of osteoglossomorphs and clupeocephalans (Osteoglossocephala).

It is interesting to speculate whether the expansion of the SINE-V family of transposable elements revealed here played a functional role in the *P. buchholzi* genome through exaptation, as has been suggested in other lineages, although this requires explicit testing and functional experimentation to validate. It is notable that the genome of the gar has a nearly equivalent proportion of SINE elements in its genome as *P. buchholzi* (2.73%). Although less total copies of the SINE-V family were detected in gar (37,444 insertions covering 0.54% of the genome) it was noted that it is still one of the most abundant TE families in the gar, and several hundred transcripts were found in the gar which contained SINE-V like elements (Braasch et al. 2016).

In general it seems that the *P. buchholzi* 'mobilome' is more similar in many respects to the pre-TGD gar genome, and to some sarcopterygian vertebrates than it is to most clupeocephalan teleosts. The TE profile of *P. buchholzi* is fairly evenly split between DNA, SINE, LINE and LTR classes. This contrasts heavily with most teleosts such as the closely related zebrafish and cavefish where DNA transposons make up the significant majority of the mobilome, or pufferfishes where LINEs make up a larger proportion of the mobilome than other elements. Pufferfishes, however, have the smallest known genomes and lowest fraction of TEs in their genomes known amongst vertebrates and as such have likely experienced very different selective pressures on TE spread than other species (Jaillon et al. 2004; Aparicio et al. 2002). The stickleback has the most similar overall proportional representation of TE classes amongst clupeocephalans to *P. buchholzi* although it has a notably smaller proportion of SINEs and an increased proportion of LINEs, like the pufferfishes. This can be seen clearly in Figure 3.4 A where the blue bars represent the fraction of DNA transposons present in *P. buchholzi* Interestingly, the eel has a more clupeocephalan-like profile of total DNA transposon content and has significantly more DNA transposons than other transposon classes (four fold difference). Notably, however, almost half of all TEs identified in the eel genome are still classified as 'Unknown', suggesting there may be novel highly derived TEs present which do not belong to any of the four major classes of TE. Only if the eel genome contiguity was significantly improved, and examined with a species-specific repeat element library could this observation be verified. Interestingly, both the eel and *P. buchholzi* show a presence/absence profile of specific TE families which aligns them more with the pre-TGD gar genome than with the post-TGD genomes of

other clupeocephalans. The shared presence of CR1/LINE3, and the absences of Nimb and Zisupton elements from *P. buchholzi*, *A. anguilla*, and gar unite these three genomes and place the loss of CR1/LINE3 and the acquisition of Nimb and Zisupton specifically on the clupeocephalan branch of the teleost tree.

Overall, these analyses do not suggest that the TGD is directly correlated with any significant shifts in TE profiles in the common ancestor of all teleosts, nor is there any evidence for any gains and losses of TE families prior to the divergence of the three basal teleost subdivisions. Lineage specific expansions of TEs and the gain and loss of TE families appears to have occurred independently in different teleost lineages, suggesting that if there was any contributory role of the TGD its effect was delayed.

The only Hox cluster in *P. buchholzi* which showed a significant invasion of TEs is the duplicated hoxby cluster. It might be the case that the loss of 3 duplicated Hox clusters early in the lineage leading to *P. buchholzi* following the TGD greatly constrained their evolution, to a degree not seen in other teleosts which retained more duplicated Hox clusters longer. Importantly, this suggests that if the TGD did alleviate functional constraints and permit invasion of TEs into the duplicated Hox clusters of teleosts it likely occurred separately in *P. buchholzi* and clupeocephalans after their divergence. It will be interesting in the future to examine each of the TE insertions in the hoxby and hoxbx clusters in detail and compare them to other teleosts to see if there are any phylogenetically informative rare genomic changes (Rokas & Holland 2000) of insertion events into either the hoxbx or hoxby cluster which would unambiguously assign orthology to either the hoxba, or hoxbb clusters of other teleosts, or whether the hoxb clusters of teleosts were still segregating as tetraploid alleles prior

to the divergence of osteoglossomorphs and clupeocephalans, as I have previously suggested (Chapter 2, Martin & Holland 2014). According to my previous work, I would not expect to find any such changes but instead that hoxbx and hoxby of osteoglossomorphs have accumulated a largely independent set of TE insertions from hoxba and hoxbb in clupeocephalans.

## Conclusions

I have produced a draft genome assembly of the osteoglossomorph *P. buchholzi* and demonstrated its utility in the comparative evolutionary analysis of both transposable elements and protein-coding genes. I have shown the transposable element profile of *P. buchholzi* exhibits a mixture of both teleost and non-teleost features and that there is no evidence for a teleost-wide shift in TE composition immediately following the TGD. This draft genome assembly will facilitate future genome-wide analyses of gene family evolution in the context of comparative genomics following the TGD between osteoglossomorphs and clupeocephalans.

**Table 3.1. *P. buchholzi* genome assembly metrics.**

Assemblies were carried out using the velvet algorithm with a range of different kmers (33, 37, 47, 57) and assembly metrics were compared between each. CEG (Core Eukaryotic Gene) representation as defined by the CEGMA pipeline were also compared between each assembly. I selected the k=37 assembly for subsequent analysis as it yielded a good balance between CEG representation and overall contiguity.

| kmer | k=33 | k=37 | k=47 | k=57 |
|---|---|---|---|---|
| **Contig N50 (bp)** | 6,512 | 6,930 | 6,495 | 5,912 |
| **Contig N90 (bp)** | 1,306 | 1,370 | 1,365 | 1,175 |
| **Number of Contigs** | 140,706 | 136,818 | 143,915 | 215,510 |
| **Longest Contig (kb)** | 501 | 410 | 185 | 132 |
| **Total Assembly (Mb)** | 438.59 | 445.40 | 453.68 | 472.40 |
| **%GC** | 41.64 | 41.72 | 41.77 | 41.88 |
| **Full CEGs (248)** | 165 (66%) | 167 (67%) | 166 (67%) | 168 (68%) |
| **Partial CEGs (248)** | 219 (88%) | 225 (91%) | 225 (91%) | 226 (91%) |

**Table 3.2. *P. buchholzi* genome assembly Pbuc_v37 read mapping statistics.**

Genomic and RNAseq reads were re-mapped to the Pbuc_v37 assembly using the software bowtie2/tophat2. Overall mapping rate considers all alignments found while concordant PE mapping rate considers only those alignments where both ends of the paired end read map adjacent to each other in the correct orientation with a gap no longer than 500bp between them.

| Library | Total Reads Used (>Q20) | Overall Mapping Rate | Total PE Reads Used (Q>20) | Concordant PE Mapping Rate |
|---|---|---|---|---|
| Genome DNA | 164,654,082 | 83.12% | 80,490,798 | 76.67% |
| Brain RNA | 29,035,415 | 75.30% | 14,245,031 | 66.20% |
| Muscle RNA | 19,290,264 | 72.10% | 9,472,207 | 63.20% |

**Table 3.3. Transposable element (TE) content in the *P. buchholzi* genome.**

Using the software RepeatMasker and the custom repeat library Pbuc_v57+Repbase2014 the total number of insertions, the number of bases in the genome they covered, and the total % of the genome covered by each of the four major classes of transposable element and those which could not be classified was calculated. The proportional representation of each TE class within the total TE composition of the genome was also calculated.

| Type | Count | bp masked | Genome % | TE proportion |
|---|---|---|---|---|
| **DNA** | 149,336 | 20,244,222 | 3.46% | 31.48% |
| **SINE** | 95,226 | 9,912,716 | 2.31% | 21.02% |
| **LTR** | 79,894 | 8,798,461 | 2.06% | 18.74% |
| **LINE** | 43,021 | 5,722,479 | 1.33% | 12.10% |
| **Unknown** | 53,426 | 7,796,838 | 1.83% | 16.65% |
| **Total** | 420,903 | 52,474,716 | 10.99% | 100.00% |

**Table 3.4. Transposable element content in the Hox clusters of *P. buchholzi*.**

Using the previously assembled Hox clusters of *P. buchholzi* an independent run of RepeatMasker using the custom Pbuc_v57+Repbase2014 library was run using each hox-cluster containing scaffold. The total number of base pairs and % coverage of the genomic sequence occupied by annotated transposable elements both within the Hox clusters (Hox TE) and outside the Hox clusters (Flanking TE) was compared. While the overall TE % in the flanking sequence adjacent to each hox cluster was similar to the genome wide average, TEs are highly excluded from within the hox clusters.

| Cluster | Hox TE (bp) | Hox Total (bp) | Hox % TE | Flanking TE (bp) | Flanking Total (bp) | Flanking % TE |
|---|---|---|---|---|---|---|
| **hoxax** | 298 | 55,231 | 0.54% | 10,929 | 139,739 | 7.82% |
| **hoxbx** | 48 | 66,149 | 0.07% | 26,539 | 213,882 | 12.41% |
| **hoxby** | 2,136 | 69,795 | 3.06% | 30,977 | 194,910 | 15.89% |
| **hoxcx** | 310 | 93,161 | 0.33% | 10,272 | 149,216 | 6.88% |
| **hoxdx** | 64 | 45,374 | 0.14% | 19,055 | 186,158 | 10.24% |
| **total** | 2,856 | 329,710 | 0.86% | 97,772 | 883,905 | 11.06% |

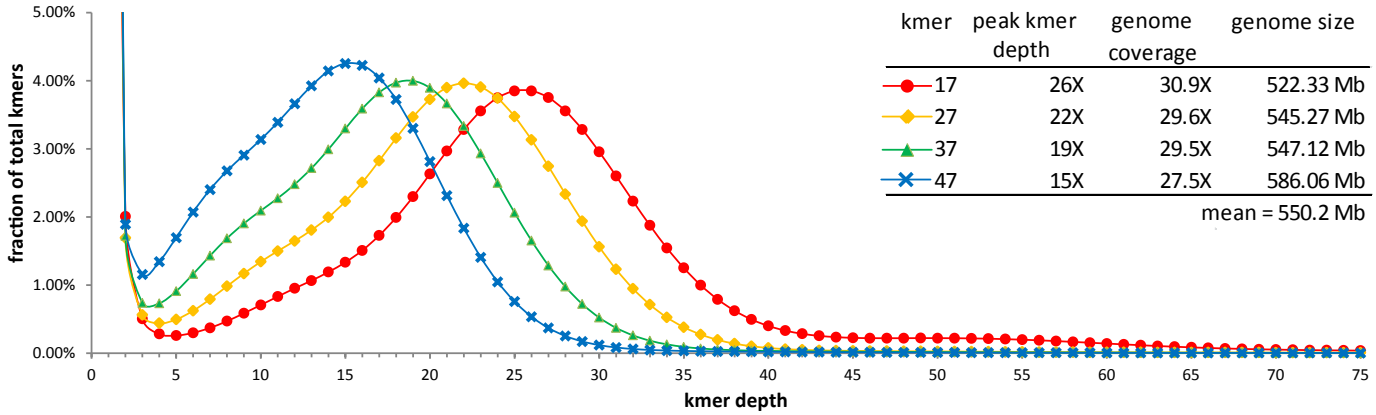



**Figure 3.1.** kmer depth plot and genome size estimation.
Using a range of kmers between 17-47 I estimated the peak kmer depth from the distribution and calculated the estimated genome coverage of the sequencing data as well as the genome size of *P. buchholzi*.

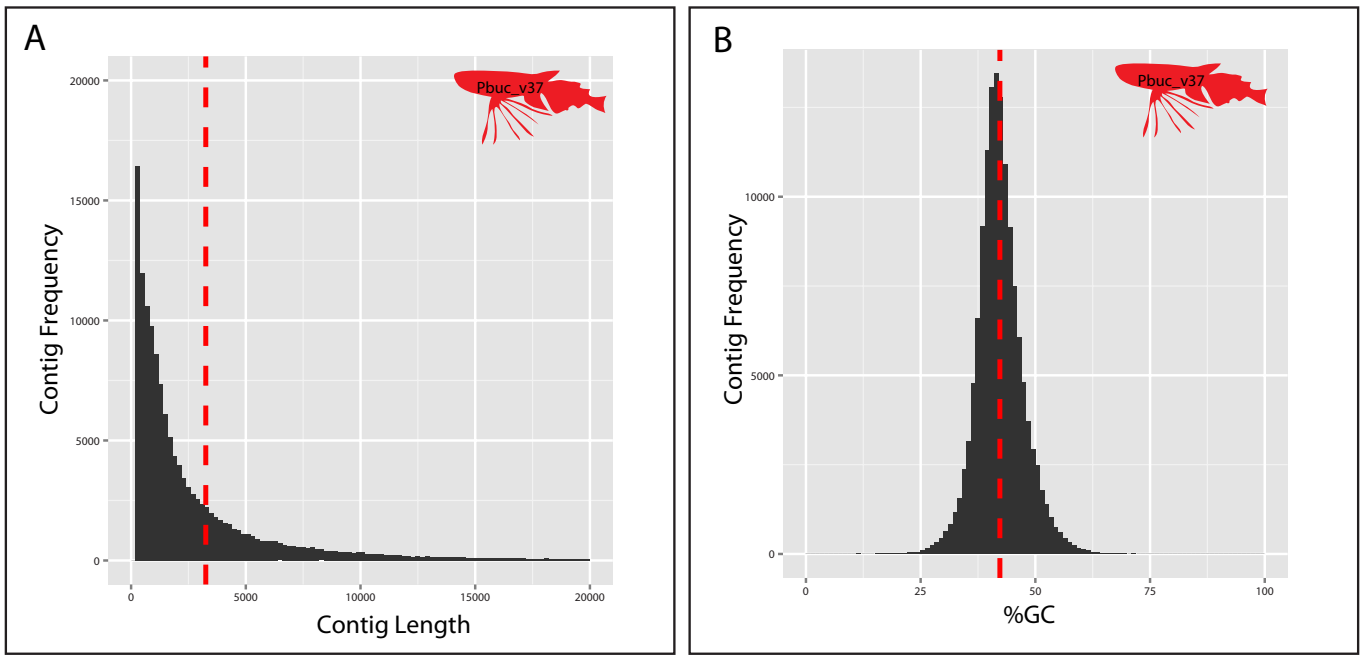



**Figure 3.2.** *P. buchholzi* genome assembly Pbuc_v37 contig length and GC content. (**A**) Distribution of contig lengths in the Pbuc_v37 assembly. Dotted red line denotes the mean contig size of 3.3kb. (**B**) Distribution of average contig %GC content. Dotted line denotes the mean %GC of 42%.

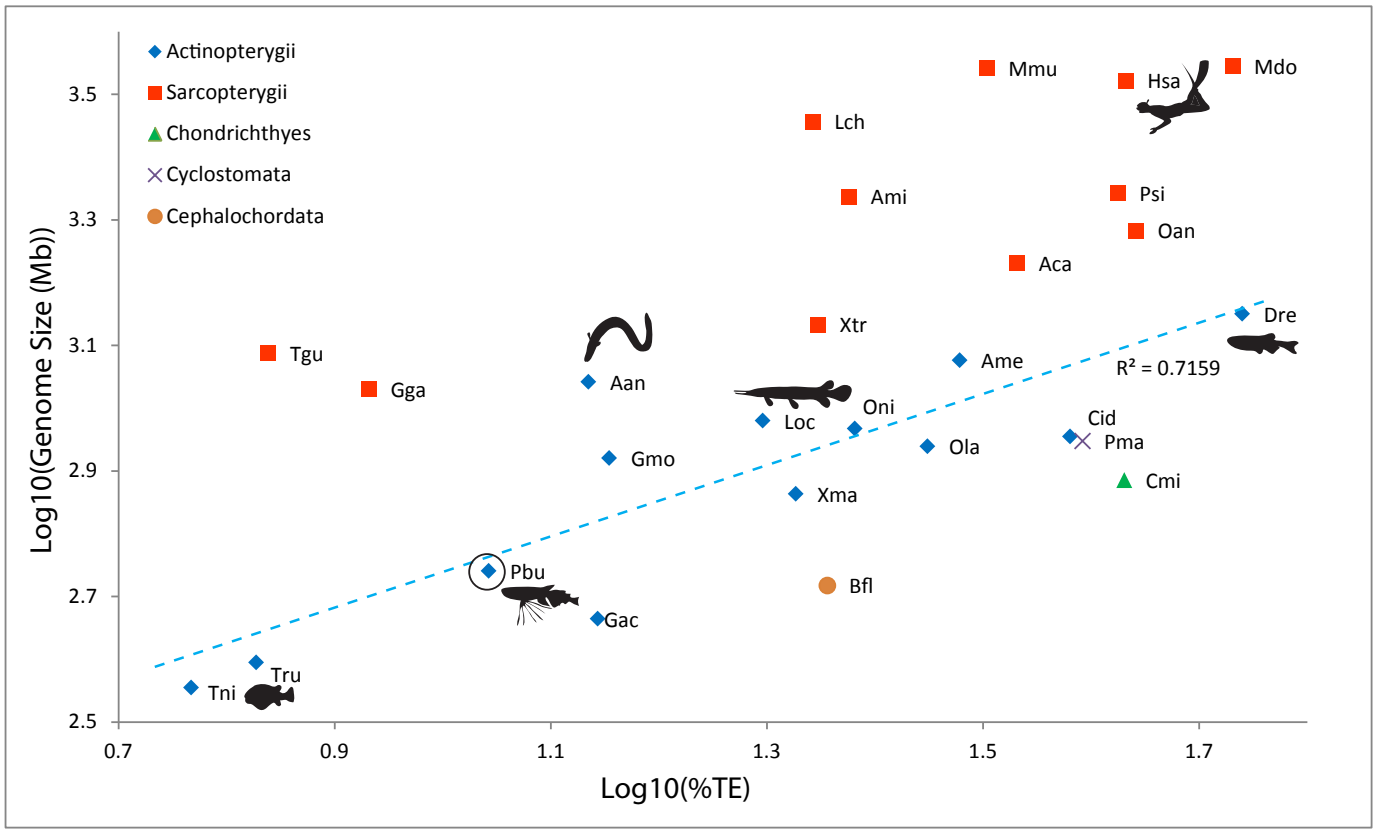



**Figure 3.3.** Proportion of transposable elements in the genome of *P. buchholzi* and other chordates. Transposable element content is highly correlated with genome size in Actinopterygii (blue diamonds). Both the genome size and TE content of *P. buchholzi* (Pbu) is intermediate between the extremely compact genomes of pufferfishes (Tni, Tru) and the large genomes of zebrafishes (Dre) and other ostariophysiians. Ostariophysiians appear to have larger genomes and higher TE content than other teleosts. Actinopterygians in general appear to have a greater proportion of TEs in their genomes compared with sarcopterygians. Species name abbreviations are as follows: Actinopterygii: T*etraodon nigroviridis* (Tni), *Takifugu rubripes* (Tru), *Pantodon buccholzi* (Pbu), *Gasterosteus aculeatus* (Gac), *Gadus morhua* (Gmo), *Anguilla anguilla* (Aan), *Lepisosteus oculeatus* (Loc), *Xiphophorus maculatus* (Xmac), O*reochromis niloticus* (Oni), *Oryzias latipes* (Ola), *Astyanax mexicanus* (Ame), *Ctenopharyngodon idella* (Cid), *Danio rerio* (Dre). Sarcopterygii: *Taenopygia guttata* (Tgu), *Gallus gallus* (Gga), *Xenopus tropicalis* (Xtr), *Alligator missisipiensis* (Ami), *Anolis caroliensis* (Aca), *Mus musculus* (Mmu), *Homo sapiens* (Hsa), *Pelodiscus sinensis* (Psi), *Ornithorhynchus anatinus* (Oan), *Monodelphis domestica* (Mdo). Cyclostomata: *Petromyzon marinus*. Chondrichthyes: *Callorhinchus milli* (Cmi). Cephalochordata: *Branchiostoma floridae* (Bfl).

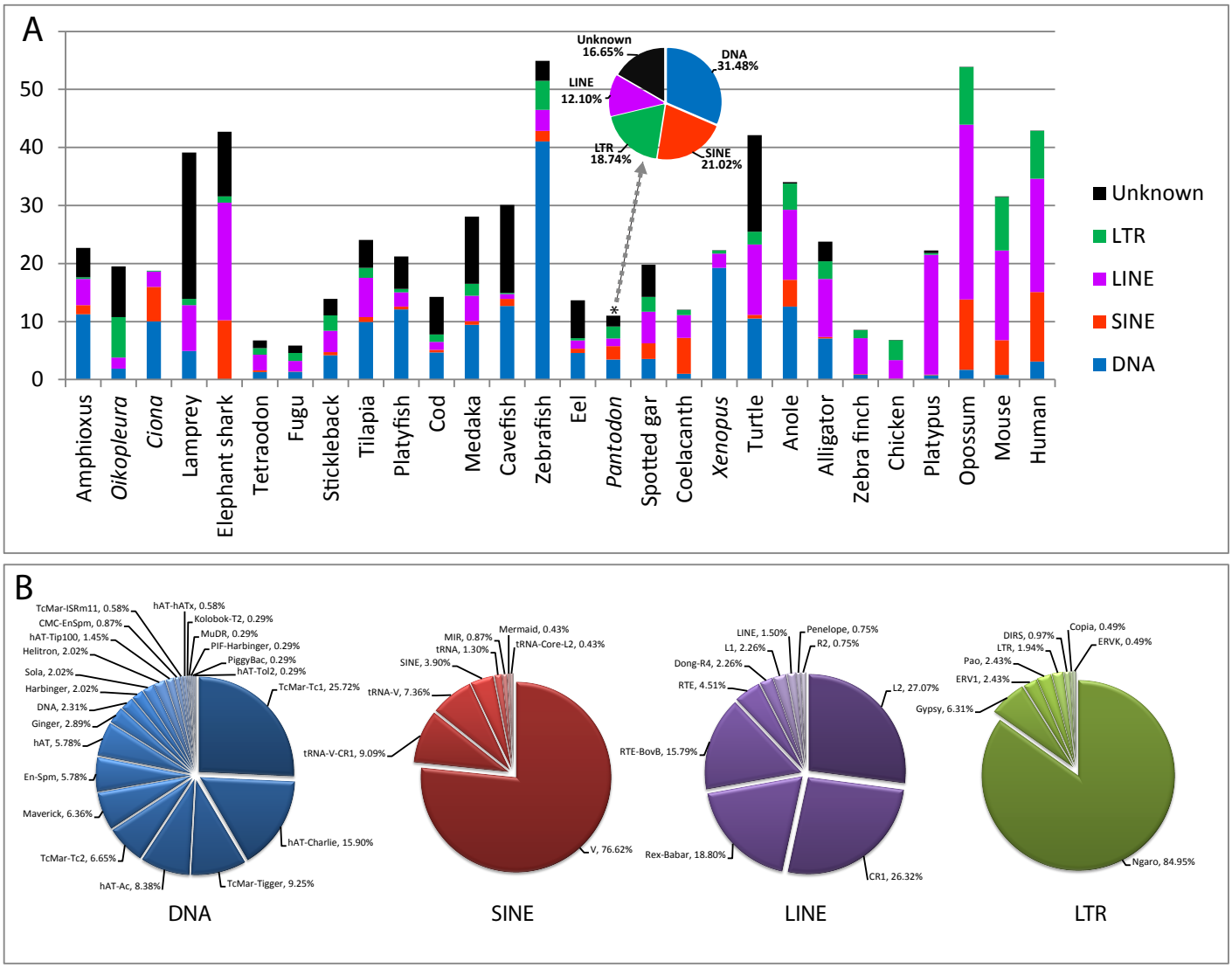



**Figure 3.4.** Diversity of transposable elements in the genome of *P. buchholzi* and other chordates. **(A)** Absolute and relative proportions of four major classes of TEs in the genome of *P. buchholzi* and other chordates. The total height of the bar represents the total % of TEs in the genome, while the relative proportion of the total TE content for each of the DNA (blue), SINE (red), LINE (purple) and LTR (green), as well as elements which could not be classified by RepeatMasker/Repbase are represented within each vertical bar. The relative proportion of TEs in *P. buchholzi* is expanded as a separate pie chart. **(B)** Diversity and proportional representation of each RepeatMasker family of TEs within the DNA (blue), SINE (red), LINE (purple) and LTR (green) classes in *P. buchholzi*.

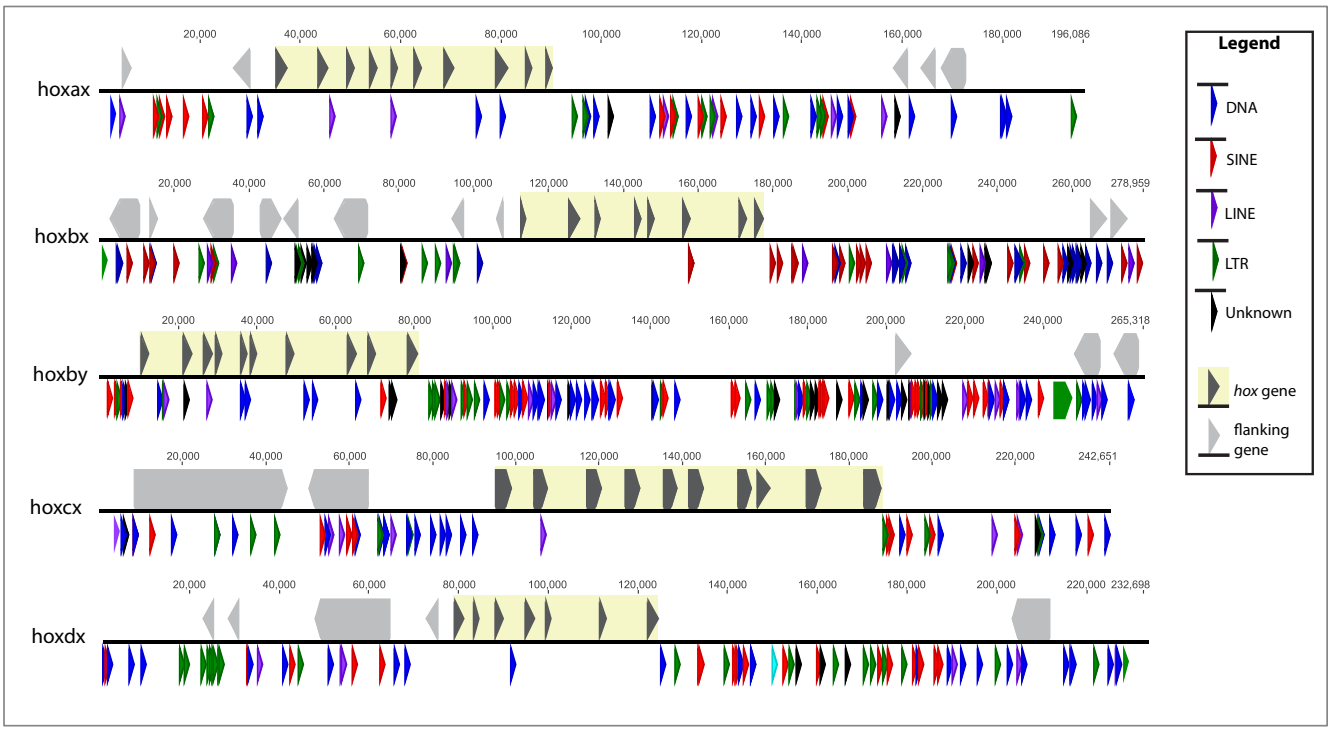



**Figure 3.5.** Transposable elements are excluded from the *P. buchholzi* Hox clusters. Transposon insertions are represented on the scale map of the *P. buchholzi* Hox clusters. Insertions of different families of transposable elements are colour coded according to the legend. Overall, transposable elements are almost completely excluded from within *P. buchholzi* Hox clusters (0.86% average), compared with the flanking region (11.06% average), suggesting strong negative selection.

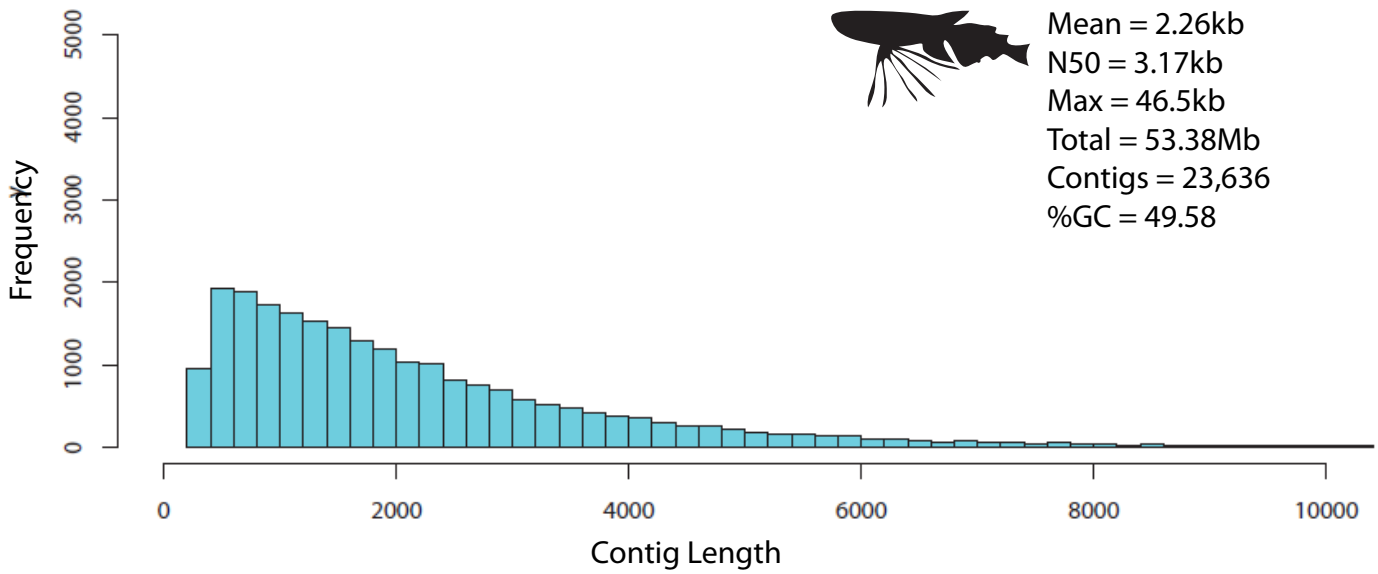



**Figure 3.6.** Transcriptome assembly of *P. buchholzi* brain and muscle total RNA.
The transcriptome was assembled with Trinity, the final putative protein-coding EST set of 23,636 transcripts with homology to the UniProtKB-SwissProt protein database was used as EST evidence in the final MAKER annotation set.

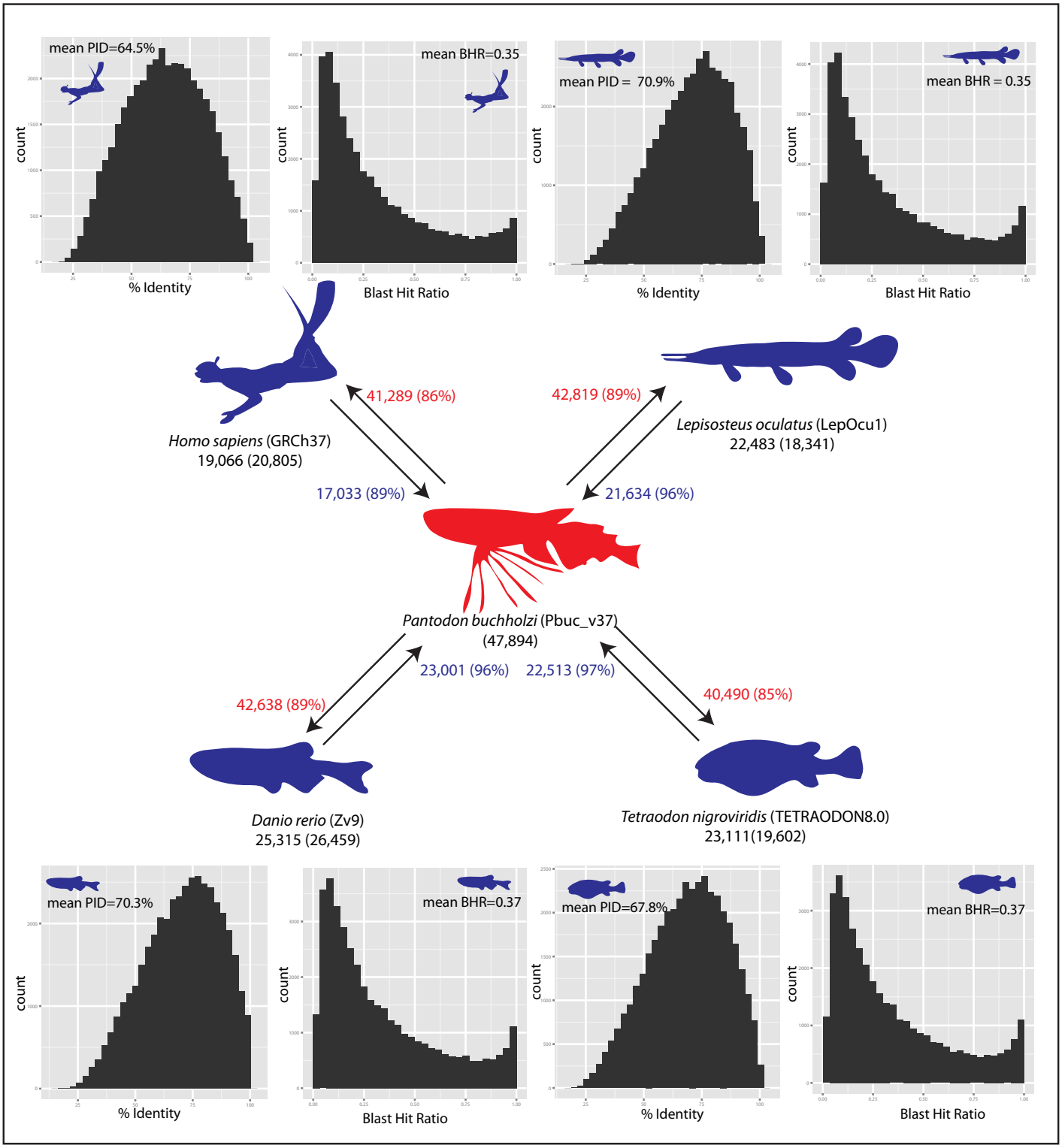



**Figure 3.7.** Reference genome BLAST analysis of the Pbuc_v37 final MAKER gene model set. Reciprocal blastp analyses were conducted using the Pbuc_v37 MAKER protein models to query the reference proteomes downloaded from the ENSEMBL BioMart webserver for human (GRCH37), gar (LepOcu1), zebrafish(Zv9), and *Tetraodon* (TETRAODON8.0), and reciprocally using each reference proteome to query the MAKER protein set. The number of significant (E=10-6) hits to Pbuc_v37 MAKER protein models are shown in red, while the number of significant hits to each reference genome is shown in blue.

# Chapter 4

## Ancient duplicated ANTP homeobox gene paralogs in the genome of *Pantodon buchholzi* reveals delayed gene loss following the teleost genome duplication


## Abstract

The evolutionary timing and dynamics of gene losses in different teleost lineages following the teleost genome duplication (TGD) are poorly understood, especially near the base of the teleost tree. Gene duplicates belonging to particular functional classes, such as growth and embryonic development, are more likely to be retained in duplicate following whole genome duplication (WGD) and therefore the evolution of developmental gene families, such as the well-known Hox and other ANTP class homeodomain-containing transcription factors, are particularly heavily influenced by WGD. Following on from large scale analyses of a *de novo* assembly of the *P. buchholzi* genome presented in Chapter 3, I carried out extensive manual annotation of the full complement ANTP class homeobox genes in this species. I revealed that *P. buchholzi* possesses at least 127 different ANTP class homeobox genes. This corresponds to a duplicate retention rate following the TGD of 18% based a reconstruction of the ANTP complement of the last neopterygian common ancestor (108 genes) and is similar to the retention rate in clupeocephalan teleosts including zebrafish (23%) and pufferfish (18%). Strikingly, after phylogenetic analysis of each of the 37 ANTP families present in vertebrates it was evident that there were significant differences in precisely which TGD paralogs were lost in *P. buchholzi* compared with clupeocephalan teleosts. Ancient TGD duplicates of 14 ANTP

subfamilies (cdx4, barx1, barx2, dlx1, emx1, lbx2, nkx2.1, hmx2, tlx1, tlx2, hoxb2, hoxb4, hoxb9, hoxb13) were retained uniquely in *P. buchholzi* and lost from all clupeocephalan teleosts. Reciprocally, several duplicated genes and even entire vertebrate subfamilies which are present in all other teleost genomes were not found in *P. buchholzi*. Finally, I identified a small set of only 8 subfamilies which retained both TGD duplicates in each of *P. buchholzi*, zebrafish and pufferfish (cdx1, mnx2, eng1, eng2, nkx2.4, nkx2.2, hmx3, hoxb5) and are therefore candidates for experiencing extremely early neofunctionalization or subfunctionalization biasing them towards long term retention prior to the radiation of the major teleost subdivisions.

## Introduction

The recent sequencing of the Hox clusters of the osteoglossomorph *P. buchholzi* revealed that it retained ancient TGD duplicates in the hoxb cluster not found in clupeocephalan teleosts, and even an additional duplicated gene pair (*hoxb13x/y*) not observed in the Hox gene replete genome of the eel (Chapter 2, Martin and Holland 2014). Intriguingly, *P. buchholzi* was also found to have lost Hox gene duplicates which have been retained in the genomes of all clupeocephalan teleosts, including one of a pair of hoxb1subfamily duplicates (*hoxb1a, hoxb1b)* which in zebrafish have experienced both subfunctionalization and neofunctionalization such that their roles in hindbrain development are both indispensable (McClintock et al. 2001; McClintock et al. 2002). Taken together these results from the fully sequenced hox clusters of elopomorphs and an osteoglossomorph highlight greater diversity in possible duplicate resolution outcomes in teleosts following the TGD, and demonstrate that the ancestor of crown teleosts (henceforth referred to as the last teleost common ancestor, or LTCA) retained almost all of their duplicated hox genes for the entire length of time between the TGD and the divergence of Osteoglossomorpha, Elopomorpha, and Clupeocephala.

It would therefore seems that the rapid gene collapse model (Sankoff et al. 2010; Inoue et al. 2015) following the TGD does not apply, at least to *hox* genes. It is also tantalizingly suggestive that the important processes of neofunctionalization and subfunctionalization which are central to the hypothesis of a role for gene duplication in the evolution of novelty might occur much later than previously thought, only after the divergence of the three major teleost subdivisions Osteoglossomorpha,

elopomorpha and clupeocephala. It is now imperative that the patterns of duplicate loss and retention in the genomes of basal teleosts are examined on a larger scale to verify if the striking pattern observed for *hox* genes applies to other gene families. If more ancient TGD duplicates absent from the genomes of all clupeocephalan teleosts genomes are found to be present in osteoglossomorphs or elopomorphs it could suggest that the LTCA was still extremely replete in unresolved gene duplicates, that diploidization was incomplete and that the LTCA was still partially karyotypically polyploid (pseudotetraploid).

Homeobox genes can be found in almost all eukaryotes including plants, and yeast and are divided into 11 classes consisting of over 100 separate gene families (Holland 2013). The ANTP class is a metazoan specific class, and with 37 gene families in vertebrates is also the largest. The history of expansion of ANTP class genes in the vertebrate lineage however does not begin with WGD. Early in bilaterian evolution, the ANTP family expanded massively through tandem gene duplication generating several clustered arrays of these developmentally important transcription factors including the Hox, NK, and ParaHox clusters (Garcia-Fernàndez 2005). This massive early expansion of ANTP class genes and the concomitant increases in developmental complexity that ensued has been suggested as a potential explanation for the explosion of morphological diversity of bilaterians in the Cambrian, the so-called Cambrian explosion (Holland 2015). In the vertebrate lineage the ANTP class expanded yet again, this time mostly as a result of the VGD1/VGD2 whole genome duplications. The pre-VGD1/VGD2 cephalochordate *B. floridae* has only 60 ANTP class genes (Takatori et al. 2008), whereas in Humans, the ANTP class has

expanded to 100 functional genes, with similar numbers found in other tetrapods (Holland et al. 2007; Zhong et al. 2008; Holland 2013). After the TGD the ANTP class expanded yet again and in the zebrafish, *D. rerio*, 132 ANTP class homeobox genes have been cataloged (Zhong et al. 2008). With each successive round of WGD in the lineage leading to teleosts (VGD1/VGD2/TGD) the size of the ANTP class has clearly grown. However, it is clear that zebrafish and other sequenced teleosts do not retain the predicted 8-fold higher number of ANTP genes compared with their pre-WGD ancestors, and therefore that gene loss has also played an important role in shaping the evolution of these transcription factor families.

Because of their strong bias towards retention following WGD and their important roles in development and evolution, I sought to characterize the ANTP complement of a representative basal teleost in the order Osteoglossomorpha, the African butterfly fish *P. buchholzi*. The patterns of duplicated gene loss or retention in this large class of genes will help answer more general questions about the relative timing of duplicate loss genome-wide. Additionally, as the evolution of this class of genes has a particularly heavy influence on the evolution development and morphology, understanding the evolutionary patterns and processes which have shaped their representation in teleost genomes may yield novel insights into the evolution of teleost features..

## Materials and Methods

### Identification of ANTP class homeobox genes in the *P. buchholzi* genome

Three parallel approaches were undertaken to bioinformatically isolate all putative ANTP class homeobox genes of *P. buchholzi*. (1) Automated annotation of the MAKER gene models derived from the *de novo* assembly of the *P. buchholzi* genome (Pbuc_v37) described in the methods section of Chapter 3 with the InterproScan pipeline (Jones et al. 2014). (2) Genome-wide tblastn searches of the Pbuc_v37 genome assembly under high stringency ($E<10^{-10}$) using the complete set of human and zebrafish homeobox proteins from HOMEODB (Zhong et al. 2008). (3) Alignment of the 456 TRAM-assembled putative homeobox-containing gene fragments reconstructed from the *P. buchholzi* raw read archive described in the methods section of Chapter 2 to the Pbuc_v37 genome assembly using blastn. The results of all automatic and manual annotation processes and alignments were as .gff files and visualized against the Pbuc_v37 genome assembly in the desktop GUI application Geneious (Biomatters Ltd.) Manual inspection of the identity and quality of each alignment to the genome was carried out, followed by manual refinement of intron-exon boundaries, start and stop codons using available evidence from MAKER (including *de novo* transcriptome alignments), tblastn searches, and the results of the InterproScan annotations.

### Phylogenetic analyses of ANTP class homeobox genes

Full length protein coding sequences of all ANTP class homeobox genes were extracted from the ENSEMBL genome browser (Hubbard et al. 2009) assembly

versions of 18 vertebrate taxa using either the ENSEMBL Biomart (http://www.ensembl.org/info/data/biomart/index.html) or ENSEMBL Gene Tree export functions. Taxa and genome assembly versions used were the teleosts: *Poecilia formosa* Poecilia_formosa-5.1.2*, Xiphophorus maculatus* Xipmac4.4.2, *Tetraodon nigroviridis* TETRAODON 8.0, Oreochromis niloticus Orenil1.0, Oryzias latipes HdrR, *Takifugu rubripes* FUGU 4.0, *Gadus morhua* gadMor1, *Gasterosteus aculeatus* BROAD S1, *Astyanax mexicanus* AstMex102, and *Danio rerio* GRCz9; the sarcopterygians: *Anolis carolinensis* AnoCar2.0*, Gallus gallus* Galgal4, *Latimeria chalumnae* LatCha1, *Homo sapiens* GRCh37, *Mus musculus* GRCm38, *Monodelphis domestica* monDom5, *Xenopus tropicalis* JGI 4.2, *Taeniopygia guttata* tae Gut3.2.4; and the basal actinopterygian *Lepisosteus oculeatus* LepOcu1. Alignments with the novel *P. buchholzi* proteins from the Pbuc_v37 assembly were carried out for each of 37 classes using the algorithm MAFFT with the L-INS-i settings (Katoh et al. 2002), and trimmed with the trimAl (Capella-Gutiérrez et al. 2009) either using the default –automated1 setting or a gap threshold of 25% whichever produced a better alignment after visual inspection. Unconstrained trees were then reconstructed using RAxML under the PROTGAMMALG amino acid substitution model running 1000 bootstrap replicates. When subfamily level branches of the ANTP family trees of interest were identified, (e.g. novel retention in *P. buchholzi* and pan-teleost retention candidates), sequences from the corresponding subfamily were re-aligned and trees were reconstructed independently using the same process, but without other subfamilies in order to produce subfamily-optimized results.

## Ancestral State Reconstructions

The ancestral gene complements of each node and therefore the gene losses along each branch of the cladogram were reconstructed manually by inferring that the ancestor must have possessed all of the genes present in any of the extant lineages branching from that node. For the teleost fish, the neopterygian gar *L. oculeatus* was used as an outgroup to infer losses. To infer the absence of gene losses along the gar lineage, the 112 ANTP genes found in gar were compared with the human and elephant shark outgroups and it was observed that they neither gained nor lost any genes following their divergence from the lineage leading to teleosts. A maximum parsimony approach was employed in order to minimize the total number of independent gene losses where it was unclear whether a gene was lost once or multiple times. In order to produce a rough estimation of absolute dates for illustrative purposes, I used the mean of the range of divergence times estimated from the mitogenome clock analyses of Hurley et al. 2007 to estimate rates of gene loss and in order to avoid complications due to potential independent diploidization of nuclear gene divergence date estimates.

## Results

### ANTP homeobox gene annotation

I pursued three parallel strategies in order to extract the full complement of ANTP class homeobox genes from the genome of *P. buchholzi.* I began by identifying all putative homeobox-containing loci in the genome. First, I identified MAKER gene models which were annotated with the Pfam domain PF00046 (Homeobox), or PANTHER class PC00119 (encompassing multiple Homeobox-containing PANTHER families) by InterproScan. In this way, I identified 253 putative homeobox-containing transcripts with a mean length of 1.37kb which were present on 248 contigs. Next, I conducted stringent blastx searches ($E=10^{-10}$) of the assembly using the manually curated homeodomain protein sequences of zebrafish and human from the HOMEODB database (Zhong et al. 2008). This resulted in the isolation of 343 contigs with significant hits. Finally, using the target-restricted assembly method (TRAM) (Johnson et al. 2010) I searched the unfiltered illumina genomic read databases and assembled the resultant read subset into a set of 456 contigs representing putative exonic homeobox fragments as previously described (Chapter 2, Martin and Holland 2014). These 456 contigs were searched by blastx against the HOMEODB database and 410 homeobox-positive TRAM fragments were retained. These short TRAM contigs were then aligned to the Pbuc_v37 assembly and 367 genomic contigs with high quality (95% identity over their whole length) TRAM alignments were extracted. 18 TRAM putative homeobox fragments did not produce an alignment to the genome assembly. These were manually inspected and 9 were discarded from the analyses as their top blast-hit in the NCBI nr database was to a

non-chordate (3 to *Arabidopsis,* 2 to *Oryza*, and 4 to bacterial strains) and therefore likely represented environmental contamination of the Illumina lane during sequencing. The remaining 9 TRAM vertebrate homeobox-like fragments which were not found in the assembly were classified according to their greatest similarity in HOMEODB. All other TRAM fragments were accounted for in the genome assembly.

After manual inspection of MAKER, HOMEODB, and TRAM evidence on each of the contigs I converged on a total homeobox-containing set of 346 genomic contigs. I roughly annotated each according to the class of homeodomains. In this way, I identified 130 ANTP-class homeobox containing contigs. I further identified contigs on which I could identify homeobox fragments from the PRD (67), LIM (24), POU (38), HNF (6), SINE (8), TALE (31), CUT (15), PROS (5), ZF (26), CERS (6), and other/unassigned (5) classes indicating that the *P. buchholzi* genome assembly contains a reasonable number of homeobox genes from all classes expected to be present in vertebrates (e.g. 322 total annotated homeobox genes zebrafish, 333 annotated homeobox containing genes in human according to HOMEODB (Zhong et al. 2008)).

After careful manual curation of the MAKER gene models overlapping the ANTP containing contigs using all available reference protein alignments, and the *P. buchholzi* brain/muscle merged transcriptome EST evidence I determined that 127 ANTP class genes were represented in the *P. buchholzi* assembly. The 127 gene set described here is the lowest possible high-confidence estimate of ANTP-class homeobox genes present in the *P. buchholzi* genome. Three further putative ANTP-like TRAM homeobox fragments were isolated which did not align >95% to any of the

other 127 putative ANTP class genes, nor anywhere else in the genome but whose top blast hit classified them as ANTP-class homeobox fragments (2 hmx-like and 1 nkx2.1-like). These might be fragments of either coding or pseudogenized homebox genes absent from the assembly, or spurious assemblies of low quality reads and were therefore not included in the high-confidence gene set. The inclusive set of ANTP class genes for which I could detect any evidence in *P. buchholzi* is therefore 127-130. This is a similar figure to the number of ANTP class genes found in well annotated genomes of the ostarioclupeomorph *D. rerio* (132) and the euteleost *T. nigroviridis* (127) and indicates a striking convergence on a similar overall duplication retention rate between an osteoglossomorph and clupeocephalan teleosts. However, while it has previously been noted that there has been systematic reciprocal loss of TGD paralogs between ostarioclupeomorphs and euteleosts like zebrafish and *T. nigroviridis* respectively (Sémon and Wolfe 2007; Garcia de la Serrana et al. 2014), I cannot determine from gene counts alone to what degree *P. buchholzi* has experienced shared or reciprocal paralog loss with other teleosts. In order to answer this question, careful phylogenetic analysis of each individual vertebrate level ANTP gene family and all TGD-generated paralogs from *P. buchholzi* and other teleosts was carried out.

**Phylogenetic analysis and divergent resolution of ANTP class homeobox genes**

In order to evaluate the degree to which the resolution of ANTP-class homeobox gene duplicates occurred either independently in *P. buchholzi* and clupeocephalan teleosts, or jointly in their common ancestor immediately following the TGD, I reconstructed gene genealogies for each of the novel *P. buchholzi* proteins. In vertebrates, there are 37 recognized families of homeobox genes within the ANTP class (Fonseca et al. 2008; Zhong et al. 2008; Holland 2013). Each family is usually represented by a single gene in the cephalochordate *B. floridae* while up to 4 genes are found in vertebrates as a result of the first two rounds of WGD common to all vertebrates (VGD1/VGD2) (Putnam et al. 2008; Takatori et al. 2008; Zhong et al. 2008). For example, within the ANTP class the CDX family is comprised of a single gene in *B. floridae* (*cdx*) and 3 genes in non-teleost vertebrates (*Cdx1*, *Cdx2* and *Cdx4*) as a result of VGD1/VGD2 followed by the loss of one gene. In teleosts however, as a result of the TGD, each family could theoretically consist of up to 8 genes with up to two co-orthologs of each vertebrate gene. For example, in zebrafish two paralogs of the cdx1 subfamily are retained (*cdx1a* and *cdx1b*). For the sake of classifying TGD level paralogs, I therefore consider each vertebrate-level gene as a subfamily. I can therefore classify *P. buchholzi* genes at the family level (e.g. CDX), the vertebrate subfamily level (e.g. cdx1, cdx2, or cdx4), and in theory the specific teleost paralogy group level (*cdx1a*, *cdx1b*).

Using the 127 manually curated protein sequences from *P. buchholzi* which had already been assigned a putative family-level homology based on the MAKER/Uniprot-Swissprot/HOMEODB blast-based annotations, together with a

broad taxonomic sampling of ANTP class genes from other vertebrates (including 10 teleosts, 8 sarcopterygians, and one pre-TGD actinopterygian), I first built 37 family-level gene trees in order to evaluate homology of the novel *P. buchholzi* ANTP proteins to particular vertebrate subfamilies. All phylogenetic analyses were carried out in a ML framework using RAxML under the PROTGAMMALG model as previously described (Chapter 2, Martin and Holland 2014). In this way I was able to unambiguously assign vertebrate subfamily level homology of *P. buchholzi* proteins. In Table 4.1, I present an updated classification scheme of all ANTP proteins from human, gar, and zebrafish alongside the newly discovered genes in *P. buchholzi*. ANTP proteins are classified by subclass (HOXL, NKL), family, and subfamily. I note that several gene names assigned to many gar and zebrafish proteins in the ENSEMBL database are not consistent with the phylogenetic reconstructions produced here, and in Table 4.1, I denote with an asterisk (*) where a proposed change in name would more accurately reflect both family-level and subfamily-level relationships. I next inspected the 37 resultant family-level trees and identified subfamilies where *P. buchholzi* retained either two, one or no duplicates and compared these results with other teleosts. In total, I found that *P. buchholzi* retained both paralogs for 19 of 108 additional TGD-generated duplicates which corresponds to a retention rate of 17.6%. This is a comparable to the ANTP retention rate I determined for zebrafish (23.1%) and *T. nigroviridis* (17.6%) and in line with previously published genome wide estimates which have suggested between 12-24% of additional TGD generated paralogs are retained in teleosts (Postlethwait et al. 2000; Jaillon et al. 2004; Woods et al. 2005; Brunet et al. 2006; Kassahn et al. 2009).

Despite maintaining an overall similar paralog retention rate to other teleosts, I revealed several striking differences in the precise pattern of TGD duplicate retention in *P. buchholzi*. This includes both the preservation of TGD duplicates never before seen in other teleosts as well as *P. buchholzi* specific gene family losses. These results are summarized in Figure 4.1. The most noteworthy observation was that I could identify 14 ANTP gene subfamilies (cdx4, hmx2, nkx2.1, lbx2, dlx1, emx1, barx1, barx2, hoxb13, hoxb9, hoxb4 and hoxb2) where *P. buchholzi* has retained both TGD duplicates while clupeocephalan teleosts retained only one. All of the family level (panel i) and subfamily level (panels ii,iii) trees for each of these *P. buchholzi* specific TGD duplicates are presented in Figure 4.1 A-G. This means that both copies of these 14 genes were retained for ~50 My between the TGD dated at 350 Mya (Jaillon et al. 2004) and the LTCA dated at 297 Mya (Hurley et al. 2007) before being specifically lost in the lineage leading to clupeocephalans. Of these subfamilies, retained duplicates for hoxb13, hoxb9, hoxb4 and hoxb2 were previously described in my analysis of the *P. buchholzi* hox clusters (Chapter 2, Martin and Holland 2014) and 3 duplicated pairs were also found in the elopomorph *A. anguilla* (hoxb9, hoxb3, hoxb2) (Henkel, Dirks, et al. 2012; Henkel, Burgerhout, et al. 2012), while this is the first time that both TGD paralogs of the remaining 8 subfamilies (cdx4, hmx2, nkx2.1, lbx2, dlx1, emx1, barx1, barx2) have ever been reported in any teleost. I carried out a further phylogenetic classification step of each subfamily with *P. buchholzi*-specific TGD duplicates and found that, although support for its classification within a particular gene subfamily was good (suggesting it is a real member of that subfamily), there was often no strong support for the 1:1 homology of

the singular clupeocephalan duplicate with either of the *P. buchholzi* duplicates. In some cases, however, I found that one or the other *P. buchholzi* duplicate segregated with the singleton retained in clupeocephalans. In the case of the cdx4 subfamily, support for inclusion of *P. buchholzi cdx4x* within the clupeocephalan clade was 88% (Figure 4.1, A). In the case of the nkx2.1 subfamily, the *P. buchholzi nkx2.1y* paralog segregated together with the clupeocephalan nkx2.1 subfamily singleton with a support of 89% (Figure 4.1, C). In the case of the dlx1 subfamily, the *P. buchholzi dlx1y* paralog clustered with the clupeocephalan singleton with a support value of 76% (Figure 4.1, E). For each of the remaining subfamilies with ancient TGD paralogs in *P. buchholzi* (lbx2, emx1, barx1, barx2) there was no significant support for one or the other paralog being more closely related to the clupeocephalan singletons (Figure 4.1, D, F, G). The discovery of these 14 genes in *P. buchholzi* shifts a significant proportion of duplicated ANTP gene losses away from the stem node between the TGD and the LTCA until after the deep divergence of the osteoglossomorpha and clupeocephala and suggests that the resolution of many more duplicated genes could have occurred significantly later than previously thought.

I next identified a small subset of 7 subfamilies (cdx1, mnx2, eng1, eng2, nkx2.4, nkx2.2, hmx3) where both TGD paralogs were retained in *P. buchholzi* as well as zebrafish, *T. nigroviridis* and most other teleosts (Figure 4.2). This indicates that, while the overall proportion of ANTP TGD duplicates retained in various teleost genomes might be reasonably high (18-23%) only a significantly smaller subset of these duplicates (6.5%) are conserved in the genomes of most teleosts. In other

words, the majority of retained TGD duplicates have experienced divergent resolution at the level of the major teleost subdivisions Clupeocephala and Osteoglossomorpha.

Because these 'pan-teleost TGD' genes do not suffer from problems of hidden paralogy they represent good candidates to test the tetralogy hypothesis from Chapter 2 by examining whether their 1:1 orthology with the post-TGD 'a' and 'b' paralogy groups of clupeocephalan teleosts could be faithfully reconstructed. By reconstructing ML gene trees, I found that in only 2 of these 7 cases did the *P. buchholzi* paralogs segregate as clear 1:1 orthologs with either the clupeocephalan teleost 'a' or 'b' post-TGD clades: cdx1 (Figure 4.2, A) and hmx3 (Figure 4.2, F). However, bootstrap support at each of these nodes was not high and these results must therefore be interpreted cautiously. I note that support values recovered for many of the other clupeocephalan 'a' or 'b' post-TGD clades was also generally quite low, although in every case these analyses did manage to recover them. In three of the remaining 'pan-teleost TGD' duplicate cases I found that both *P. buchholzi* duplicates segregated together with only the clupeocephalan 'b' clade: eng1 (Figure 4.2, C), nkx2.4 (Figure 4.2, E) and nkx2.2 (Figure 4.2, G). The remaining two cases revealed that one paralog segregated with the clupeocephalan 'a' clade, while the other segregated outside both the 'a' and 'b' clades as might be expected from a pre-TGD paralog: mnx2 (Figure 4.2, B) and eng2 (Figure 4.2, D). Bootstrap support for nodes containing the *P. buchholzi* sequences were also quite low in these cases.

Overall, results of these phylogenetic analysis could not conclusively resolve the question of homology of the *P. buchholzi* sequences with either the 'a' or 'b' paralogy groups of clupeocephalan teleosts, although some show more support than

others. This is consistent with my expectation under the tetralogy model of independent diploidization of many of these loci following the divergence of osteoglossomorphs and clupeocephalans (Chapter 2, Martin and Holland 2014). However due to the short length of the sequences used for phylogenetic reconstruction and the relatively long branch between the singular osteoglossomorph species and the group of clupeocephalan teleosts, it is far from conclusive and would benefit from synteny data and reconstruction of gene genealogies of neighbouring genes, or ideally even genome-wide.

**Ancestral state reconstructions and rates of gene loss**

I manually reconstructed the ANTP homeobox gene complement in the ancestral state of the last neopterygian common ancestor (LNCA), the last clupeocephalan common ancestor (LCCA), and the last common ancestor of crown teleosts (LTCA) and mapped putative gene losses along each branch of the teleost tree (Figure 4.3). My reconstruction of the LNCA revealed it possessed 112 ANTP class homeobox genes and that 4 entire subfamilies (cdx2, pdx2, hoxa6 and hoxd2) were lost selectively in the teleost lineage, possibly prior to the TGD, while the gar demonstrated remarkable evolutionary conservation, suffering no apparent losses. Prior to the TGD the gene count in teleosts was therefore 108 and following the TGD it was doubled to 216. Leveraging this new genomic data from *P. buchholzi* I could infer that the LTCA retained at least 171 genes, or 58% of TGD-generated duplicates. Following the divergence of osteoglossomorphs and clupeocephalans, 21 losses occurred on the lineage leading to the LCCA which retained at least 150 genes, corresponding to a TGD duplicate retention rate of 39%. Subsequently, along the

zebrafish lineage an additional 17 gene losses occurred resulting in a total complement of 133 genes and a retention rate of 23%, while along the *T. nigroviridis* lineage 23 losses occurred resulting in a complement of 127 genes and retention rate of 18%. Since diverging from the LTCA, this data suggests *P. buchholzi* lost 44 ANTP class genes resulting in its final complement of 127 genes.

There are several caveats associated with producing absolute divergence time estimates between lineages (Graur & Martin, 2004). In the case of giving an absolute date for the TGD this is especially difficult as the molecular clock of mutational changes which delineate paralogs generated by the TGD and which is generally used to time the duplication event itself, does not actually start 'ticking' until after the allele to paralog transition associated with diploidization (Doyle and Egan 2010). In osteoglossomorphs such as *P. buchholzi* I have previously shown that there is evidence that diploidization of different regions of the genome occurred in a stepwise, segmental fashion and independently from other teleosts (Chapter 2, Martin and Holland 2014) further suggesting that the use of different genes and different teleost lineages may recover vastly different dates for the TGD. The discovery of 14 *P. buchholzi* specific paralogs shifted the earliest possible timing of their loss in other teleosts to in the region of 50 My, but perhaps as little as 10 My following the TGD. Together with data from eel hox clusters, the loss of 21 genes in the clupeocephalan lineage representing 18% of all TGD paralogs can now be shown to have occurred sometime in the between the LTCA and the LCCA, 50-120 My following the TGD.

Next, by using the mean of the range of divergence times estimated by Hurley et al. 2007 I estimated the gene loss rate along key branches of the early teleost tree

(Figure 4.4). With the addition of *P. buchholzi* as a representative osteoglossomorph I was able to estimate gene loss rates for the first ~50 My following the TGD before the divergence of osteoglossomorphs and clupeocephalans separately from the next ~70 My prior to the divergence of ostarioclupeomorphs and euteleosts. Overall, I found that the gene loss rate is still apparently highest in the first ~50 My after the TGD with a loss of 46 genes indicating a gene loss rate of 0.87 genes/My or 0.4% of TGD-generated duplicates/My. In the following ~70 My between the LTCA and the LCCA the loss rate drops by approximately half to 0.29 genes/My or 0.17% of TGD-generated duplicates/My. Subsequently, in the lineage leading to zebrafish the loss rate drops again to ~0.08 genes/My and in *Tetraodon* to ~0.1 genes/Ma. In the *P. buchholzi* lineage the average gene loss rate over the full ~300 My since it split from the LTCA was ~0.15 genes/My.

## Discussion

While inferring a gene loss is patently more difficult than demonstrating presence, I note several putative gene losses from *P. buchholzi* which, should they be verified, represent another significant departure from the situation seen in clupeocephalans. For example I found several subfamilies (dbx1, lbx1, msx2, and tlx3) where only one paralog was identified in *P. buchholzi* while other teleosts retained both paralogs. Perhaps more surprisingly, in other cases *P. buchholzi* appears to have lost both members of an entire vertebrate-level subfamily. No paralogs of the meox1, dlx3, dlx4, nkx2.8, nkx3.3, or nkx2.3 subfamilies could be detected in the *P. buchholzi* genome while at least one paralog could be found in the genomes of clupeocephalan teleosts. The lineage specific loss of an entire vertebrate subfamily is not unprecedented as it has been shown to have occurred in teleosts previously with the lineage specific loss of the cdx2 subfamily (Mulley et al. 2006) and has even occurred multiple times independently as in the case of the pdx2 subfamily which has been lost independently in both teleosts and tetrapods (Mulley and Holland 2010). The apparent loss of both dlx3 and dlx4 subfamilies is notable in that it is the first example where both paralogs of more than one member of an ANTP family has been lost in teleosts. Previous studies have also suggested that dlx3 and dlx4 subfamilies are more highly derived than other dlx gene clusters both at the level of their protein coding sequences, and conservation of *cis*-regulatory DNA (Sumiyama et al. 2002; Ghanem et al. 2003; MacDonald et al. 2010) but they have not been lost from other vertebrates. It is still an outstanding question how a gene which has been conserved for hundreds of millions of years, presumably due to

strong selective constraint, could subsequently be lost. One possible explanation is that lingering redundancy between other members of the gene family could compensate for the loss of a previously fixed paralog by re-consolidating previously partitioned gene functions in a process known as 'function shuffling' (McClintock et al. 2002). It is therefore interesting that I identified an additional member of the dlx1 subfamily in the *P. buchholzi* genome which is not present in other teleosts and that the total number of dlx genes in *P. buchholzi* is 6 which is the same number as found in pre-TGD vertebrates. The loss of dlx3 and dlx4 could therefore have been compensated for by preservation of dlx genes from other subfamilies in *P. buchholzi* and this could present an interesting case to test the function shuffling model.

In most cases, I could not assign clear 1:1 orthology of the *P. buchholzi* ANTP duplicates with duplicates in other teleosts. These results show that over half (58%) of all ANTP class homeobox gene duplicates were still retained in the ancestral teleost genome approximately 50 My following the TGD. This supports a slow and stepwise model of diploidization and gene loss following the TGD, in contrast to the rapid gene collapse model. The discovery of 14 previously unobserved TGD duplicates in *P. buchholzi* implies that they were preserved in the teleost lineage for approximately 50 My following the TGD before the divergence of Osteoglossomorpha and Clupeocephala. The presence of so many newly discovered ancient TGD duplicates amongst one class of genes in a single species of basal teleost highlights the ancient duplication history of all teleost genes and calls into question the reality of the existence of "single-copy" genes in teleosts.

It is also interesting to note that if simple phenetic approaches such as reciprocal best hit (RBH) with BLAST approaches had been used to carry out homeobox gene classification, most of the reciprocal losses detected using phylogenetic methods would have gone unnoticed. This is because where reciprocal loss of TGD-generated paralogs occurs in two lineages, the remaining singletons would recover each other as their respective RBH matches and the pattern could be misconstrued as an early shared loss prior to lineage divergence. This highlights again the importance of reconstructing accurate gene tree genealogies with phylogenetic methods prior to inferring patterns of gene losses following duplication.

I note that although the gene loss rate inferred in the teleost stem lineage between the TGD and the LTCA is apparently more rapid than any other period I measured, it is an absolute maximum as I cannot rule out the possibility that any apparently shared losses in *P. buchholzi* and clupeocephalans are not in fact independent losses in each lineage. Furthermore, I expect that with the addition of more osteoglossomorph genome data it will be found that different osteoglossomorph lineages have experienced divergent resolution just as has been found between ostarioclupeomorphs and euteleosts and that the last osteoglossomorph common ancestor likely had significantly more ANTP class genes than I can infer from examining *P. buchholzi* alone. Consequently, the number of ANTP genes conserved in the genome of the LTCA is also probably significantly higher than I can infer from *P. buchholzi* alone. As more genome data becomes available our view of the gene content in the LTCA, and the gene loss rate in the initial ~50 My following the TGD is likely to shift considerably.

## Conclusions

In this chapter, I have phylogenetically classified 127 novel members of the ANTP class of homeobox genes in *P. buchholzi* and demonstrated that several genes which have been reduced to a single copy in clupeocephalan teleosts retained two copies in the *P. buchholzi* genome. This indicates that they were still present in the genome of the LTCA, and highlights the ancient duplication history of all teleost genes. Overall, I have shown that the LTCA still retained at least 58% of ANTP duplicates ~50 My following the TGD. Furthermore, there was rarely any strong support for the 1:1 orthology of any *P. buchholzi* duplicates with other clupeocephalan duplicates. This is consistent with a model of independent diploidization in osteoglossomorphs and clupeocephalans (Martin and Holland 2014), and suggests that one reason that so many duplicates could remain unresolved in the genome of the LTCA could be that they were still segregating as tetraploid alleles, rather than independent paralogous loci. These results have important implications for attempts to reconstruct the phylogenetic relationships of basal teleosts using nuclear gene data and suggests that the widespread practice of using 'single-copy' genes during phylogenetic reconstruction of teleosts relationships in order to 'avoid complications due to the TGD' is fundamentally flawed.

**Table 4.1.** Classification of all ANTP proteins in *P. buchholzi,* human, gar, and zebrafish.

For each gene within each family and subfamily I list either the recognized gene name and ENSEMBL protein ID for each of the human, gar and zebrafish proteins, or the putative gene name, genomic contig loci, and any supporting transcriptome evidence for novel proteins in *P. buchholzi*. Rows are organized first by subclass (HOXL, NKL), followed by family (e.g. Cdx), followed by subfamily (e.g. cdx1, cdx2 etc.) and then each member of the subfamily in human, gar, zebrafish and *P. buchholzi* are listed in separate columns. I note subfamilies with *P. buchholzi*-specific duplicates or putative pan-teleost duplicates in the rightmost columns. (4 proceeding pages)

| | | | Genes in pre-TGD subfamily | | | Genes in post-TGD subfamily | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Human | Gar | Zebrafish | Pantodon | | | | | |
| **Subclass** | **Family** | **Subfamily** | **Gene Name (ENSEMBL_ID)** | **Gene Name (ENSEMBL_ID)** | **Gene Name (ENSEMBL_ID)** | **Gene Name** | **Genome Contig** | **Transcriptome Contig** | ***Pantodon* only paralogs** | ***Pantodon* specific losses** | ***Pan* /Ostario./ Eutel. Retention** |
| | Barhl | BARHL1 | BARHL1 (ENSP00000263610) | barhl1* [barhl1b] (ENSLOCP00000006769) | barhl1a (ENSDARP00000051473) | barhl1x | NODE_2010937 | *comp52549_c0_seq1* | | | |
| | | | | | barhl1b (ENSDARP00000016114) | barhl1y | NODE_36237 | *comp42483_c0_seq1* | | | |
| | | BARHL2 | BARHL2 (ENSP00000359474) | barhl2 (ENSLOCP00000005994) | barhl2 (ENSDARP00000093436) | barhl2 | NODE_988097 | *comp28075_c0_seq1* | | | |
| | | | | | X | X | X | *X* | | | |
| | Barx | BARX1 | BARX1 (ENSP00000253968) | barx1 (ENSLOCP00000017328) | barx1 (ENSDARP00000014797) | **barx1x** | NODE_1151617 | - | √ | | |
| | | | | | X | **barx1y** | NODE_307760; NODE_4443286 | - | | | |
| | | BARX2 | BARX2 (ENSP00000281437) | barx2 (ENSLOCP00000002868) | barx2 (ENSDARP00000120269) | **barx2x** | NODE_681865; NODE_558790 | *comp419927_c0_seq1* | √ | | |
| | | | | | X | **barx2y** | NODE_574362 | - | | | |
| | Bsx | BSX | BSX (ENSP00000344285) | bsx (ENSLOCP00000006615) | bsx (ENSDARP00000090755) | bsx | NODE_395935 | *comp323498_c0_seq1* | | | |
| | | | | | X | X | X | X | | | |
| | Dbx | DBX1 | DBX1 (ENSP00000436881) | dbx1* [dbx1b] (ENSLOCP00000005875) | dbx1a (ENSDARP00000023732) | dbx1 | NODE_302165 | - | | √ | |
| | | | | | dbx1b (ENSDARP00000013350) | X | X | *X* | | | |
| | | DBX2 | DBX2 (ENSP00000331470) | dbx2 (ENSLOCP00000020295) | dbx2 (ENSDARP00000057729) | dbx | NODE_1016269 | *comp501363_c0_seq1* | | | |
| | | | | | X | X | X | *X* | | | |
| | Dlx | DLX1 | DLX1 (ENSP00000354478) | dlx1* [dlx1a] (ENSLOCP00000009229) | dlx1a (ENSDARP00000011033) | **dlx1x** | NODE_56374; NODE_156405 | *comp63089_c3_seq4; comp53090_c0_seq2* | √ | | |
| | | | | | X | **dlx1y** | NODE_285933 | *comp63089_c3_seq17;* | | | |
| | | DLX2 | DLX2 (ENSP00000234198) | dlx2* [dlx2b] (ENSLOCP00000009211) | dlx2a (ENSDARP00000103004) | dlx2x | NODE_1025790; NODE_256252 | *comp256447_c0_seq1* | | | |
| | | | | | dlx2b (ENSDARP00000007080) | dlx2y | NODE_250444 | *comp27408_c0_seq1* | | | y |
| | | DLX3 | DLX3 (ENSP00000389870) | dlx3* [dlx3b] (ENSLOCP00000016404) | x | X | X | *X* | | √ x 2 | |
| | | | | | dlx3b (ENSDARP00000022291) | X | X | *X* | | | |
| | | DLX4 | DLX4 (ENSP00000240306) | dlx4* [dlx4a] (ENSLOCP00000016398) | dlx4a (ENSDARP00000122639) | X | X | *X* | | √ x 2 | |
| | | | | | dlx4b (ENSDARP00000096665) | X | X | *X* | | | |
| | | DLX5 | DLX5 (ENSP00000222598) | dlx5* [dlx5a] (ENSLOCP00000013515) | dlx5a (ENSDARP00000100492) | dlx5x | NODE_650611; NODE_1309274 | *comp36931_c0_seq1* | | | |
| | | | | | X | X | X | *X* | | | |
| | | DLX6 | DLX6 (ENSP00000428480) | dlx6* [dlx6a] (ENSLOCP00000013503) | dlx6a (ENSDARP00000093553) | dlx6x | NODE_342098; NODE_1309274 | *comp2902_c0_seq1; comp217761_c0_seq1* | | | |
| | | | | | X | X | X | *X* | | | |
| | Emx | EMX1 | EMX1 (ENSP00000258106) | emx1 (ENSLOCP00000011406) | emx1 (ENSDARP00000057809) | **emx1x** | NODE_179863 | *comp59942_c0_seq1* | √ | | |
| | | | | | X | **emx1y** | NODE_388694 | - | | | |
| | | EMX2 | EMX2 (ENSP00000450962) | emx2 (ENSLOCP00000011841) | emx2 (ENSDARP00000058035) | emx2 | NODE_645033; NODE_616168 | *comp98547_c0_seq1* | | | |
| | | | | | X | X | X | *X* | | | |
| | | X | X | emx3 (ENSLOCP00000014949) | emx3 (ENSDARP00000018933) | emx3 | NODE_151638; NODE_281918 | *comp248080_c0_seq1* | | | |
| | | | | | X | X | X | *X* | | | |
| | En | EN1 | EN1 (ENSP00000295206) | eng1* [eng1b] (ENSLOCP00000003773) | eng1a (ENSDARP00000023487) | **eng1x** | NODE_495381; NODE_572207 | *comp148371_c0_seq1* | | | √ |
| | | | | | eng1b (ENSDARP00000012379) | **eng1y** | NODE_990420; NODE_1089411 | *comp45187_c0_seq1* | | | |
| | | EN2 | EN2 (ENSP00000297375) | eng2* [eng2b] (ENSLOCP00000015391) | eng2a (ENSDARP00000028905) | eng2x | NODE_1947072 | *comp26185_c0_seq1* | | | √ |
| | | | | | eng2b (ENSDARP00000056747) | eng2y | NODE_205550; NODE_109127 | - | | | |
| | Hhex | HHEX | HHEX (ENSP00000282728) | hhex (ENSLOCP00000008533) | hhex (ENSDARP00000140473) | hhex | NODE_314414; NODE_471432 | *comp176277_c0_seq1* | | | |
| | | | | | X | X | X | *X* | | | |
| | Hlx | HLX | HLX (ENSP00000355870) | hlx (ENSLOCP00000020245) | hlx1 (ENSDARP00000116160) | hlx | NODE_887530; NODE_1336457 | *comp47961_c0_seq2* | | | |
| | | | | | X | X | X | *X* | | | |
| | Lbx | LBX1 | LBX1 (ENSP00000359212) | lbx1* [lbx1a] (ENSLOCP00000014309) | lbx1a (ENSDARP00000045350) | lbx1 | NODE_134223; NODE_357761 | *comp487526_c0_seq1* | | √ | |
| | | | | | lbx1b (ENSDARP00000014246) | X | X | *X* | | | |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| NKL Subclass | Lbx | LBX2 | LBX2 (ENSP00000366789) | lbx2 (ENSLOCP00000002192) | lbx2 (ENSDARP00000029690) | **lbx2x** | NODE_1108199 | - | √ | | |
| | | | | | *X* | **lbx2y** | NODE_393308; NODE_56008 | *comp52976_c0_seq1* | | | |
| | Msx | MSX1 | MSX1 (ENSP00000372170) | msx1* [msxe] (ENSLOCP00000012235) | msx1a* [msxe] (ENSDARP00000012308) | msx1 | NODE_419342 | *comp26621_c0_seq1* | | | |
| | | | | | *X* | x | x | *x* | | | |
| | | MSX2 | MSX2 (ENSP00000239243) | msx2* [msxb] (ENSLOCP00000013291) | msx2a* [msxd] (ENSDARP00000110660) | msx2x | NODE_990566 | *comp256264_c0_seq1* | | √ | |
| | | | | | msx2b* [msxa] (ENSDARP00000073690) | x | x | *x* | | | |
| | | MSX3 | x | msx3 [msxc] (ENSLOCP00000005170) | msx3a* [msxc] (ENSDARP00000024644) | msx3x | NODE_773851 | *comp326620_c0_seq1* | | | √ |
| | | | | | msx3b* [msxb] (ENSDARP00000015954) | msx3y | NODE_849703 | *comp53282_c0_seq1* | | | |
| | Nanog | NANOG | NANOG (ENSP00000229307) | nanog (ENSLOCP00000001251) | nanog (ENSDARP00000102376) | nanog | NODE_985444; NODE_199381 | - | | | |
| | | | | | x | x | x | x | | | |
| | Nk1 | NKX1-1 | NKX1-1 (ENSP00000407978) | nkx1.1* [nkx1.2b] (ENSLOCP00000013491) | nkx1.1* [nkx1.2lb] (ENSLOCP00000013491) | nkx1.1 | NODE_576472; NODE_94652 | - | | | |
| | | | | | *x* | x | x | *x* | | | |
| | | NKX1-2 | NKX1-2 (ENSP00000451945) | nkx1.2* [nkx1.2la] (ENSLOCP00000009821) | nkx1.2* [nkx1.2la] (ENSDARP00000108802) | nkx1.2 | NODE_90328 | - | | | |
| | | | | | *x* | x | x | *x* | | | |
| | | *novel in actinopterygii?* | x | pnx* [novel transcript] (LOC102697275) | pnx (ENSDARP00000034835) | pnx | NODE_641422 | - | | | |
| | | | | | *x* | x | x | *x* | | | |
| | Nk2.1 | NKX2-1 | NKX2-1 (ENSP00000346879) | nkx2.1 (ENSLOCP00000017254) | nkx2.1 (ENSDARP00000025546) | **nkx2.1x** | NODE_258184 | *comp53868_c0_seq1* | | | |
| | | | | | *x* | **nkx2.1y** | NODE_300354 | *comp166936_c0_seq1* | √ | | |
| | | NKX2-4 | NKX2-4 (ENSP00000345147) | nkx2.4* [nkx2.4a] (ENSLOCP00000020858) | nkx2.4a (ENSDARP00000104286) | nkx2.4x | NODE_431471 | *comp256395_c0_seq1* | | | |
| | | | | | nkx2.4b (ENSDARP00000004227) | nkx2.4y | NODE_155257 | *comp166936_c0_seq1* | | | √ |
| | Nk2.2 | NKX2-2 | NKX2-2 (ENSP00000366347) | nkx2.2 [nkx2.2a] (ENSLOCP00000020856) | nkx2.2a (ENSDARP00000117693) | nkx2.2x | NODE_68327 | *comp41170_c0_seq1* | | | |
| | | | | | nkx2.2b (ENSDARP00000068914) | nkx2.2y | NODE_958566 | *comp229510_c0_seq1* | | | √ |
| | | NKX2-8 | NKX2-8 (ENSP00000258829) | nkx2.8 [nkx2.2b] (ENSLOCP00000017252) | nkx2.8 (BC091676.1) | x | x | *x* | | | |
| | | | | | *x* | x | x | *x* | | √ x 2 | |
| | Nk3 | NKX3-1 | NKX3-1 (ENSP00000370253) | nkx3.1 (ENSLOCP00000018485) | nkx3.1 (ENSDARP00000115852) | nkx3.1 | NODE_43710 | - | | | |
| | | | | | *x* | x | x | *x* | | | |
| | | NKX3-2 | NKX3-2 (ENSP00000371875) | nkx3.2 (ENSLOCP00000012087) | nkx3.2 (ENSDARP00000054821) | nkx3.2 | NODE_154032; NODE_284484 | - | | | |
| | | | | | *x* | x | x | *x* | | | |
| | | *x* | *x* | nkx3.3 (ENSLOCP00000011004) | nkx3.3 (ENSDARP00000107444) | x | x | *x* | | | |
| | | | | | *x* | x | x | *x* | | √ x 2 | |
| | Nk4 | NKX2-5 | NKX2-5 (ENSP00000327758) | nkx2.5 (ENSLOCP00000011169) | nkx2.5 (ENSDARP00000016308) | nkx2.5 | NODE_845617; NODE_342086 | *comp63089_c3_seq11* | | | |
| | | | | | *x* | x | x | *x* | | | |
| | | NKX2-6 | NKX2-6 (ENSP00000320089) | nkx2.6* [NKX2-3] (ENSLOCP00000018486) | nkx2.6* [nkx2.7] (ENSDARP00000004716) | nkx2.6 | NODE_141516 | - | | | |
| | | | | | *x* | x | x | *x* | | | |
| | | NKX2-3 | NKX2-3 (ENSP00000342828) | nkx2.3 (ENSLOCP00000011023) | nkx2.3 (ENSDARP00000057093) | x | x | *x* | | | |
| | | | | | *x* | x | x | *x* | | √ x 2 | |
| | Nk5/Hmx | HMX1 | HMX1 (ENSP00000383516) | hmx1 (ENSLOCP00000012688) | hmx1 (ENSDARP00000102248) | hmx1 | NODE_356071 | - | | | |
| | | | | | *x* | x | x | *x* | | | |
| | | HMX2 | HMX2 (ENSP00000341108) | hmx2 (ENSLOCP00000009942) | hmx2 (ENSDARP00000108860) | **hmx2x** | NODE_789955 | - | | | |
| | | | | | *x* | **hmx2y** | NODE_197826 | - | √ | | |
| | | HMX3 | HMX3 (ENSP00000350549) | hmx3* [hmx3a] (ENSLOCP00000009957) | hmx3a (ENSDARP00000095383) | hmx3x | NODE_789955 | - | | | |
| | | | | | hmx3b (ENSDARP00000127346) | hmx3y | NODE_160048 | *comp221531_c0_seq1* | | | √ |
| | | HMX4 | *x* | hmx4 (ENSLOCP00000012684) | hmx4 (ENSDARP00000010904) | hmx4 | NODE_356071 | - | | | |
| | | | | | *X (present in Astyanax and Euteleosts)* | x | x | *x* | | √ | |
| | | NKX6-1 | NKX6-1 (ENSP00000295886) | nkx6.1 (ENSLOCP00000014936) | nkx6.1 (ENSDARP00000035373) | nkx6.1x | NODE_134268 | *comp403295_c0_seq1* | | | |
| | | | | | *x* | x | x | *x* | | | |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Nk6 | NKX6-2 | NKX6-2 (ENSP00000357581) | nkx6.2 (ENSLOCP00000008975) | nkx6.2 (ENSDARP00000106735) | nkx6.2x | NODE_203318; NODE_706108 | *comp125862_co_seq1* | |
| | | | | | *x* | x | x | *x* | |
| | | NKX6-3 | NKX6-3 (ENSP00000429553) | nkx6.3 (ENSLOCP00000018502) | nkx6.3 (ENSDARP00000107903) | nkx6.3x | NODE_42962; NODE_693053 | - | |
| | | | | | *x* | x | x | *x* | |
| | Noto | NOTO | NOTO (ENSP00000381486) | noto (ENSLOCP00000011345) | noto* [flh] (ENSDARP00000022555) | noto | NODE_1337680 | - | |
| | | | | | x | x | x | x | |
| | Tlx | TLX1 | TLX1 (ENSP00000359215) | tlx1 (ENSLOCP00000014302) | tlx1 (ENSDARP00000003103) | **tlx1x** | NODE_1216478 | - | **√** |
| | | | | | *x* | **tlx1y** | NODE_482468 | - | |
| | | TLX2 | TLX2 (ENSP00000233638) | tlx2 (ENSLOCP00000002277) | tlx2 (ENSDARP00000009610) | **tlx2x** | NODE_530309 | - | **√** |
| | | | | | *x* | **tlx2y** | NODE_485392 | - | |
| | | TLX3 | TLX3 (ENSP00000296921) | tlx3* [novel gene] (ENSLOCP00000011819) | tlx3a [tlx3 2 of 2] (ENSDARP00000008694) | tlx3x | NODE_506972; NODE_54042 | *comp47312_c1_seq1* | |
| | | | | | tlx3b (ENSDARP00000009156) | x | x | *x* | **√** |
| | Vax | VAX1 | VAX1 (ENSP00000358207) | vax1 (ENSLOCP00000011783) | vax1 (ENSDARP00000039283) | vax1 | NODE_1302232; NODE_482756 | *comp59942_c0_seq5* | |
| | | | | | x | x | x | *x* | |
| | | VAX2 | VAX2 (ENSP00000234392) | vax2 (ENSLOCP00000001846) | vax2 (ENSDARP00000076061) | vax2 | NODE_110279; NODE_56629; NODE_637893 | *comp213217_c0_seq1* | |
| | | | | | x | x | x | *x* | |
| | Ventx | VENTX | VENTX (ENSP00000357556) | | | | | | |
| | | | *orthology of ventx and ved uncertain* | | ved (ENSDARP00000140685) | ved | NODE_253514 | - | |
| | | *novel in actinopterygii?* | *x?* | vox (ENSLOCP00000008804) | vox (ENSDARP00000019127) | x | x | x | |
| | | | | | vent (ENSDARP00000011087) | x | x | x | |

**Figure 4.1 A:** CDX
**Panel i**: CDX family tree

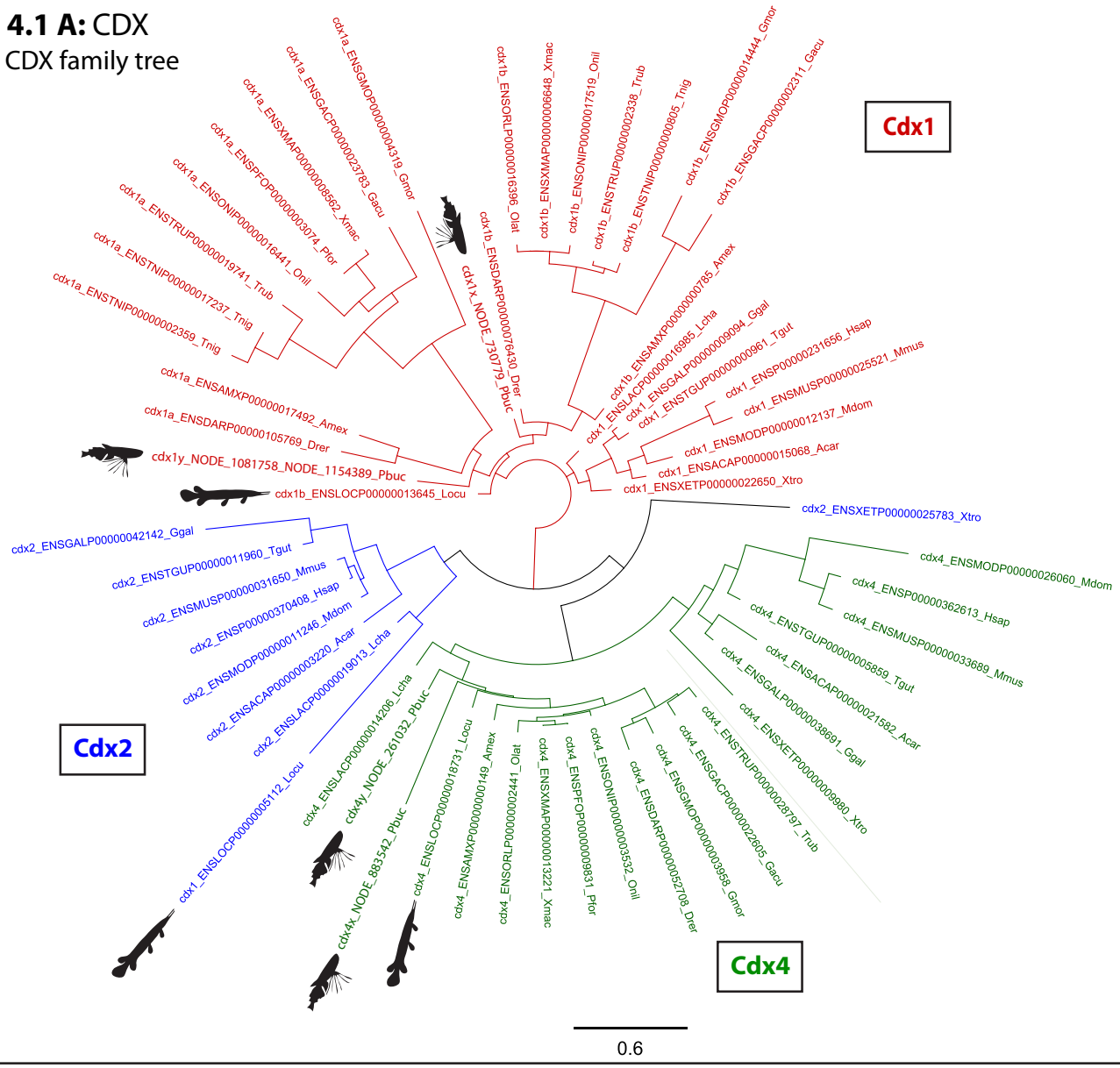



**Figure 4.1 A:** CDX
**Panel ii:** cdx4 subfamily tree

**Figure 4.1 B:** HMX

**Panel i:** HMX family tree

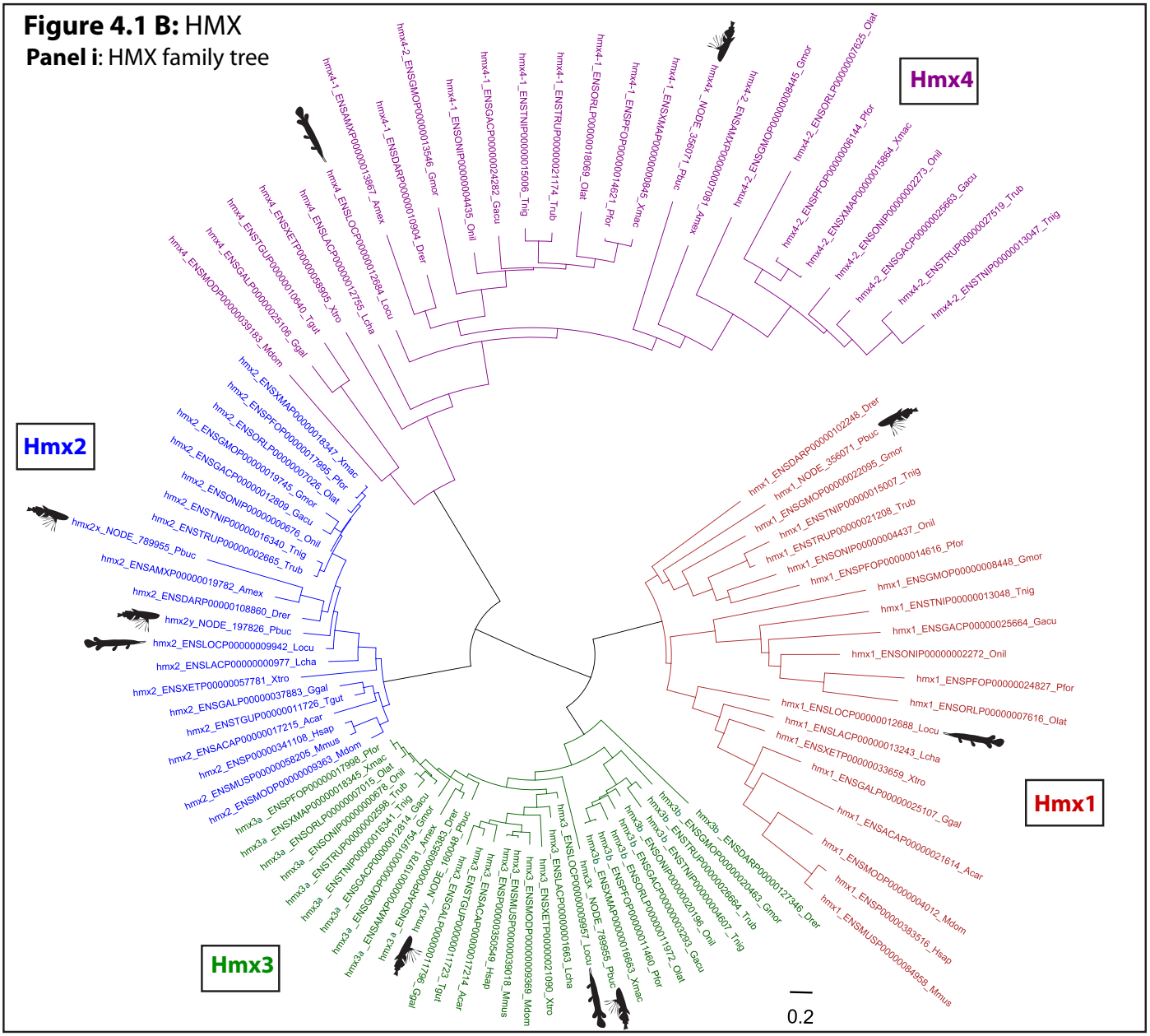



**Figure 4.1 B:** HMX

**Panel ii:** hmx2 subfamily tree

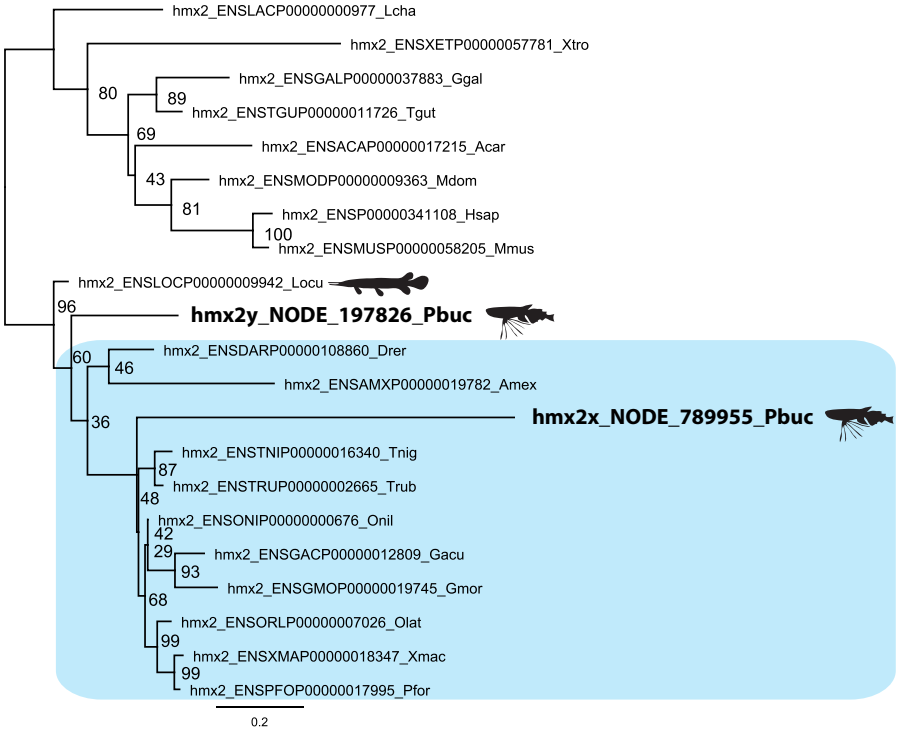


**Figure 4.1 C:** NK2.1
**Panel i:** NK2.1 family tree

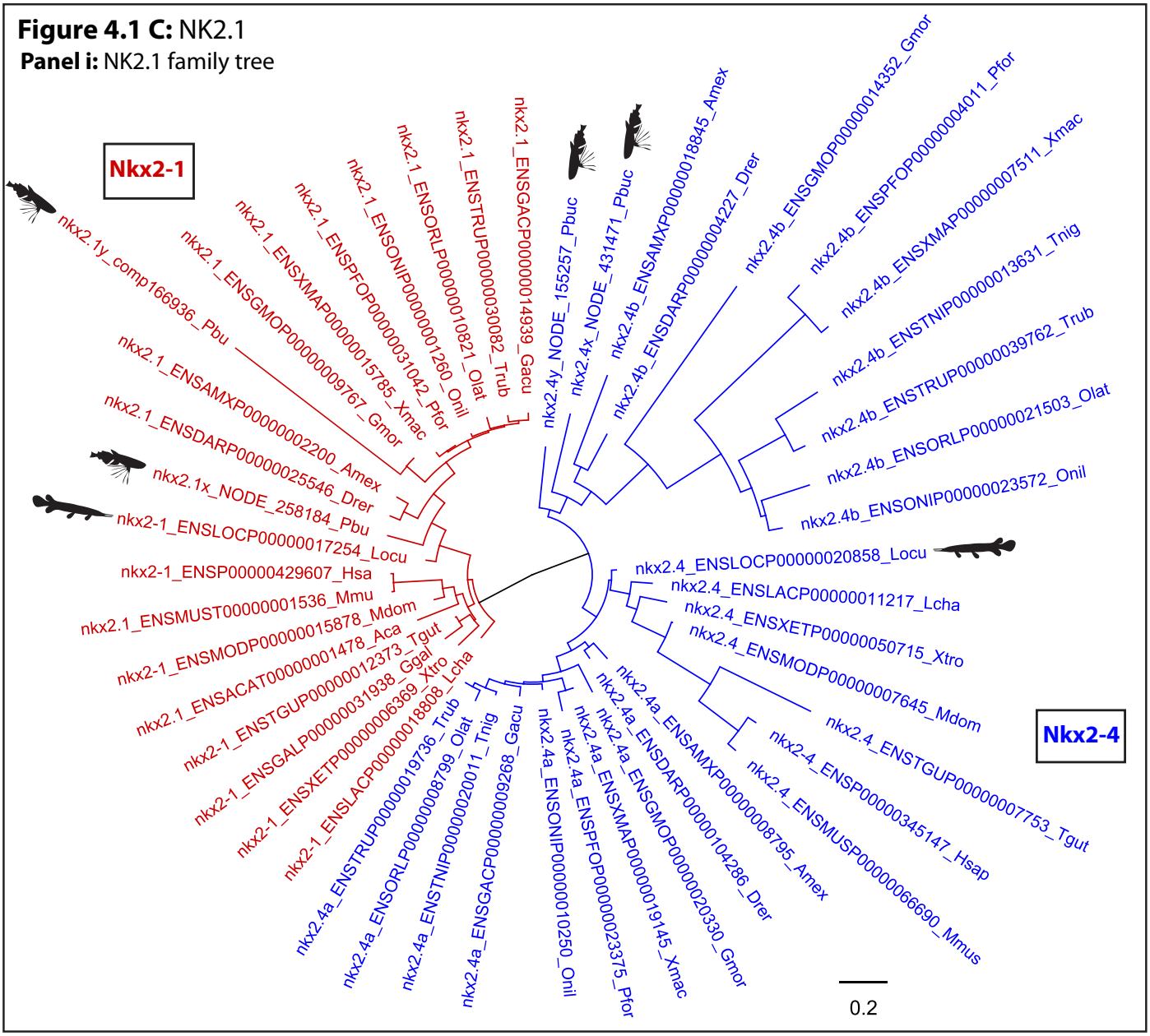



**Figure 4.1 C:** NK2.1
**Panel ii:** nkx2.1subfamily tree

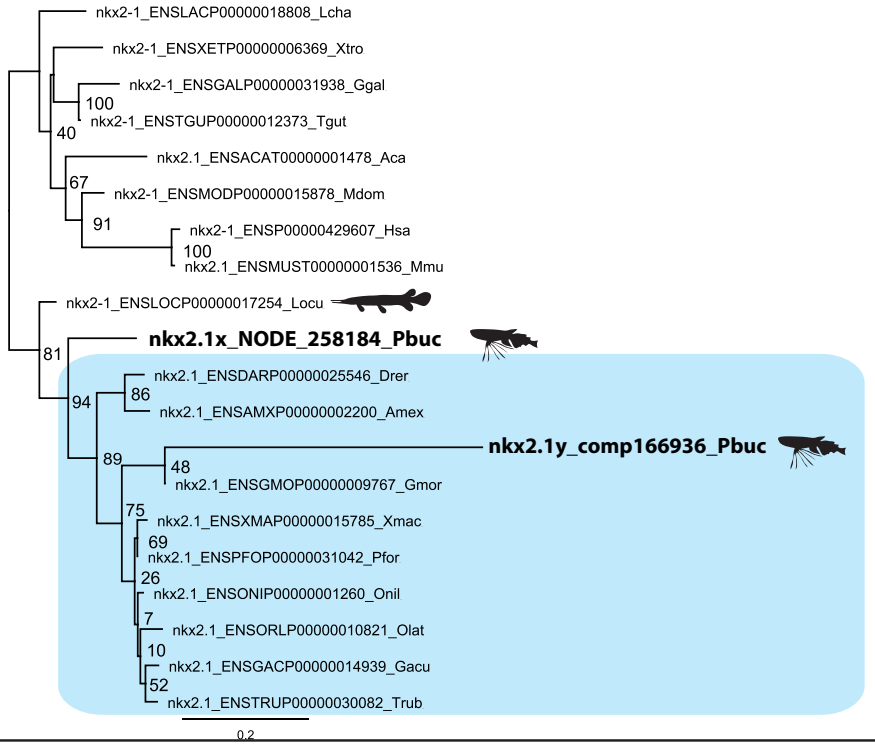


**Figure 4.1 D:** LBX
**Panel i**: LBX family tree

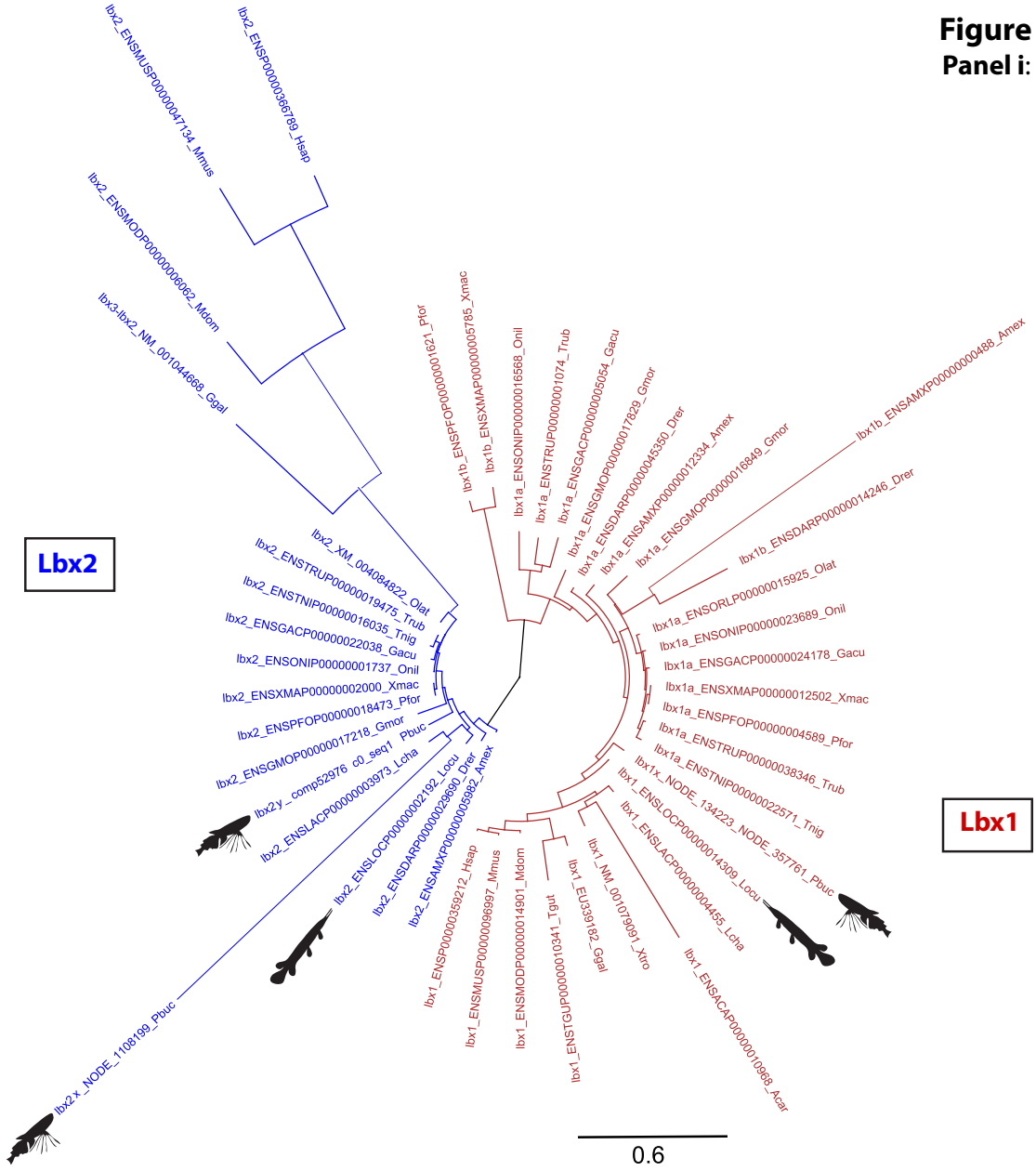



**Figure 4.1 D:** LBX
**Panel ii:** lbx2 subfamily tree

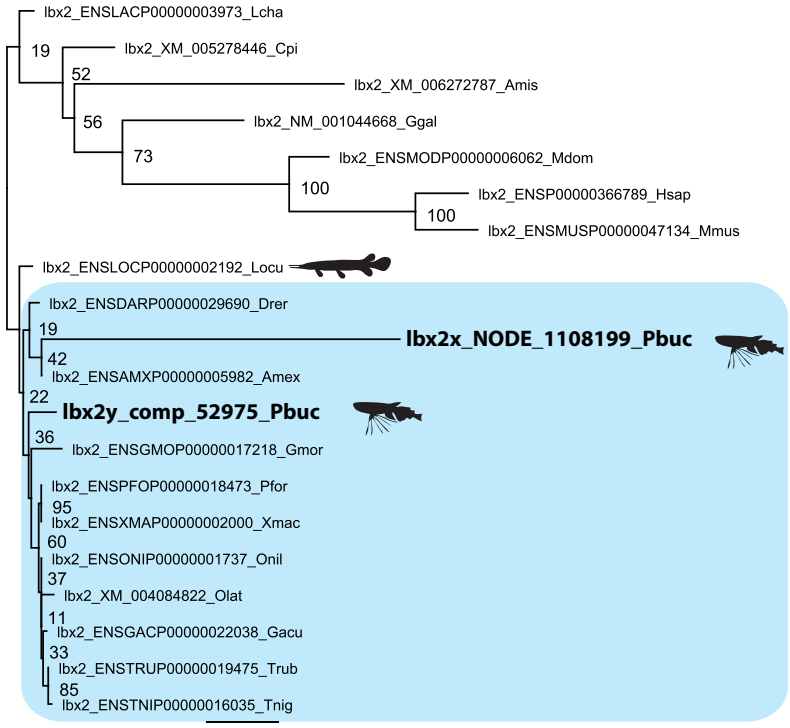


**Figure 4.1 E:** DLX

**Panel i:** DLX1/4/6 family tree

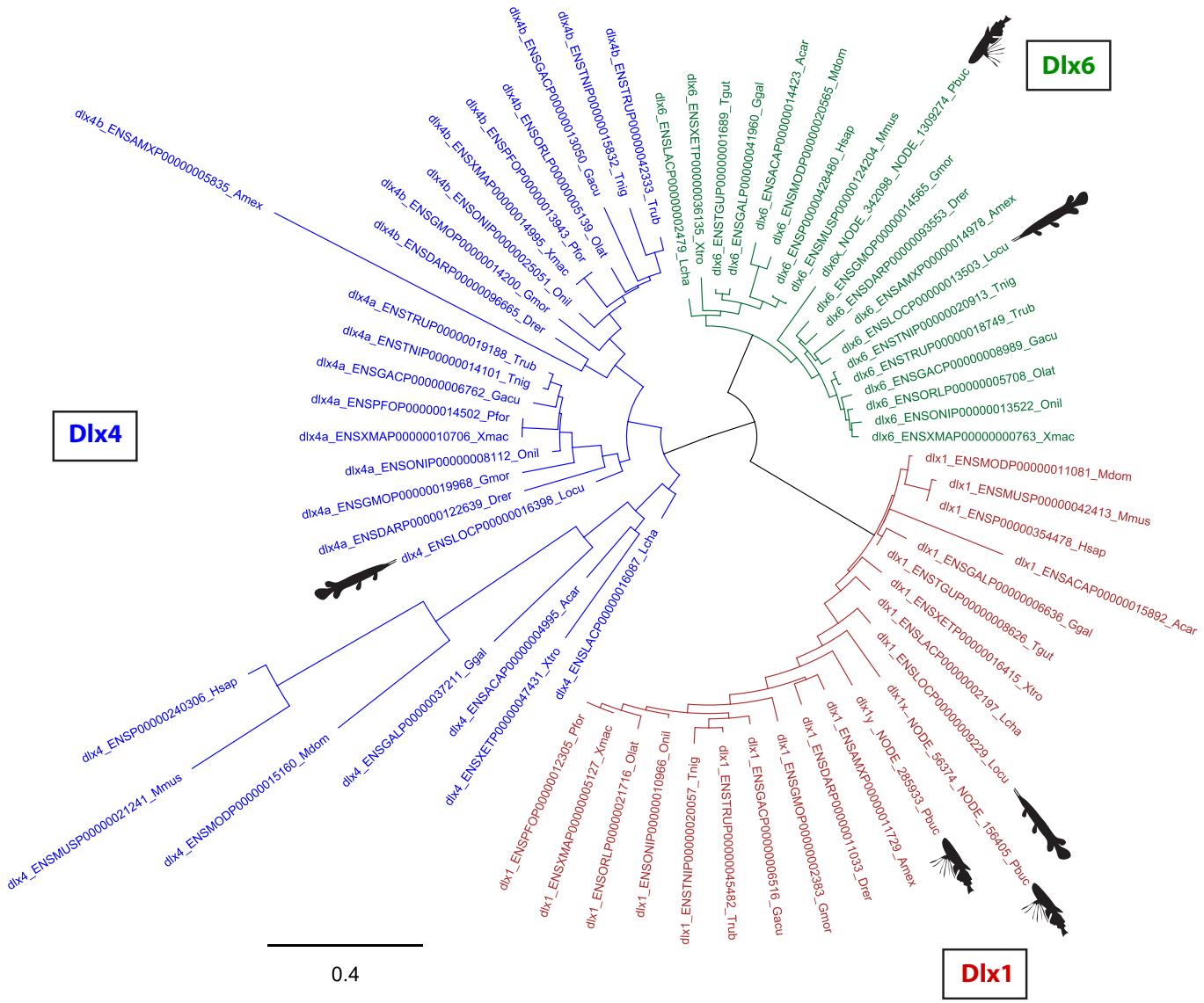



**Figure 4.1 E:** DLX

**Panel ii**: dlx1 subfamily tree

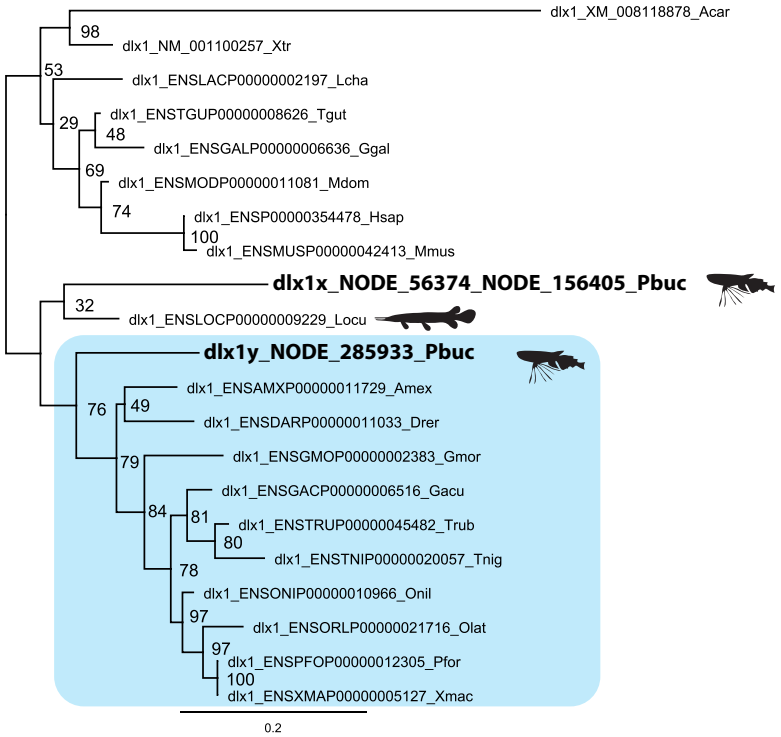


**Figure 4.1 F:** EMX

**Panel i:** EMX family tree

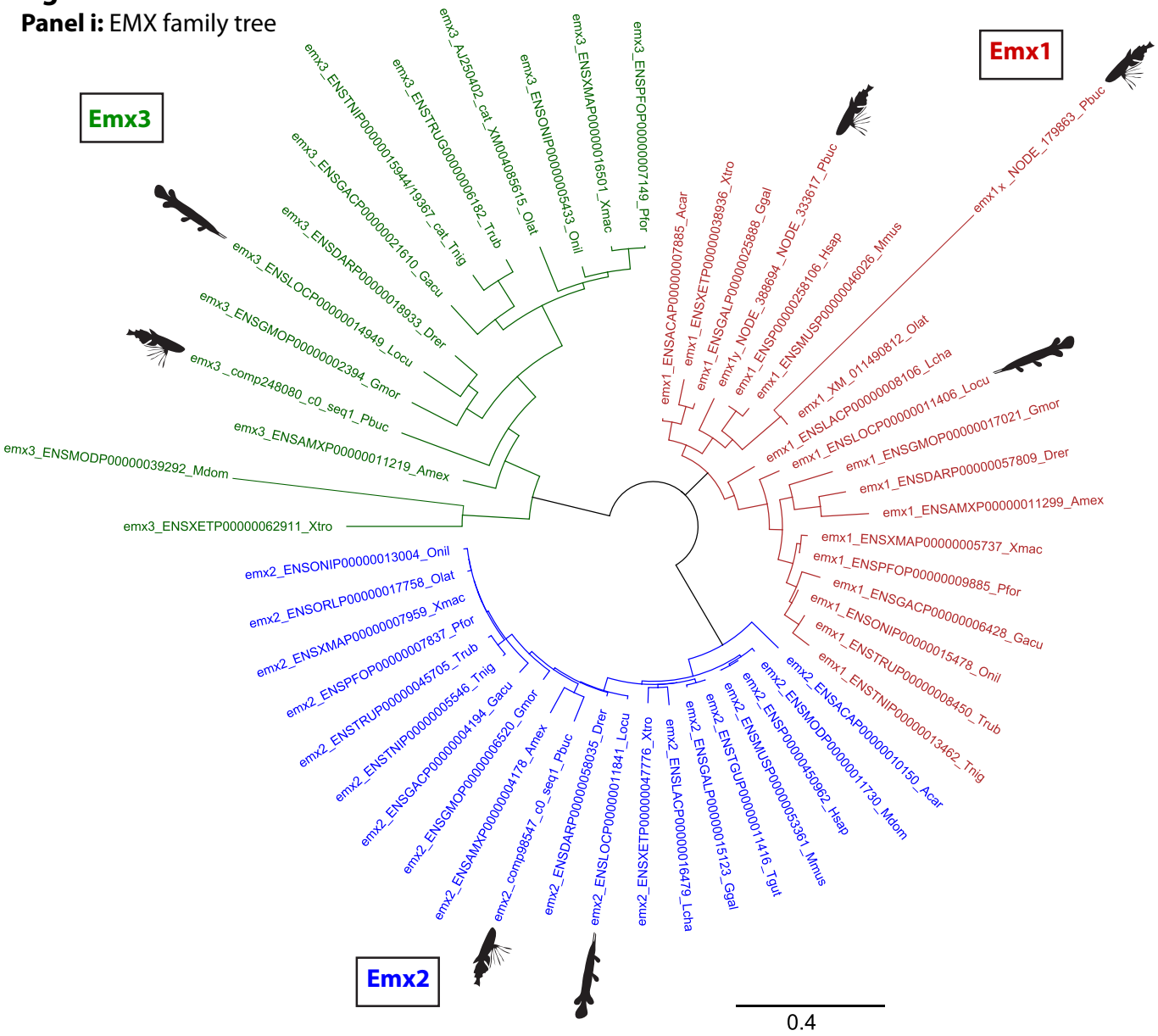



**Figure 4.1 F:** EMX

**Panel ii:** emx1 subfamily tree

**Figure 4.1 G:** BARX
**Panel i:** BARX family tree

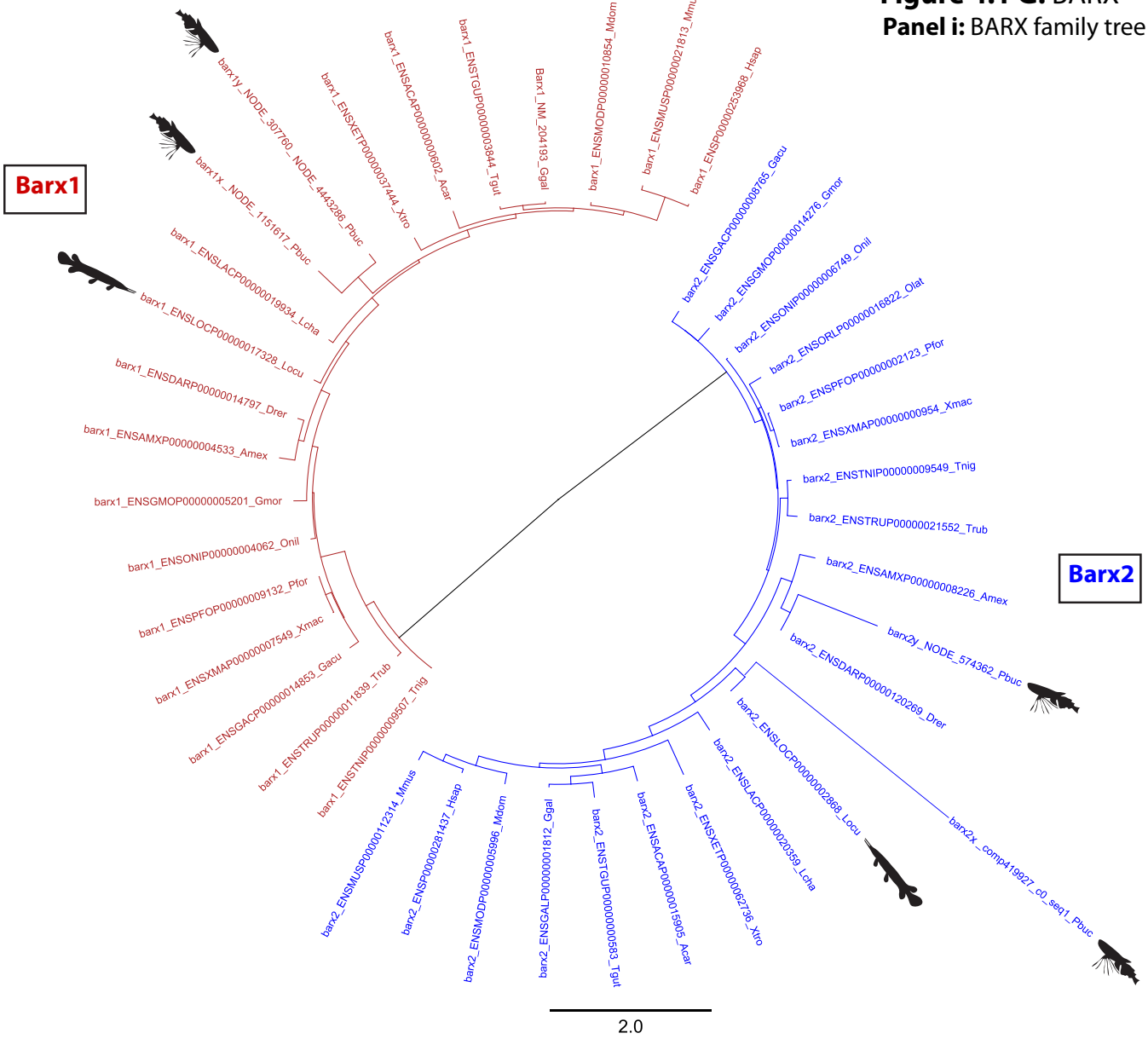
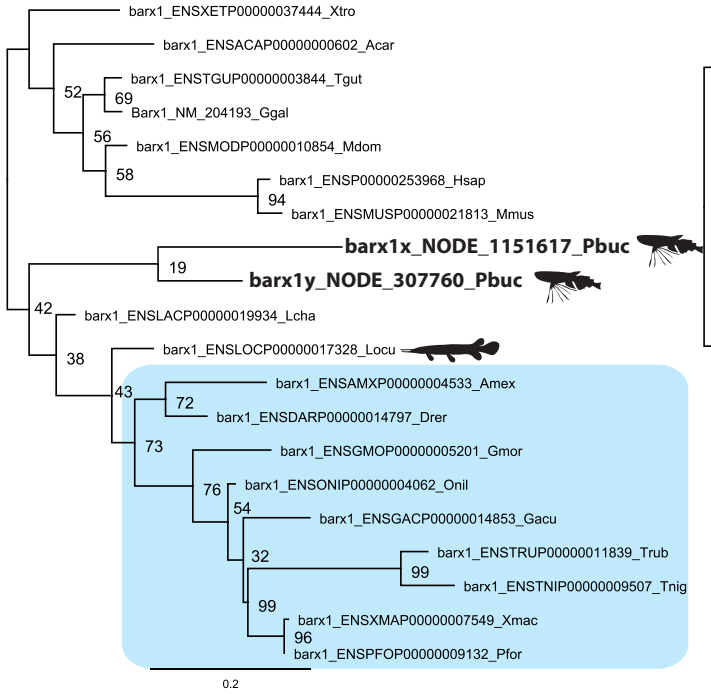



**Figure 4.1 F:** BARX

**Panel ii:** barx1 subfamily tree




**Panel iii**: barx2 subfamily tree

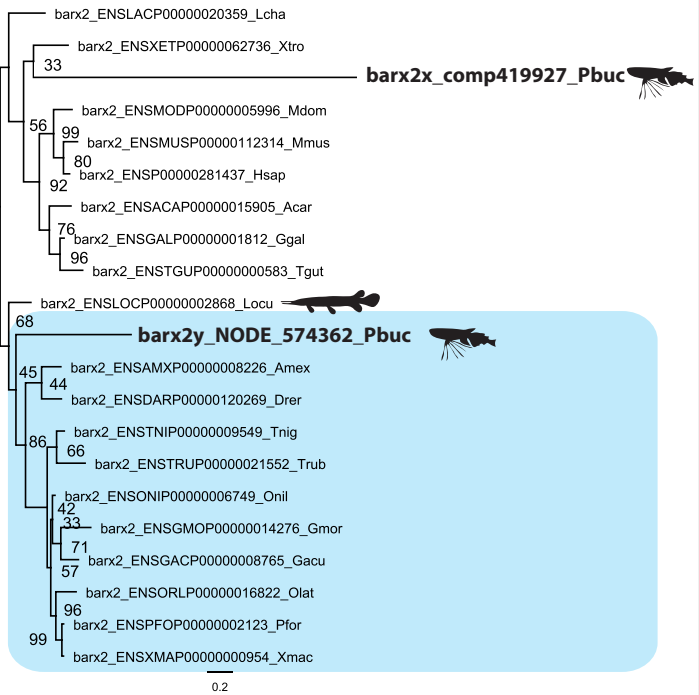
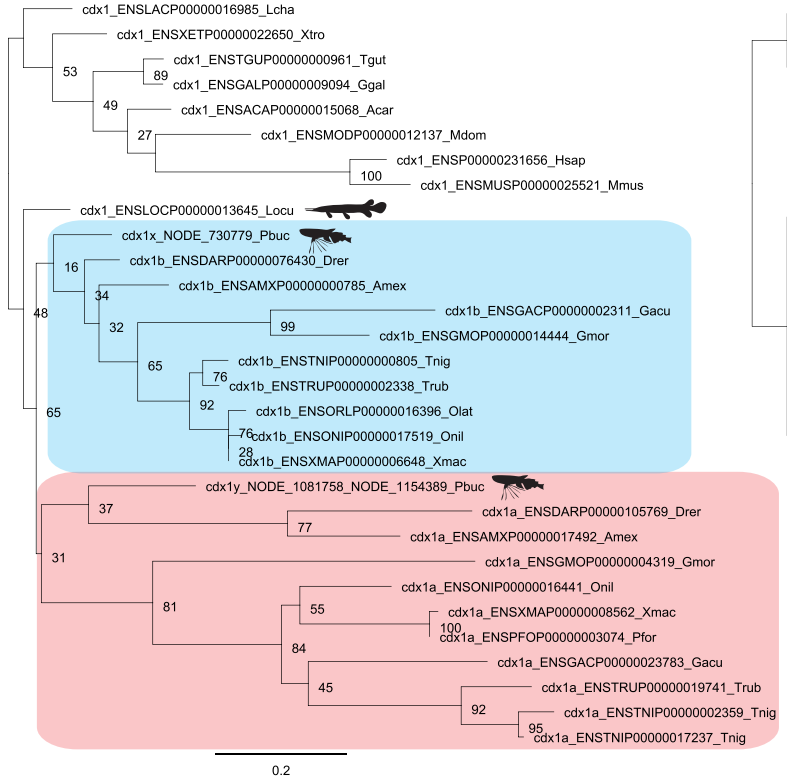
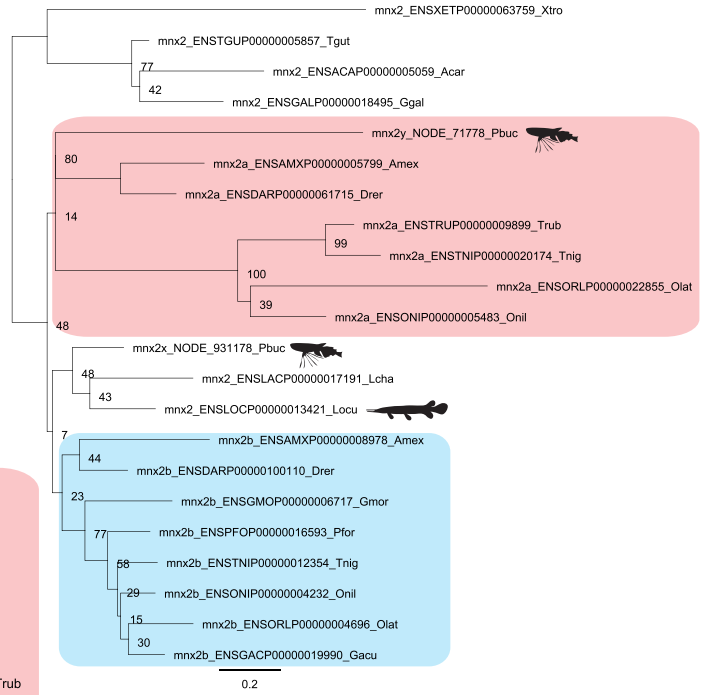


**Figure 4.1.** TGD paralogs retained only in *P. buchholzi* and lost in clupeocephalan teleosts.
ANTP family level trees (panel **i**) and vertebrate subfamily level trees (panel **ii**) are shown for cases where *P. buchholzi* was the only teleost to retain two post-TGD duplicates. (**A**) i. CDX ii. cdx4 (**B**) i. HMX ii. hmx2 (**C**) i. NK2-1 ii. nkx2.1 (**D**) i. LBX ii. lbx2 (**E**) i. DLX1/4/6 subtree of the full DLX family ii. dlx1 (**F**) i. EMX ii. emx1 (**G**) i. BARX ii. barx1 iii. barx2. Trees were built using the protein coding sequence of each novel *P. buchholzi* gene and the protein coding sequences of every member of the gene family which could be extracted from the ENSEMBL database for the following taxa 10 teleost taxa: *Danio rerio* (Drer), *Astyanax mexicanum* (Amex) *Gadus morhua* (Gmor), *Oreochromis niloticus* (Onil), *Xiphophorus maculeatus* (Xmac), *Poecelia formosa* (Pfor), *Oryzias latipes* (Olat), *Gasterosteus acueleatus* (Gacu), *Tetraodon nigroviridis* (Tnig), and *Takifugu rubripes*. One pre-TGD actinopterygian *Lepisosteus oculeatus* (Locu), and 8 sarcopterygians: *Latimeria chalumnae* (Lcha), *Xenopus tropicalis* (Xtro), *Anolis caroliensis* (Acar), *Gallus gallus* (Ggal), T*aenopygia guttata* (Tgut), *Monodelphis domestica* (Mdo), *Mus musculus* (Mmu), and *Homo sapiens* (Hsap). Protein coding sequences were aligned using MAFFT L-INSI, trimmed automatically using the program trimAl, and visually inspected. Gene trees were reconstructed in a maximum-likelihood framework with RAxML and the PROTGAMMALG amino acid substitution model with 1000 bootstrap replicates. Bootstrap support values for the unrooted ANTP family level trees are not shown. Bootstrap support values for each vertebrate subfamily level tree are shown adjacent to each node. The relative position of the novel *P. buchholzi* sequences are denoted by the silhouette ( ) and the pre-TGD outgroup gar is denoted by the silhouette ( ).

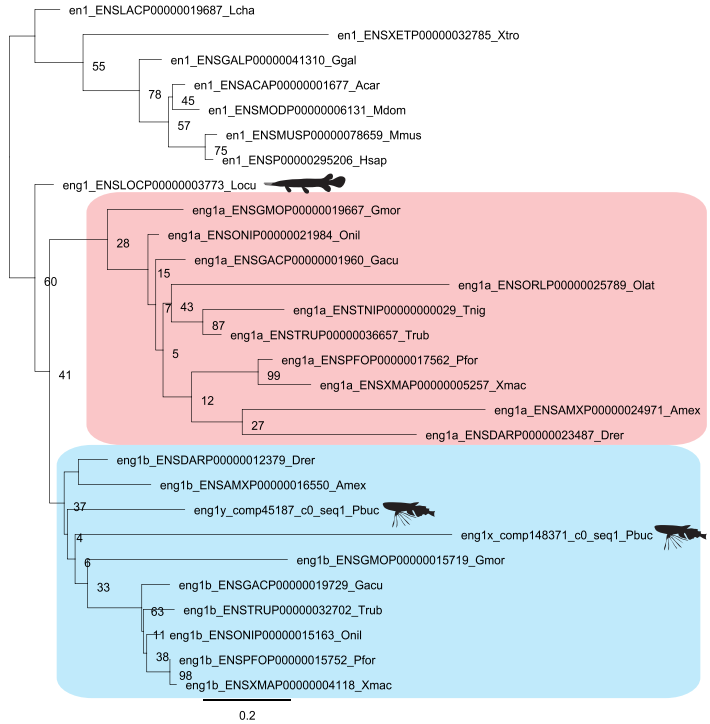
## **A:** cdx1 subfamily tree



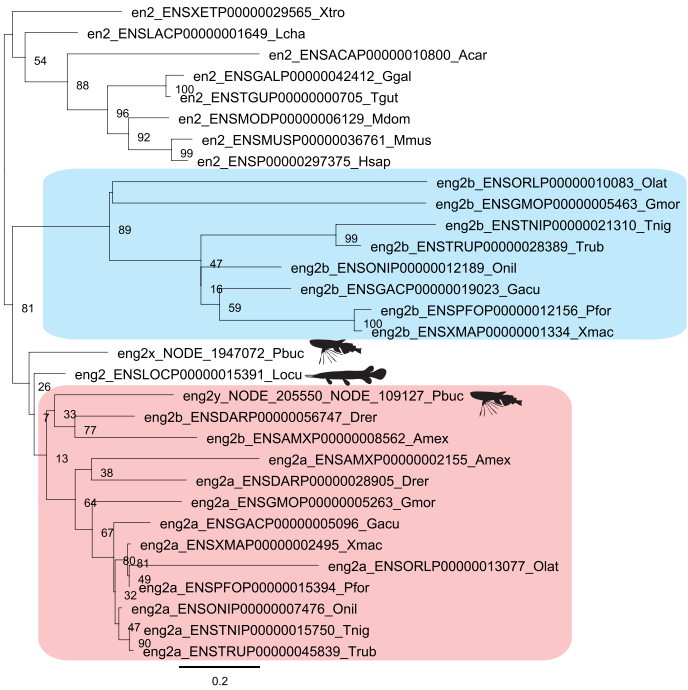

## **B:** mnx2 subfamily tree




## **C:** eng1 subfamily tree

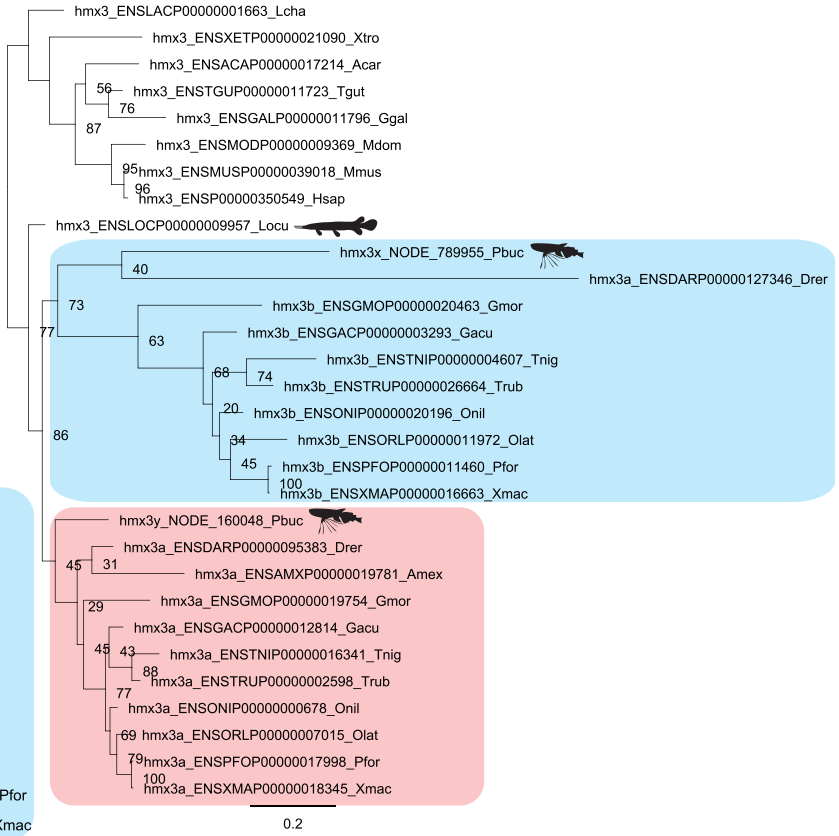



## **D:** eng2 subfamily tree



## **E:** nkx2.4 subfamily tree

## **F:** hmx3 subfamily tree




## **G:** nkx2.2 subfamily tree

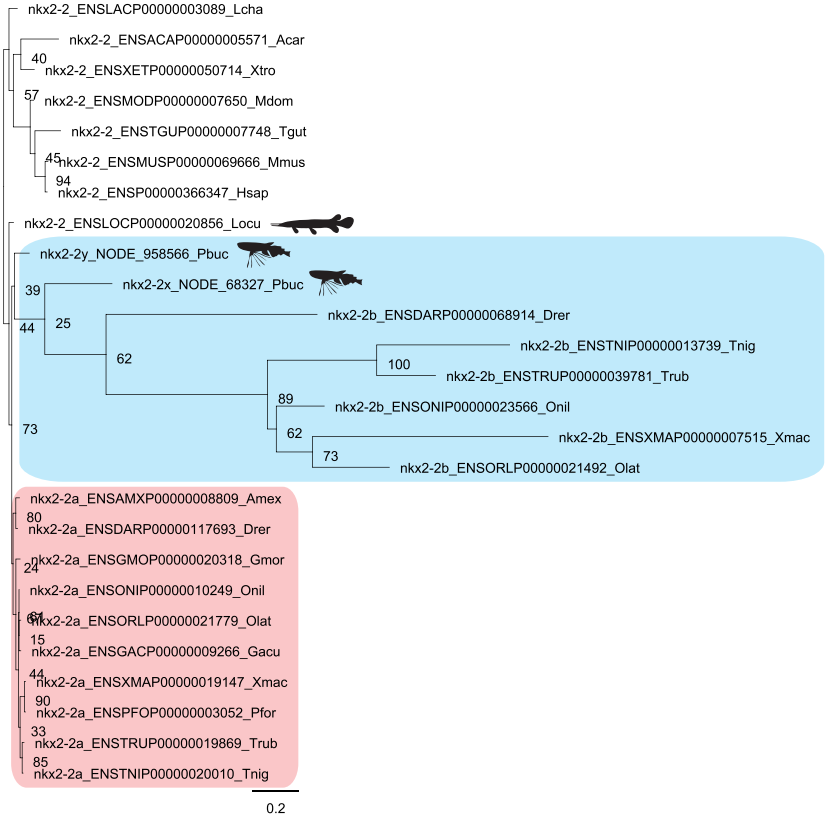
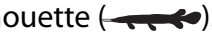


**Figure 4.2.** TGD paralogs retained by both *P. buchholzi* and clupeocephalan teleosts.

Gene genealogies for the cdx1 **(A),** mnx2 **(B)**, eng1 **(C)**, eng2 **(D)** nkx2.4 **(E),** hmx3 **(F),** nkx2.2 **(G)** homeobox gene subfamilies for which I could identify two paralogs retained in both *P. buchholzi* and at least one ostarioclupeomorph and euteleost taxon. Gene trees were reconstructed in a maximum-likelihood framework with RAxML and the PROTGAMMALG amino acid substitution model with 1000 bootstrap replicates. Clupeocephalan post-TGD 'a' and 'b' paralogy groups are highlighted in red and blue repectively, and the relative position of the novel *P. buchholzi* sequences are denoted by the silhouette ( ) and the gar by the silhouette (  ). Only in the cases of cdx1 (**A**) and hmx3 **(F)** did both *P. buchholzi* sequences segregate in a reciprocally monophyletic clade with clupeocephalan 'a' and 'b' paralogy groups although support for these clades overall are quite low.

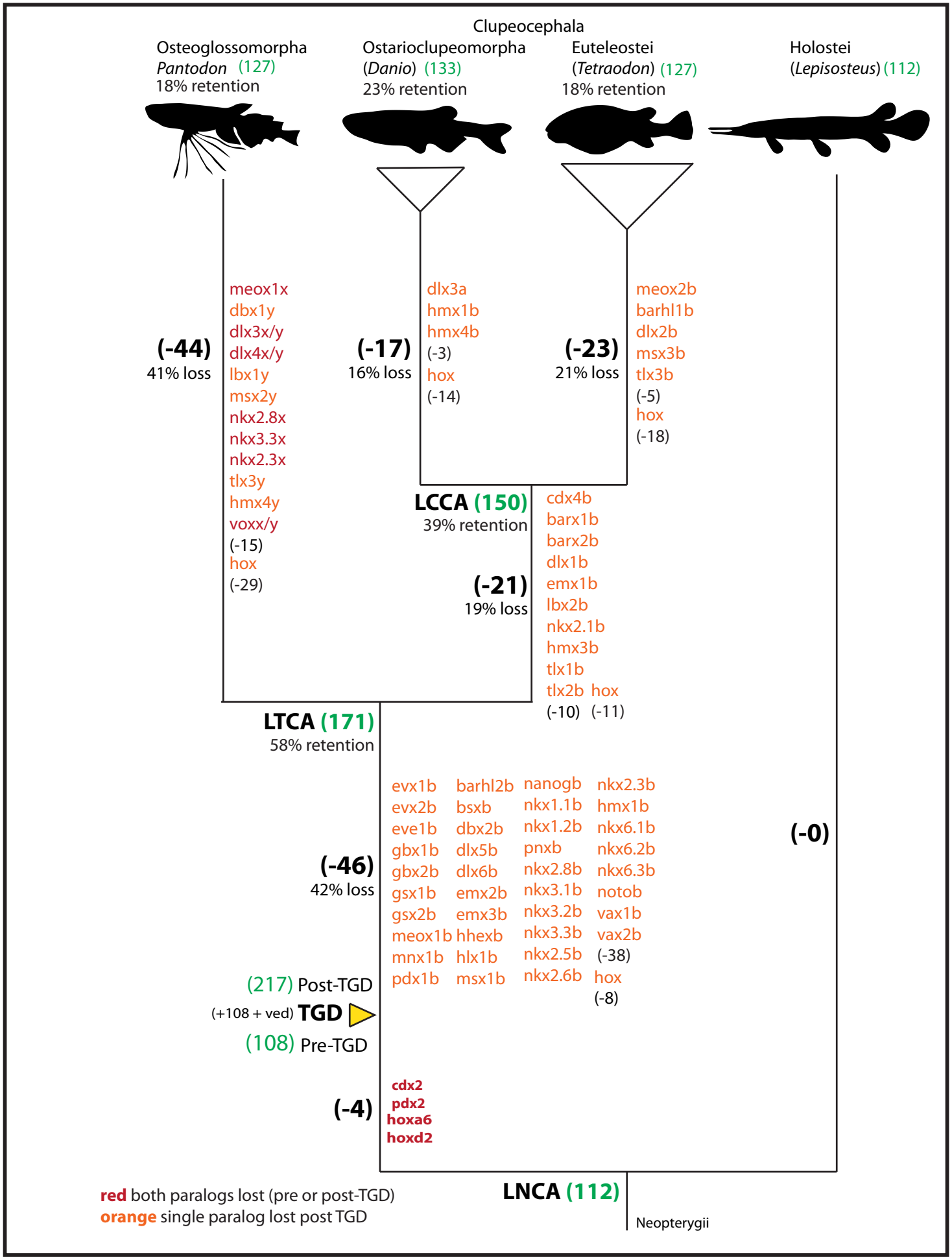

Clupeocephala
Osteoglossomorpha
*Pantodon* (127)
18% retention
Ostarioclupeomorpha
(*Danio*) (133)
23% retention
Euteleostei
(*Tetraodon*) (127)
18% retention
Holostei
(*Lepisosteus*) (112)
(-44)
41% loss
meox1x
dbx1y
dlx3x/y
dlx4x/y
lbx1y
msx2y
nkx2.8x
nkx3.3x
nkx2.3x
tlx3y
hmx4y
voxx/y
(-15)
hox
(-29)
(-17)
16% loss
dlx3a
hmx1b
hmx4b
(-3)
hox
(-14)
(-23)
21% loss
meox2b
barhl1b
dlx2b
msx3b
tlx3b
(-5)
hox
(-18)
LCCA (150)
39% retention
(-21)
19% loss
cdx4b
barx1b
barx2b
dlx1b
emx1b
lbx2b
nkx2.1b
hmx3b
tlx1b
tlx2b
(-10)
hox
(-11)
LTCA (171)
58% retention
(-46)
42% loss
evx1b
evx2b
eve1b
gbx1b
gbx2b
gsx1b
gsx2b
meox1b
mnx1b
pdx1b
barhl2b
bsxb
dbx2b
dlx5b
dlx6b
emx2b
emx3b
hhexb
hlx1b
msx1b
nanogb
nkx1.1b
nkx1.2b
pnxb
nkx2.8b
nkx3.1b
nkx3.2b
nkx3.3b
nkx2.5b
nkx2.6b
nkx2.3b
hmx1b
nkx6.1b
nkx6.2b
nkx6.3b
notob
vax1b
vax2b
(-38)
hox
(-8)
(-0)
(217) Post-TGD
(+108 + ved) TGD
(108) Pre-TGD
(-4)
cdx2
pdx2
hoxa6
hoxd2
LNCA (112)
Neopterygii
red both paralogs lost (pre or post-TGD)
orange single paralog lost post TGD

**Figure 4.3**. Evolution of ANTP class homeobox genes in teleosts after the TGD.
The ANTP complement at ancestral nodes LNCA (Last Neopterygian Common Ancestor), LTCA (Last Teleost Common Ancestor), and LCCA (Last Clupeocephalan Common Ancestor) were reconstructed by comparative genome analysis. At each ancestral node, and in terminal nodes for *P. buchholzi*, *D. rerio, T. nigroviridis*, and *L. oculeatus* the total number of genes inferred to have been present is shown in green. Along each branch of the tree, both the total number, and names of genes which experienced losses of paralogs following the TGD are mapped in red where both paralogs or a whole subfamily are lost, and in orange where only one paralog has been lost. Separate hox family genes are not shown individually for simplicity, but are counted separately in the totals.

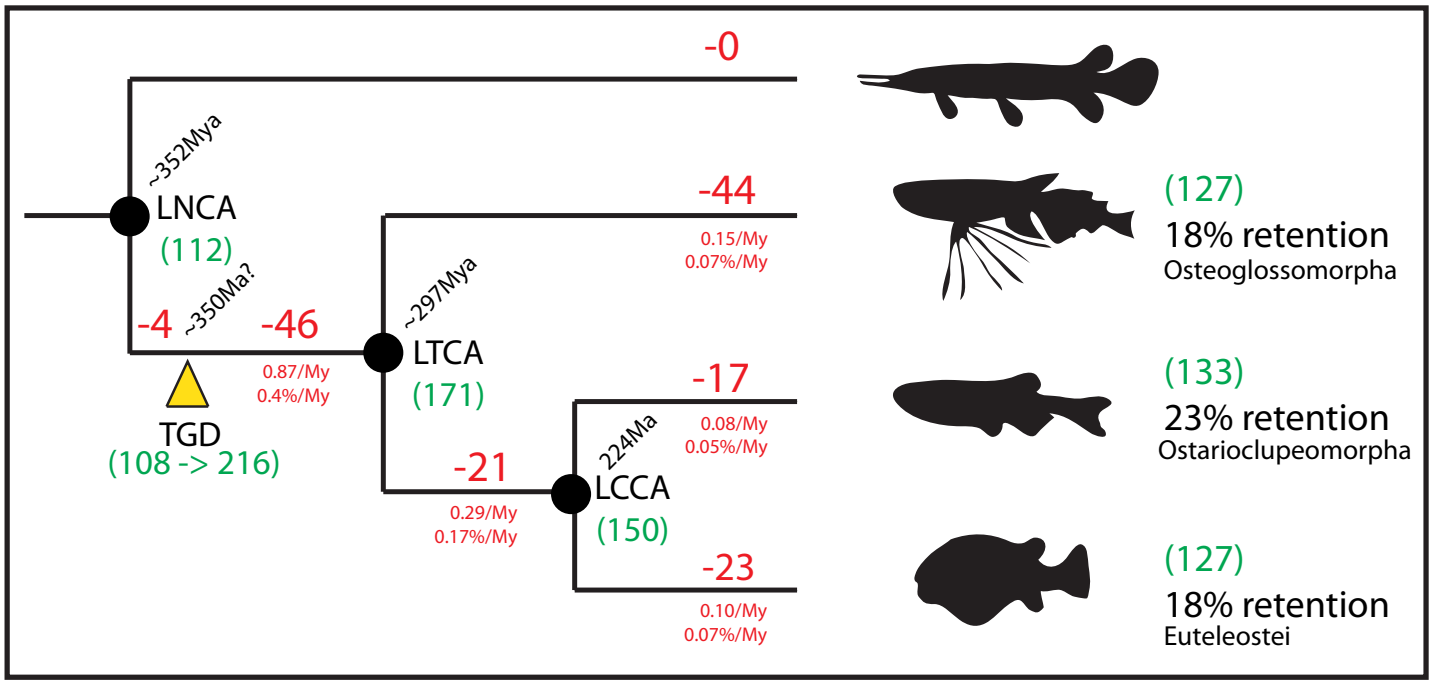



**Figure 4.4.** Relative rates of ANTP class homeobox gene paralog loss along key branches following the TGD.

In red below each branch, both the absolute rate in genes per million years (My) and relative rate as % of TGD-generated ANTP-class duplicates per My of loss between the teleost genome duplication (TGD) and the last teleost common ancestor (LTCA), between the LTCA and the last clupeocephalan common ancestor (LCCA), and along the branch leading to zebrafish and *T. nigroviridis* as representatives of the Ostarioclupeomorpha and Euteleostei respectively. In red above each branch the total number of ANTP gene losses in each lineage is shown. In green at each node the number of reconstructed ANTP genes inferred to be present is shown, and the total number of ANTP genes in each lineage investigated is shown adjacent to their respective silhouettes. The mean of the range of divergence times estimated from the mitogenome clock analyses of Hurley et al. 2007 was used to make estimates of gene loss rates to avoid potential complications due to potential independent diploidization of nuclear gene divergence date estimates such as those used in Near et al. 2012, while acknowledging this is a very rough estimation of absolute dates for illustrative purposes only.

# Chapter 5

## A bioinformatic survey of Hox genes in the draft genome sequences of three osteoglossomorph teleosts reveals extensive divergent resolution of Hox clusters following the teleost genome duplication


## Abstract

Teleost fishes have all shared an ancient genome duplication (TGD) resulting in the duplication of all genes. Osteoglossomorpha, or 'bony-tongues', are an ancient teleost subdivision which diverged from the other major teleost subdivisions (Elopomorpha, Clupeocephala) close to the root of the teleost tree. Divergent resolution of TGD paralogs is a potential driving force for speciation, but although there are >27,000 species of extant teleosts, there are significant differences in the number of extant species in osteoglossomorphs, elopomorphs, and clupeocephalans. The degree to which divergent resolution of TGD paralogs has occurred between taxa within the Osteoglossomorpha has never previously been measured. I have constructed *de novo* draft genome assemblies of three osteoglossomorph taxa representing three different families: *Osteoglossum bicirrhosum* (Osteoglossidae) *Chitala ornata* (Notopteridae), and *Gnathonemus petersii* (Mormyridae) and assayed their Hox gene contents. Hox genes are developmentally important transcription factors which occur on colinear clusters in the genome, and have previously been shown to have experienced divergent resolution in other teleosts. Each osteoglossomorph taxon I investigated retained more Hox genes than any clupeocephalan taxa (excluding those which underwent supplementary rounds of

WGD following the TGD), but less than have been found in eels (Elopomorpha). My results indicate that the last osteoglossomorph common ancestor was extremely replete in Hox genes (retaining 70 genes) and that divergent resolution at both the level of individual gene losses, and whole cluster losses, has been rampant in the osteoglossomorph lineage. Furthermore, I find that both TGD paralogs of the hoxb13 and hoxa5 subfamilies can be found in at least one osteoglossomorph genome, but not in the genomes of any other teleosts, suggesting they may be osteoglossomorph-specific retentions. These data increases the number of Hox genes inferred to be present in the genome of the last teleost common ancestor to 75, and suggesting that the majority of Hox gene losses occurred only after the deep divergence of the three major teleost subdivisions rather than in the immediate wake of the TGD.

## Introduction

It is currently unknown to what extent divergent resolution of TGD paralogs has occurred either between families of osteoglossomorphs or between osteoglossomorphs, elopomorphs, and clupeocephalans. I set out to examine whether there is any evidence for divergent resolution of TGD-generated gene duplicates between 5 families of osteoglossomorph teleosts and to examine the extent of divergent resolution of TGD paralogs which occurred both between these deeply diverging families within Osteoglossomorpha as well as between major teleost subdivisions Osteoglossomorpha and Clupeocephala. If divergent resolution of TGD paralogs between osteoglossomorph families has occurred to a different extent than between clupeocephalan families, it could have led to different effects on the rates of speciation and subsequent radiations within each of these major subfamilies.

Interestingly, divergent resolution and reciprocal loss has occurred not only at the level of individual genes, but also at the level of entire Hox clusters with the loss of the hoxdb cluster in ostariophysiians (Woltering and Durston, 2006), and again independently in salmonids (Mungpakdee et al., 2008), while the hoxcb cluster has been lost from percomorphs including pufferfishes, cichlids and sticklebacks (Amores et al., 2004; Kurosawa et al., 2006; Hoegg et al., 2007; Thomas-Chollier and Ledent, 2008). The first study to investigate the Hox clusters of earlier diverging teleost subdivisions was a PCR study of the osteoglossomorph *Hiodon alosoides* which concluded that the 41 Hox gene fragments they amplified could be assigned to 8 Hox clusters, suggesting that *H. alosoides* might have suffered no Hox cluster losses following the TGD (Chambers et al., 2009). Subsequently, the Hox clusters of two

anguilliform elopomorphs, the European eel (*A. anguilla*) and the Japanese eel (*A. japonica*) were found to have retained more genes (70 and 71 respectively) than any other teleost, and are the only teleosts with completely sequenced Hox clusters which were confirmed to have retained all 8 post-TGD clusters (Guo et al., 2010; Henkel et al., 2012a). From these data it seemed possible that the basal teleost subdivisions experienced less total gene loss and therefore less potential instances of divergent resolution than clupeocephalans, and possibly that the loss of complete Hox clusters was a clupeocephalan-specific trait rather than a predictable outcome following the TGD. However, I have recently shown that the monotypic osteoglossomorph in the family Pantodontidae, *P. buchholzi*, has a significantly different Hox cluster organization compared with other teleosts and has reduced the number of Hox clusters present in its genome to only 5, making it the teleost with the fewest Hox clusters described to date (Martin and Holland, 2014). Interestingly, despite having fewer Hox clusters, both ancient TGD paralogs of 4 genes in the duplicated hoxb clusters (*hoxb2x/y, hoxb4x/y, hoxb9x/y, hoxb13x/y*) were discovered in *P. buchholzi* which are not present in the genomes of clupeocephalan teleosts, one of which (*hoxb13x/y*) is also absent from the otherwise Hox-replete genomes of eels. This demonstrated that divergent resolution both at the level of individual genes and entire clusters has occurred between the three major teleost subdivisions and hints that osteoglossomorph Hox clusters might be significantly different from both elopomorph and clupeocephalan Hox clusters. Therefore, contrary to the initial idea that basal teleosts experienced less gene and cluster loss than clupeocephalans, I found that at least one osteoglossomorph demonstrated a degree of gene loss divergent resolution

on a scale greater than seen in other teleosts. However, *P. buchholzi* represents just one of the six families of osteoglossomorphs and it is currently unknown whether it is representative of the situation in osteoglossomorphs more broadly or whether it represents a unique departure. Pantodontidae is a deeply branching family within Osteoglossomorpha and experienced a long period of independent evolution following its divergence from other osteoglossomorphs ~200 Mya. Moreover, two allopatric populations of *P. buchholzi* have been described which were estimated to have diverged ~57 Mya and yet exhibited one of the most extreme cases of morphological stasis on a geological timescale yet observed in any vertebrate (Lavoué et al., 2010). It has also been noted that the five remaining Hox clusters in *P. buchholzi* show an exclusion of transposable elements to a greater degree than other teleost fish and therefore are likely under extremely strong selective constraint (Chapter 3).

A longstanding hypothesis has been that WGD can accelerate speciation through the reciprocal loss of different paralogs in different populations, leading to hybrid dysfunction, or sterility (Lynch and Force, 2000; Presgraves, 2010). In teleosts, it has been proposed that reciprocal gene loss following the TGD ~350 Mya could have played an important role in the radiation of this lineage which today constitutes over half of all vertebrate biodiversity (Taylor et al., 2001; Hoegg et al., 2004; Jaillon et al., 2004). It is thought that divergent resolution can occur both at the level of reciprocal non-functionalization (gene-loss) between lineages or in the reciprocal subfunctionalization (partitioning of parental gene functions between two daughter genes) for example through the complementary degeneration of *cis*-

regulatory elements in accordance with the duplication-degeneration-complementation (DDC) model (Force et al., 1999; Postlethwait et al., 2004). Previous studies have investigated the degree of divergent resolution which has occurred between clupeocephans, the largest teleost subdivision consisting of ostarioclupeomorphs (zebrafish, carp, cavefish etc.) and euteleosts (medaka, cichlids, pufferfishes etc.), and have shown that there was indeed abundant genome-wide reciprocal gene loss between two of the most speciose superorders, the Ostariophysii and the Percomorpha (Sémon and Wolfe, 2007; Garcia de la Serrana et al., 2014). There is also evidence of ongoing reciprocal gene loss of Hox genes within the percomorph family Cichlidae (Thomas-Chollier and Ledent, 2008). Importantly, while the TGD is unambiguously a synapomorphy of all teleosts (Taylor et al., 2003; Henkel et al., 2012a; Martin and Holland, 2014), species richness is not evenly spread across the teleost tree and together the accelerated speciation rate in the Ostariophysii (~9,000 species) and Percomorpha (~16,000 species) accounts for ~88% of extant teleost biodiversity (Santini et al., 2009). In addition to the Clupeocephala, two other major teleost subdivisions are recognized: the Elopomorpha and the Osteoglossomorpha (Nelson, 2006). Systematists do not agree on whether osteoglossomorphs, elopomorphs, or a clade containing both are the sister lineage to clupeocephalans, but the respective monophyly of each major subdivision is consistently well supported (Le et al., 1993; Inoue et al., 2001, 2004; Hurley et al., 2007; Broughton, 2010; Broughton et al., 2013). Osteoglossomorphs are also the most species-poor subdivision of teleosts, with only ~221 extant members occurring within 6 families: Hiodontidae (2 species), Pantodontidae (1

species), Osteoglossidae (8 species), Notopteridae (8 species), Gynmarchidae (1 species) and Mormyridae (~201 species) (Nelson, 2006).

I have generated low coverage *de novo* draft genome assemblies of three osteoglossomorphs teleosts from three different families, *O. bicirrhosum* (Osteoglossidae), *C. ornata* (Notopteridae), and *G. petersii* (Mormyridae) and conducted a reciprocal-blast based bioinformatic survey of their Hox gene complements. Together with the completely sequenced Hox gene clusters of *P. buchholzi* (Pantodontidae) (Chapter 2, Martin & Holland, 2014) and data from a PCR screen of Hox genes in *H. alosoides* (Hiodontidae) (Chambers et al., 2009) I can for the first time evaluate the extent of divergent resolution of the Hox clusters which has occurred between taxa spanning osteoglossomorph diversity. Because of conserved collinearity of Hox genes within clusters in the vertebrate genome, and their long history of serving as a proxies of the duplication status of vertebrate genomes generally, and their short highly conserved Homeodomain-encoding homeobox sequences which are easily recovered with bioinformatic search tools, they will serve as a useful study system to begin the examination of reciprocal gene loss and divergent resolution in the Osteoglossomorpha.

## Materials and Methods

### Animal husbandry

Juvenile unsexed specimens of each of *C. ornata, G. petersii*, and *O. bicirrhosum* were obtained from the online fish retailer Fareham Tropicals (London, UK) and maintained in freshwater aquaria at the University of Oxford as described in Chapter 2. Animals were sacrificed according to Home Office approved Schedule 1 methods using an overdose of MS222 (400mg/L) followed by decapitation.

### Genomic DNA preparation

Genomic DNA was prepared from 25-50mg of axial skeletal muscle and skin tissue from a single individual of undetermined sex of each species using a combined proteinase K and PCIA extraction protocol as detailed in Chapter 2.

### Genome sequencing and assembly

For each of the three species, high quality genomic DNA was extracted and used to produce a 100bp paired-end library with an average insert size of 180bp for sequencing according to the Illumina HiSeq2000 protocol at the Wellcome Trust Center for Human Genomics, Oxford. These libraries were individually tagged and then pooled in equimolar amounts and sequenced on a single lane of the HiSeq2000 platform. Quality control was carried out as described for the *P. buchholzi* genome assembly described in Chapter 3. Using the short-read assembler velvet (Zerbino and Birney, 2008) and a range of kmers (27, 37, 47, 57) multiple *de novo* assemblies were produced. Assembly metrics and protein coding representation assessed with

the software CEGMA (Parra et al. 2007) as described in Chapter 3 were calculated and the metrics for the best assembly of each species is presented in Table 5.1. Genome size estimates were carried out using jellyfish (Marçais and Kingsford, 2011) with a kmer size of 27 as described in Chapter 3.

**Identification of Hox genes in osteoglossomorph genome assemblies**

The 43 Hox gene sequences of *L. oculeatus* from the LepOcu1 assembly version, manually annotated in Braasch et al., 2016 (appendix 2), were used as a query input to the tblastn algorithm (Altschul et al., 1990) and used to search databases comprising the genome assemblies of *O. bicirrhosum*, *C. ornata*, and *G. petersii* respectively. Using an E-value cut-off of $10^{-6}$, 363, 312, and 321 putative Hox-containing contigs were extracted from the genome assemblies of *O. bicirrhosum*, *C .ornata*, and *G. petersii* respectively. In order to sort out contigs which contained Hox genes from the false positive hits against other proteins with high sequence similarity (such as non-hox ANTP class genes) these putative hox-containing contigs were reciprocally used to query the NCBI nr database (http://blast.ncbi.nlm.nih.gov/Blast.cgi) using the software BLAST2GO using the blastx algorithm and an E-value cut-off of $10^{-10}$ (Götz et al., 2008). Candidate contigs with top blastx hits to other vertebrate Hox protein sequences in the nr database were retained and those which had a top-hit to a different non-hox sequence in the database were discarded. Those which were retained were subsequently used once more to query the predicted proteome of the LepOcu1 *L. oculeatus* assembly on ENSEMBL and the top hit to gar was compared with the top

hit to nr to determine the level of homology assignment of the putative osteoglossomorph hox contigs.

**Ancestral state reconstructions**

The ancestral hox gene complements of each node and therefore the hox gene losses along each branch of the cladogram were reconstructed manually by inferring that the ancestor must have possessed all of the genes present in any of the extant lineages branching from that node. The LOCA was inferred by using the novel predictions described here from *O. biccirhosum*, *G. petersii*, *C. ornata* here in addition to the previously characterized hox clusters of *P. buchholzi* (Martin and Holland, 2014, chapter 2) and the PCR survey in *H. alosoides* (Chambers et al., 2009). The LECA was inferred using the annotated hox clusters of *A. anguilla* (Henkel et al., 2012a) and *A. japonica* (Guo et al., 2010). The LCCA was reconstructed using the Hox genes annotated in the assemblies of the clupeocephalans: *P. formosa* Poecilia_formosa-5.1.2*, X. maculatus* Xipmac4.4.2, *T. nigroviridis* TETRAODON 8.0, *O. niloticus* Orenil1.0, *O. latipes* HdrR, *T. rubripes* FUGU 4.0, *G. morhua* gadMor1, *G. aculeatus* BROAD S1, *A. mexicanus* AstMex102, and *D. rerio* GRCz9 used previously for phylogenetic analyses (chapter 4). Because several additional clupeocephalan genomes have been sequenced a rapid survey of the ENSEMBL gene trees for each hox gene was done in clupeocephalans to verify no additional Hox genes absent from the clupeocephalan teleosts above was automatically annotated in these novel genomes, but a manual check of each new genome to confirm the ENSEMBL pipeline had correctly annotated each hox gene was not carried out. For the teleost fish, the neopterygian gar *L. oculeatus* was used as an

outgroup to infer losses which may have occurred either prior to the TGD or between the TGD and the early divergence of the three major teleost subdivisions. A maximum parsimony approach was employed in order to minimize the total number of independent gene losses where it was unclear whether a gene was lost once or multiple times. In order to produce a rough estimation of absolute dates for illustrative purposes, I used the mean of the range of divergence times estimated from the mitogenome clock analyses of Hurley et al. 2007 to estimate rates of hox gene loss.

## Results

### Genome assembly metrics

Using the *de novo* assembly software velvet (Zerbino and Birney, 2008) a range of assemblies using incremental kmers (27, 37, 47, 57) were produced for each osteoglossomorph genome and their general assembly metrics were compared. I estimated coding sequence representation by searching for 248 conserved core eukaryotic genes (CEGs) in each assembly using the CEGMA pipeline (Parra et al., 2007). In each species, I selected the best assembly defined by whichever kmer produced the highest N50 for all further analyses, as the CEG representation did not vary by more than 1 gene in each assembly version. Genome assembly metrics for these best assemblies are presented in Table 5.1. I also estimated genome size for each species using the kmer counting method (Li et al., 2010; Liu et al., 2013). In each case, the estimated genome sizes were similar to previously published genome size estimates based on flow cytometry (Hinegardner and Rosen, 1972). I estimated the genome size of *O. bicirrhosum* at 1142Mb, which is slightly higher than the flow

cytometry estimate of 1pg=977Mb. I estimated the genome size of *C. ornata* at 910Mb, slightly lower than the flow cytometry estimated its size of 1.08pg=1055Mb. Finally, I estimated the genome size of *G. petersii* at 1091Mb, which is only slightly lower than the published estimate of 1.2pg=1172Mb based on flow cytometry. The expected actual coverage of these genomes is therefore in the range of 11-14X. It is notable that the genome assembly sizes of these three species are lower than their estimated genome sizes based on both kmer counting and flow cytometry. My assembly of *O. bicirrhosum* covers only 52% of the estimated genome size. Similarly, my assembly of *C. ornata* covers 66%, and the *G. petersii* assembly covers 61%. Correspondingly, in each genome I was able to detect at least partial models of 76-83% of the CEGMA set of CEGs. This suggests these draft genomes are highly fragmented. However, because Hox genes are short 2 exon long protein coding genes and typical teleost median intron lengths range from 120-250 bp (Moss et al., 2011) the N50 values of these assembies (1.9-2.8kb) are still in a reasonable range for finding Hox genes. Gene finding in draft genomes can be difficult. Many genes which appear to be absent from a draft genome can be discovered later with the addition of greater sequencing depth, different types of sequencing data (i.e. Sanger, 454, Illumina), or improved assembly methods and this problem does not exclude even high quality reference genomes such as the human (Denton et al., 2014). I therefore expect that my estimates of Hox gene numbers in these assemblies to be conservative underestimates. However, demonstrating that a gene is present is far easier than demonstrating its absence and as my results indicate I was able to find many known and several novel genes even in these fragmented assemblies.

**Fishing for Hox in fish: the next-generation**

In order to screen the novel osteoglossomorph genome assemblies which were produced I designed a simple, rapid reciprocal-best hit (RBH) based bioinformatic screening method for surveying teleost draft genome assemblies for Hox genes.

I took advantage the manually annotated Hox clusters of the spotted gar *L. oculeatus*, a pre-TGD neopterygian fish and a member of the sister lineage to Teleostei (Braasch et al., 2016, appendix 2) , to search for similar sequences in the novel osteoglossomorph genome sequences. The gar genome was recently sequenced, and it was shown that it experienced no Hox gene losses since its divergence from its most recent common ancestor with teleosts, and only two Hox gene losses (*HoxA14*, *HoxD14*) since it diverged from the last common ancestor of bony vertebrates (Braasch et al., 2016, appendix 2). Moreover, it was also shown through comparative dN/dS estimates that the Hox protein coding sequences of gar were all evolving significantly more slowly than any of their teleost co-orthologs (Braasch et al., 2016, appendix 2), and in many cases were evolving even more slowly than their orthologs in Coelacanth, previously considered to have the slowest evolving Hox clusters amongst bony vertebrates (Koh et al., 2003; Amemiya et al., 2010). Using the reciprocal blast methodology described above, I extracted 95, 80, and 64 Hox-containing contigs from *O. bicirrhosum*, *C. ornata* and *G. petersii* respectively. Because the homeodomain sequence is highly conserved both across proteins classes and between distantly related species (Bürglin, 2011; Holland, 2013) it was not unexpected that many of the contigs returned top blastx hits to related non-

Hox homeobox genes, particularly within the ANTP class. This also gives confidence that the sensitivity of the original gar-seeded tblastn searches was high enough to find most of the Hox genes in the Osteoglossomorph genome sequences, since many more distantly related sequences were also extracted. After selecting only the contigs which return top blastx hits in nr to Hox sequences, I repeated the search using the putative osteoglossomorph hox-encoding contigs specifically against the predicted proteome of the pre-TGD actinopterygian *L. oculeatus*. With this method I found almost complete concordance between the top nr blastx hit and the top gar tblastn hit at the level of the Hox subfamily or vertebrate-tier paralogy group (PG1-13), and could usually assign the exons to a vertebrate level cluster (HoxA-D). However, if there was any discordance between the nr and gar blast searches at the level of cluster assignment I did not assign a putative cluster level homology, but preserving the subfamily level information and putative vertebrate paralogy group level information. Furthermore, due to inherent limitations of this phenetic approach, I was not able to confidently assign TGD-cluster level homology (i.e. *D. rerio* hoxa*a* vs hoxa*b* cluster) due to results of this blast-based approach which often suggested the osteoglossomorph gene fragments were more similar to pre-TGD Hox genes as post-TGD Hox genes in the teleost 'a' or 'b' paralogy groups. Furthermore, as I have previously shown and have discussed extensively in Chapter 2 of this thesis, it can be notoriously difficult, and even theoretically impossible to reconstruct 1:1 orthology relationships between many TGD duplicates in osteoglossomorphs and clupeocephalans, even with careful phylogenetic methods. Figure 5.1 depicts a summary of all putative Hox genes discovered in the genome assemblies of each of

the three osteoglossomorph genomes, including previously published results from *H. alosoides* and *P. buchholzi*, and is mapped onto an osteoglossomorph phylogeny based on the topology recovered in (Lavoué et al., 2010).

In total, I was able to identify contigs containing the putative full coding sequence (exon1 and exon2) of 15, 25, and 29 Hox genes in *O. bicirrhosum*, *C. ornata*, and *G. petersii* respectively (Figure 5.1). I additionally identified 42, 25, and 17 isolated Hox exon1-containing contigs, and 38, 29, and 20 isolated Hox exon2-containing contigs in *O. bicirrhosum*, *C. ornata*, and *G. petersii* respectively. In total, I found evidence for at least 59 Hox genes in *O. bicirrhosum* representing at least 7 clusters. Only a single contig did not pass my reciprocal blast test and could not be confidently assigned cluster-level homology, a *hox4*-like exon1 fragment. In *O. bicirrhosum* I found evidence for two HoxA clusters (16 genes which retained both duplicates of at least the hoxa13, hoxa11, hoxa10, hoxa9, and hox2 subfamilies), two HoxB clusters (17 genes which retained both duplicates of hoxb13, hoxb9, hoxb8, hoxb6, hoxb5, hoxb3 and hoxb2 subfamilies) and two HoxC clusters (20 genes with both duplicates of hoxc13, hoxc12, hoxc10, hoxc9, hoxc8, hoxc6, hoxc5, and hoxc4 subfamilies) and only one HoxD cluster (7 genes with no apparent duplicates) (Figure 5.1).

In *G. petersii* I found evidence for at least 51 genes which could be assigned to at least 6 clusters including two HoxA clusters (15 genes with duplicates of hoxa13, hoxa11, hoxa10, hoxa9, and hoxa2 subfamilies) and two HoxC clusters (17 genes with duplicates of hoxc13, hoxc12, hoxc10, hoxc9, hoxc8, hoxc6, hoxc5 and hoxc4 subfamilies) but only one HoxB cluster (10 genes with no putative duplicates) and

one HoxD cluster (6 genes with no putative duplicates). Four exon2-containing contigs could not be reliably assigned a cluster-level homology (one *hox13*-like, one *hox7*-like and two *hox3*-like) (Figure 5.1). Finally, in *C. ornata* I found at least 58 Hox genes and evidence for 8 clusters. These genes could be assigned to two HoxA clusters (12 genes with duplicates of hoxa13, hoxa11, and hoxa10 subfamilies), two HoxB clusters (14 genes with duplicates of hoxb13, hoxb9, hoxb2, and hoxb1 subfamilies), two HoxC clusters (16 genes with duplicates of hoxc13, hoxc12, hoxc11, hoxc9, hoxc8, hoxc6, hoxc5, and hoxc4 subfamilies) and two HoxD clusters (14 genes with duplicates of hoxd12, hoxd11, hoxd9 and hoxd4 subfamilies). Two hox exon2-containing contigs could not be assigned a cluster level homology (one *hox3*-like and one *hox1*-like) (Figure 5.1).

**Ancestral state reconstruction**

Based on these new data from three osteoglossomorph genomic screens, combined with the fully sequenced Hox clusters of *P. buchholzi* and the PCR screen from *H. alosoides* (Chambers et al., 2009) I was able to make the first reconstruction of the putative Hox complement of the last osteoglossomorph common ancestor (LOCA) (Figure 5.2). Furthermore, combined with the recently annotated Hox clusters of the gar *L. oculeatus* representing the Holostei, two eels (*A. anguilla, A. japonica*) representing Elopomorpha, and a broad range of genomic data available for clupeocephalan Hox clusters, I also made new reconstructions of the ancestral Hox complements of the last neopterygian common ancestor (LNCA), the last elopomorph common ancestor (LECA), the last clupeocephalan common ancestor (LCCA) and finally the last teleost common ancestor (LTCA) common to each of the

major subdivisions (Figure 5.2). I infer that the LNCA had 43 Hox genes, and that two genes (*HoxA6*, *HoxD2*) were likely lost prior to the TGD as no orthologs of these genes have been found in any extant teleost, including any of the 3 new osteoglossomorphs investigated here. In contrast, by comparing the complement recovered in the gar with those of the slow evolving sarcopterygian fish *L. chalumnae* (Koh et al., 2003) and the chondrichthyan *C. milii* (Ravi et al., 2009) it was determined that the gar lineage apparently experienced no Hox gene losses in the ensuing ~352 million years since its divergence with teleosts (Hurley et al., 2007) demonstrating remarkable evolutionary stability (Braasch et al., 2016, appendix 2). Following the TGD the Hox repertoire of the ancestral teleost increased from 41 genes organized in 4 genomic clusters to 82 genes organized in 8 genomic clusters. I found that following the divergence of the osteoglossomorph subdivision from other teleosts the LOCA still retained at least 70 Hox genes organized in 8 clusters. This is a comparable retention rate to the LECA which retained at least 72 Hox genes, even though in eels 70-71 Hox genes are present in a single extant species, while in Osteoglossomorphs the most I have evidence for in a single species is 59 genes in *O. bicirrhosum*. This suggests that lineage specific gene loss continued in osteoglossomorphs, while this process seems to have all but ceased at least within anguilliform elopomorphs. When I reconstructed the Hox complement of the LCCA I found evidence for at least 63 genes, which is considerably fewer than either the LOCA or the LECA. Furthermore, the greatest number of Hox genes observed in extant clupeocephalans (excluding those which underwent additional WGD beyond the TGD) is 49 (in *D. rerio*) while I found that there was evidence for a higher Hox

gene retention rate in every one of these three new osteoglossomorph genomes (59, 51, 58 genes). Notably, however, *P. buchholzi* has retained only 45 genes spanning 5 clusters (Chapter 2, Martin & Holland, 2014). This indicates that the situation in *P. buchholzi* is not representative of all osteoglossomorphs but rather that it experienced significant lineage specific losses and divergent resolution compared with other osteoglossomorphs, including the loss of a hoxa, hoxc and hoxd cluster which does not appear to be shared with any other osteoglossomorph. Likewise, evidence for lineage specific losses and divergent resolution within each of the three new osteoglossomorph genomes investigated here was found. These data suggests a hoxd cluster was lost from *O. bicirrhosum*, and both a hoxb and a hoxd cluster were lost independently in *G. petersii*, while no clusters were lost from *C. ornata*. This suggests a degree of reciprocal cluster loss between families of osteoglossomorphs even greater than has been observed between families of clupeocephalans which have lost at most one hox cluster per lineage.

Finally, an updated reconstruction of the Hox complement of the ancestor common to Osteoglossomorpha, Elopomorpha, and Clupeocephala – the putative last teleost common ancestor (LTCA) was carried out (Figure 5.2). Using this reconstruction, it can be shown that the LTCA retained at least 75 Hox genes organized in 8 genomic clusters. Both TGD paralogs of the hoxb13 and hoxa5 subfamilies are found only in osteoglossomorphs and not in eels or any clupeocephalan genomes and therefore could be inferred to be present in the LTCA, taking the total from 73 based on a previous reconstruction using Clupeocephala and Elopomorpha alone to the new total of 75. Therefore, after the TGD a maximum of 7

Hox genes were lost before the divergence of the 3 major subdivisions (one paralog each of the hoxa4, hoxa7, hoxb10, hoxc1, hoxd1, hoxd8, and hoxd13 subfamilies). Subsequently, 5 genes were lost in the lineage leading to the LOCA (one paralog each of the hoxa1, hoxa3, hoxb7, hoxb10, and hoxc3 subfamilies). Only 3 genes were lost in the branch leading from the LTCA to the LECA (one paralog of the hoxa5, hoxb13 and hoxd3 subfamilies). Finally, in the branch between the LTCA and the LCCA common to zebrafish, pufferfish, medaka and most other extant teleosts 12 Hox genes were lost (one paralog of each of the hoxa1, hoxa3, hoxb2, hoxb4, hoxb7, hoxb9, hoxb13, hoxc3, hoxd3, hoxd8, hoxd10 and hoxd12 subfamilies). Therefore, more Hox genes were lost in the lineage leading to the LCCA after they split from other teleosts (12) than in the period of time between the TGD and the LTCA (7), and significantly more than in the lineage leading to the LOCA (5) and LECA (3).

By using the mean of the range of estimates of divergence times based on (Hurley et al., 2007) of 352 Mya for the LNCA, 297 Mya for the LTCA, 252 Mya for the LOCA, 241 Mya for the LECA and 224 Mya for the LCCA, and the estimate of the timing of the TGD based on genome wide dating of TGD paralog divergence times in *T. nigroviridis* (Jaillon et al., 2004) of 350 Mya, I roughly estimated the gene loss rates along each of the three lineages leading to the LOCA, LECA, and LOCA after their divergence from the LTCA, and compared this with the Hox gene loss rate experienced in the period of time between the TGD and the LTCA (Figure 5.3). The loss of 7 Hox genes on the branch between the TGD and LTCA corresponds to a gene loss rate of 0.13/My, or 0.32% of the 41 additional TGD-generated Hox duplicates per My. For osteoglossomorphs, the loss of 5 genes between the LTCA

and the LOCA corresponds to a loss rate of 0.11/My, or 0.30% of TGD duplicates/My. The loss rate in elopomorphs following divergence from the TGD was only 0.05/My or 0.14%/My, while the loss rate in clupeocephalans was highest at 0.16/My or 0.44% of TGD duplicates/My. This suggests that within the first ~50 My between the TGD and the LTCA the rate of TGD Hox paralog loss (0.32%/My) was approximately the same as the rate of loss in the subsequent ~45My in the lineage leading to the last osteoglossomorph common ancestor (0.30%/My). In contrast, it slowed significantly in the lineage leading to elopomorphs (0.14%/My) while it accelerated in the lineage leading to clupeocephalans (0.44%/My). Within osteoglossomorphs the most rapid rate of post TGD Hox gene loss in the subsequent 252 My was within the Pantodontidae with 0.32%/My, while the slowest was in *O. bicirrhosum* at 0.12%/My. This suggests that the rate of Hox gene loss in the first ~50 My immediately following the TGD prior to the divergence of the three major teleost subdivisions was not significantly faster than the rate of loss in the following 45 Ma leading to the LOCA, and was in fact slower than the subsequent gene loss rate in each of the families of osteoglossomorphs I investigated in the subsequent 252 My.

## Discussion

Prior to the advent of next-generation sequencing methods which have brought the costs of producing genome-scale data into a range attainable by single laboratories (Metzker, 2009), the most popular method for surveying gene content in a novel genome of interest was PCR amplification of conserved domains using degenerate primers. This method is still in recent use and is reasonably successful in producing a first estimate of Hox gene diversity (Luo et al., 2007; Chambers et al., 2009; Guo et al., 2010; Martin and Holland, 2014). However, due to the inherent limitations and biases of PCR using degenerate primers on unknown, sometimes evolutionarily distant sequences, in every case which has been followed up with either whole genome sequencing or targeted BAC sequencing it has been shown that PCR-based Hox screens produce a great underestimate of the real gene numbers. Furthermore, the strengths of targeting a highly conserved region (usually the homeobox) with degenerate primers are counteracted by the drawbacks of having a short DNA sequence with relatively few phylogenetically informative sites for subsequent phylogenetic analysis which can cause ambiguity in orthology assignment.

The discovery of so many duplicated Hox genes in the genomes of these three osteoglossomorphs, combined with the data from eel and *P. buchholzi* has increased the minimum number of Hox genes known to be present in the LTCA to 75. Notably, two duplicates – hoxa5 and hoxb13 were not found in the otherwise Hox gene rich genomes of eels. This demonstrates that only a minuscule fraction (6/41=15%) of duplicated Hox genes were lost prior to the branching of the three major teleost

subdivisions. It also demonstrates that both the last osteoglossomorph common ancestor (70 genes, 71% ret; 29% loss) and last elopomorph common ancestor (73 genes, 78% ret; 21% loss) retained a similar total number of genes even though there are apparently many lineage specific losses in osteoglossomorphs, while anguillids experienced next to no losses. Interestingly, based on these figures both the LOCA and LECA retained more Hox genes than LCCA. These estimates however are prone to change as more teleosts genomes are sequenced. I predict that it is likely that many apparent superorder-level losses (e.g. 13 Hox gene losses in the clupeocephala) will be found to be more lineage-restricted than previously thought and increase the number of ancestral genes which can be inferred to be present at each ancestral node recursively throughout the teleost tree. Indeed, it would not be surprising to discover that some combination of teleosts have retained both paralogs of the six Hox genes currently thought to have been lost prior to the LTCA (hoxa4, hoxa7, hoxc1, hoxd1, hoxd8 and hoxd13) and that the hox clusters remained completely unreduced for ~50My between the TGD and the divergence of the three major teleost subdivisions.

Although the approach of rapid bioinformatic screening described here is effective in discovering and counting novel Hox genes in newly sequenced genomes, the fragmentary nature of these low coverage assemblies and the purely phenetic blast-based approach pose some challenges to fully describing the hox complements of these novel genomes. Although hox coding regions are relatively short compared to most other genes, it was still not possible to recover the full putative cds on a single contig for all of the Hox genes I inferred to be present in each of these novel

genomes. Full putative cds recovery on a single contig was only possible for a total of 69 of the 175 newly identified Hox genes. For a further 78 genes, putative exon 1 and exon 2 sequences were identified, but they were broken across two contigs. In these cases evidence to pair specific exon 1 and exon 2 fragments is lacking and would need to be verified by either PCR or ideally higher depth genome sequencing. Therefore, in the cases where the genes appear to be duplicated (i.e. *C. ornata hoxc8ex1-like1, hoxc8ex1-like2, hoxc8ex2-like1, hoxc8ex2-like2*) the putative full length coding sequences cannot currently be reconstructed reliably. Furthermore, in cases where only a single gene appears to be present (i.e. *G. petersii hoxb1ex1-like*, *hoxb1ex2-like*) it remains unclear whether these exons represent one full coding sequence or whether there are additional exon1 and exon2 sequences missing from the present dataset and the exons discovered actually correspond to two genes. Finally, for 28 of the 175 newly identified Hox genes only one of two putative exons could be recovered in these genome assemblies. The most likely interpretation in these cases is that the missing exon was simply not present in the assembly, rather than a single-exon loss or a failure to recover the contig using the blast methodology. However, alternative explanations which make the failure to recover the contig by blast plausible include evolutionary divergence to the point where it is not recognizable due to strong selection, or even pseudogenization of the entire gene coupled with some selective constraint on preserving the one exon recovered due to a secondary role of the hox cds which may act, for example, as a regulatory element for other genes in the cluster.

Although the fragmentary nature of the assembly prohibited the full cds reconstruction of all Hox genes in each of these novel genomes, it was still possible to recover some evidence of linkage between Hox genes in contigs which span the intergenic region between two paralogy groups was recovered. In *G. petersii* I identified one contig containing the full coding sequence of a *hoxb8*-like gene linked to exon1 of a *hoxb7*-like gene, and one contig containing exon2 of a *hoxb6*-like gene linked to the full cds of a *hoxb5*-like gene. In *C. ornata* I identified a single contig spanning two Hox genes including the full putative cds of a *hoxb2*-like and a *hoxb1*-like gene. No such contigs were found in *O. bicirrhosum*. Except for these specific cases, one cannot say for certain that Hox genes belonging to apparently singleton Hox clusters are not actually spread over 2 clusters which have experienced complementary gene losses. Furthermore, completely reciprocal gene losses between paralogy groups across a duplicated Hox cluster such that no duplicated genes have been retained despite the retention of both clusters (absolute divergent resolution between the clusters) has never previously been described but cannot be definitively ruled out with this data.

While it is difficult to say with certainty that any gene which I was unable to find with the bioinformatic screening approach was truly lost from the genome, it is notable that in both *G. petersii* and *O. bicirrhosum* my results are consistent with complete cluster losses. Only a maximum of a single gene for each paralogy group can be unambiguously assigned to the hoxd cluster in *O. bicirrhosum*, while in *G. petersii* I have evidence for only a single hoxb and a single hoxd cluster (Figure 5.1). I have previously shown that *P. buchholzi* has lost 3 whole Hox clusters (hoxa, hoxc,

hoxd), and duplicates of the hoxc and hoxd clusters have been lost in percomorphs and ostariophysiians respectively. It therefore seems plausible that *O. bicirrhosum* has lost a duplicated hoxc cluster, while *G. petersii* has lost both a duplicated hoxb and hoxd cluster. Based on mitogenomic phylogenies of osteoglossomorphs which places Osteoglossiforms, Pantodontidae, and H. alosoidestidae as successive sister lineages to a monophyletic clade consisting of Notopteridae and Mormyridae (Lavoué et al., 2010), this suggests that there was extensive reciprocal loss of whole Hox clusters between families. Interestingly, the putative loss of a duplicated hoxb cluster from the genome of *G. petersii* is the first observation of its kind. Even in *P. buchholzi* which underwent extensive Hox cluster loss, only the duplicated hoxb clusters remained. If it is verified in either subsequent improved, higher quality genome assemblies or targeted resequencing experiments, the loss of hoxb would be the first documented loss of a duplicated hoxb cluster and show that after the TGD any of the four duplicated clusters could be reduced to singletons in the osteoglossomorph lineage.

It was suggested that one possible explanation for the high retention of Hox genes in eel was constraint necessitated by the evolution of its elongated body plan (Henkel et al., 2012b). These new data suggest that not only the last elopomorph common ancestor, but also the last osteoglossomorph common ancestor was also extremely replete in Hox genes suggesting that factors besides the evolution of the anguilliform body plan were able to prevent gene losses within the hox clusters of two major teleost subdivisions for ~100 My prior. It will be interesting to determine if the subsequent high degree of divergent resolution apparent in the Hox clusters of

osteoglossomorphs after their family-level divergence ~252 My ago also occurred in elopomorphs or whether the entire subdivision maintained an eel-like Hox-rich genome. If other non-anguilliform elopomorphs such as tarpon, bonefishes, and ladyfishes also retained as many Hox genes as eels it would suggest that there may be a subdivision-wide constraint on Hox cluster evolution following the TGD in elopomorphs setting them apart from osteoglossomorphs and clupeocephalans.

Overall, the results presented here contradict to the prevailing model of gene loss following WGD which suggests that 'gene collapse' occurs rapidly following a WGD (Sankoff et al., 2010; Inoue et al., 2015) and favours a model of slow and stepwise gene loss during a prolonged diploidization process. Interestingly, the rate of gene loss in the clupeocephalan branch following divergence from the LTCA (0.44%/My) was actually more rapid than within the first 50My following the TGD (0.32%). The higher rate of Hox gene loss in the clupeocephalan lineage than either the osteoglossomorph or elopomorph lineages is yet another indication that clupeocephalans experienced different evolutionary pressures following the TGD than either of the other subdivisions. The only lineage in which there is less evidence for slow, stepwise gene loss than rapid gene collapse is the elopomorpha where the majority of all losses which occurred in eels occurred prior to the LTCA, after which Hox gene loss crawled to a near standstill (0.01%/My). However, because so few genes were lost in total in eels this still does not fit the expectation of a 'rapid gene collapse' model.

## Conclusions

In conclusion, I have shown evidence for divergent resolution of Hox clusters between representative taxa of 5 families of osteoglossomorph teleosts. While there is evidence for up to 8 clusters in the Notopterid *C. ornata* and the goldeye *H. alosoides*, the Mormyrid *G. petersii* appears to have lost a hoxb and a hoxd cluster, the Osteoglossid *O. bicirrhosum* appears to have lost a hoxd cluster, and the monotypic member of the Pantodontidae, *P. buchholzi* has lost one of each hoxa, hoxc and hoxd cluster. Together, this suggests that an even greater degree of divergent resolution of Hox clusters has occurred between families within Osteoglossomorpha than observed in either Elopomorpha or Clupeocephala. Ancestral state reconstructions further revealed that the Hox complements of the LOCA and the LECA were both more replete than the LCCA, and that there is no evidence for an increased rate of Hox gene loss in the first 50 My following the TGD along any other branch of either the osteoglossomorph or clupeocephalan trees and that in fact the rate of Hox gene loss increased specifically in the clupeocephalan lineage following its divergence from other teleosts.

**Table 5.1.** Osteoglossomorph genome assembly metrics.

Assembly contiguity and protein-coding representation assessments for selected *de novo* assemblies of 3 novel osteoglossomorph genomes.

| | ***O. bicirrhosum*** | ***C. ornata*** | ***G. petersii*** |
|---|---|---|---|
| **PE reads** | 64,562,111 | 70,583,995 | 69,889,639 |
| **Data (Gb)** | 12.91 | 14.12 | 13.98 |
| **Genome coverage estimate** | 13.09X | 14.29X | 11.09X |
| **Genome size estimate (Mb)** | 1141.97 | 910.4 | 1090.76 |
| **Best kmer** | 47 | 27 | 47 |
| **Contig N50 (bp)** | 1,936 | 2,880 | 2,816 |
| **Contigs** | 436,545 | 338,181 | 398,683 |
| **Total Assembly (Mb)** | 594.35 | 604.07 | 669.08 |
| **% GC** | 45.76 | 41.86 | 43.82 |
| **Full CEGs** | 109 (44%) | 103 (41%) | 120 (48%) |
| **Partial CEGs** | 189 (76%) | 204 (82%) | 205 (83%) |

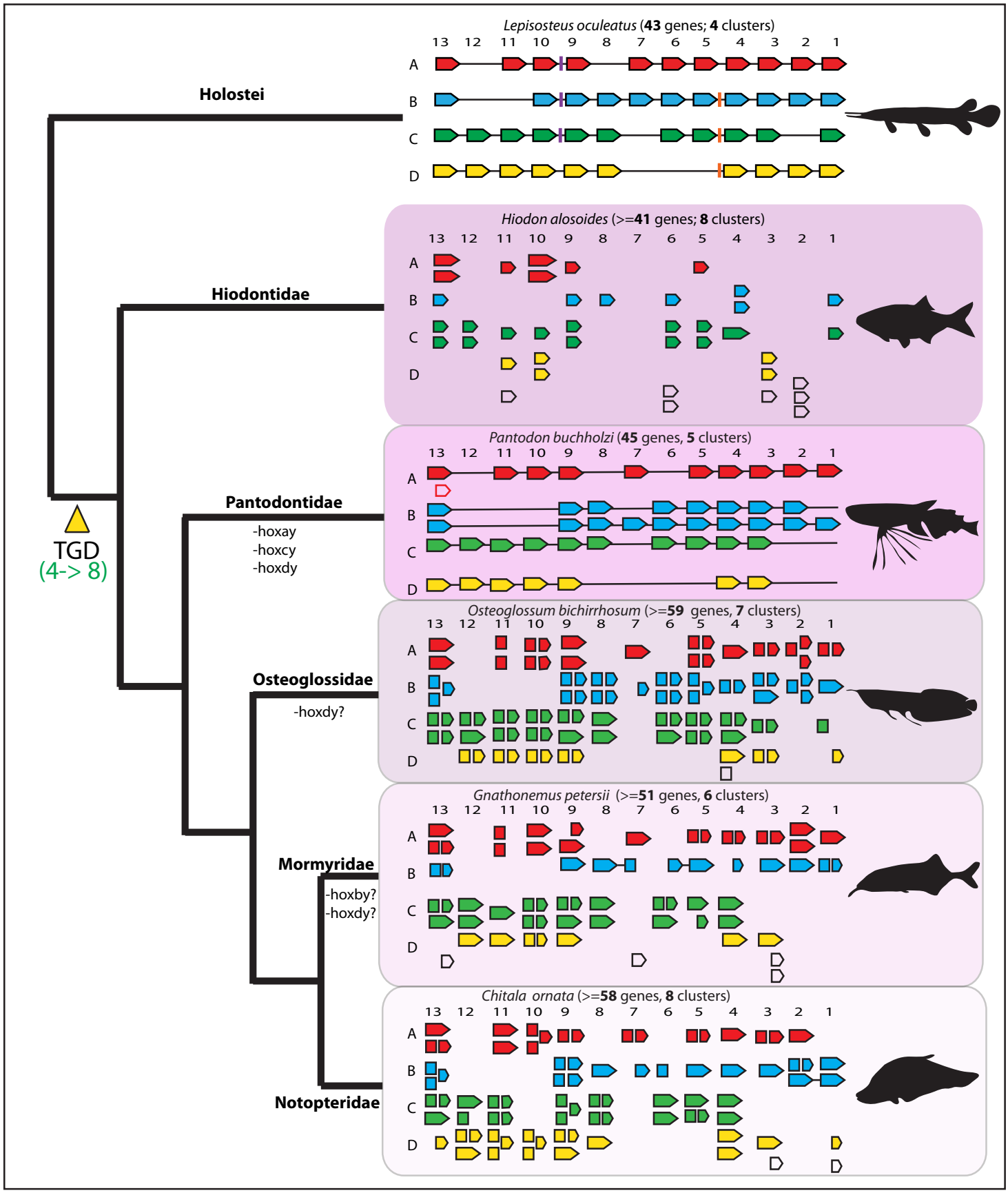

Lepisosteus oculeatus (43 genes; 4 clusters)
13 12 11 10 9 8 7 6 5 4 3 2 1
A
B
C
D
Holostei
Hiodon alosoides (>=41 genes; 8 clusters)
Hiodontidae
Pantodon buchholzi (45 genes, 5 clusters)
Pantodontidae
-hoxay
-hoxcy
-hoxdy
TGD
(4-> 8)
Osteoglossum bichirrhosum (>=59 genes, 7 clusters)
Osteoglossidae
-hoxdy?
Gnathonemus petersii (>=51 genes, 6 clusters)
Mormyridae
-hoxby?
-hoxdy?
Chitala ornata (>=58 genes, 8 clusters)
Notopteridae

**Figure 5.1.** Hox cluster complements in five osteoglossomorph families.
Hox gene complements from *O. bicirrhosum*, *G. petersii*, and *C. ornata* recovered in this study are presented alongside the previously surveyed Hox gene complements from *P. buchholzi* and *H. alosoides* to present a first estimate of subdivision-wide changes in Hox gene content in the Osteoglossomorpha. The HoxA cluster is shown in red, the HoxB cluster is shown in blue, the HoxC cluster is shown in green, and the HoxD cluster is shown in yellow. Genes which were not assigned to particular clusters but could be assigned putative paralogy groups (PG1-13) are shown as unfilled boxes in the corresponding column of their most likely paralogy group. Putative exon 1 sequences are shown as squares, exon 2 sequences shown as arrow shaped pentagons. Where the full putative cds was recovered an arrow shaped pentagon of twice the length is shown. Putative cluster losses inferred from the absence of any TGD-level gene duplicates are mapped along each branch of the osteoglossomorph tree. The loss of a hoxa, hoxc, and hoxd cluster occurred in *P. buchholzi* the loss of a hoxd cluster occurred in *O. bicirrhosum*, the loss of a hoxd and a hoxb cluster occurred in *G. petersii*, while we recovered TGD paralogs of every cluster in *C. ornata* suggesting it did not experience any Hox cluster losses. The Hox complement of the gar *L. oculeatus*, a pre-TGD holostean fish is the most closely related outgroup to the duplicated Hox clusters of teleosts.

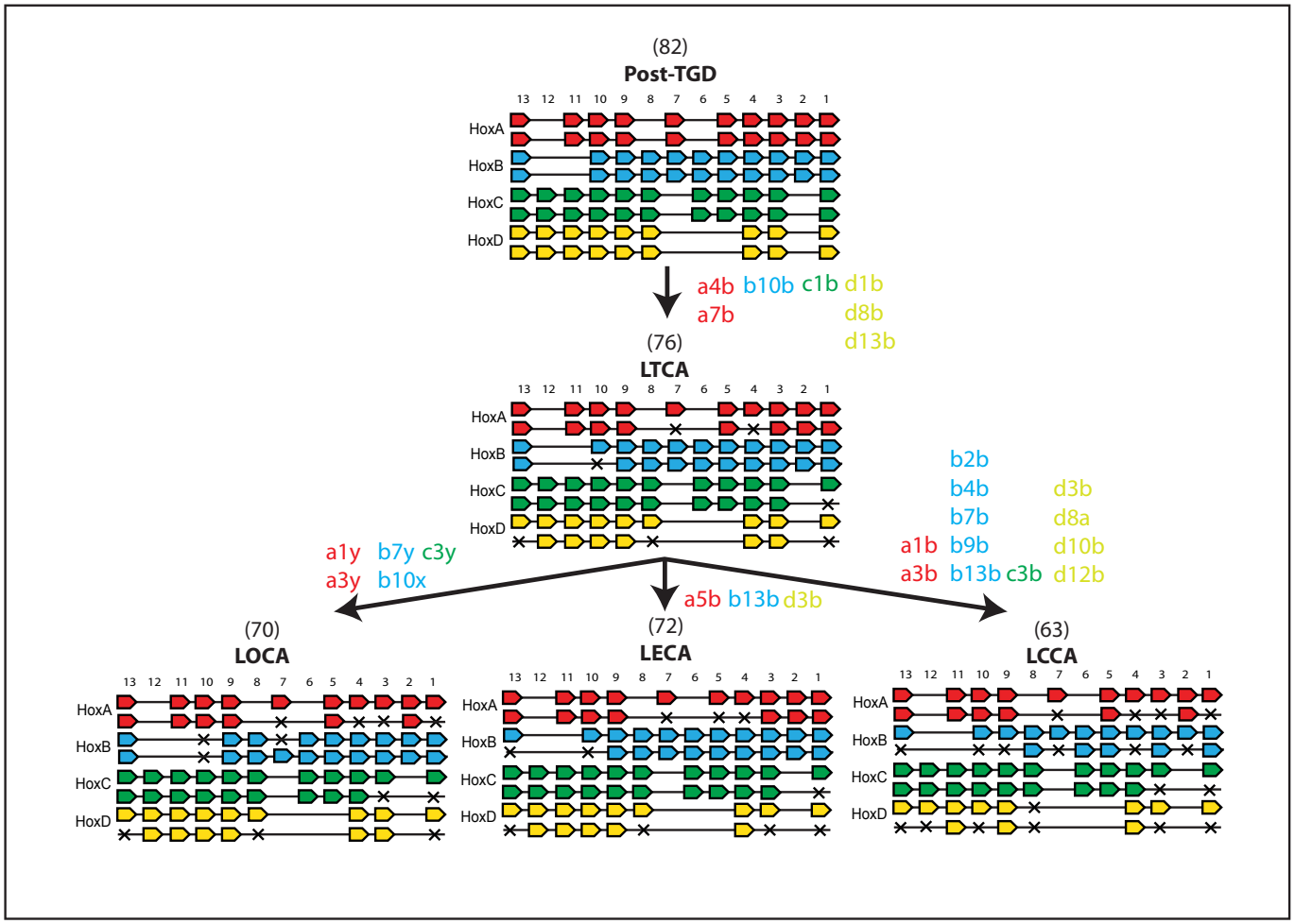



**Figure 5.2.** Reconstruction of ancestral Hox complements.

Reconstructions of the Hox complements of the last common ancestor of crown teleosts (i.e. the last teleost common ancestor (LTCA)), the last osteoglossomorph common ancestor (LOCA), the last elopomorph common ancestor (LECA), and the last clupeocephalan common ancestor (LCCA), with gene losses mapped along each branch of the major teleost subdivisions. Genes belonging to the HoxA clusters are shown in red, HoxB in blue, HoxC in green, and HoxD in yellow. Within cluster schematics losses are represented by an 'x', and the genes which were lost specifically in the lineages between each reconstructed ancestral state are named and colour coded to their Hox cluster.

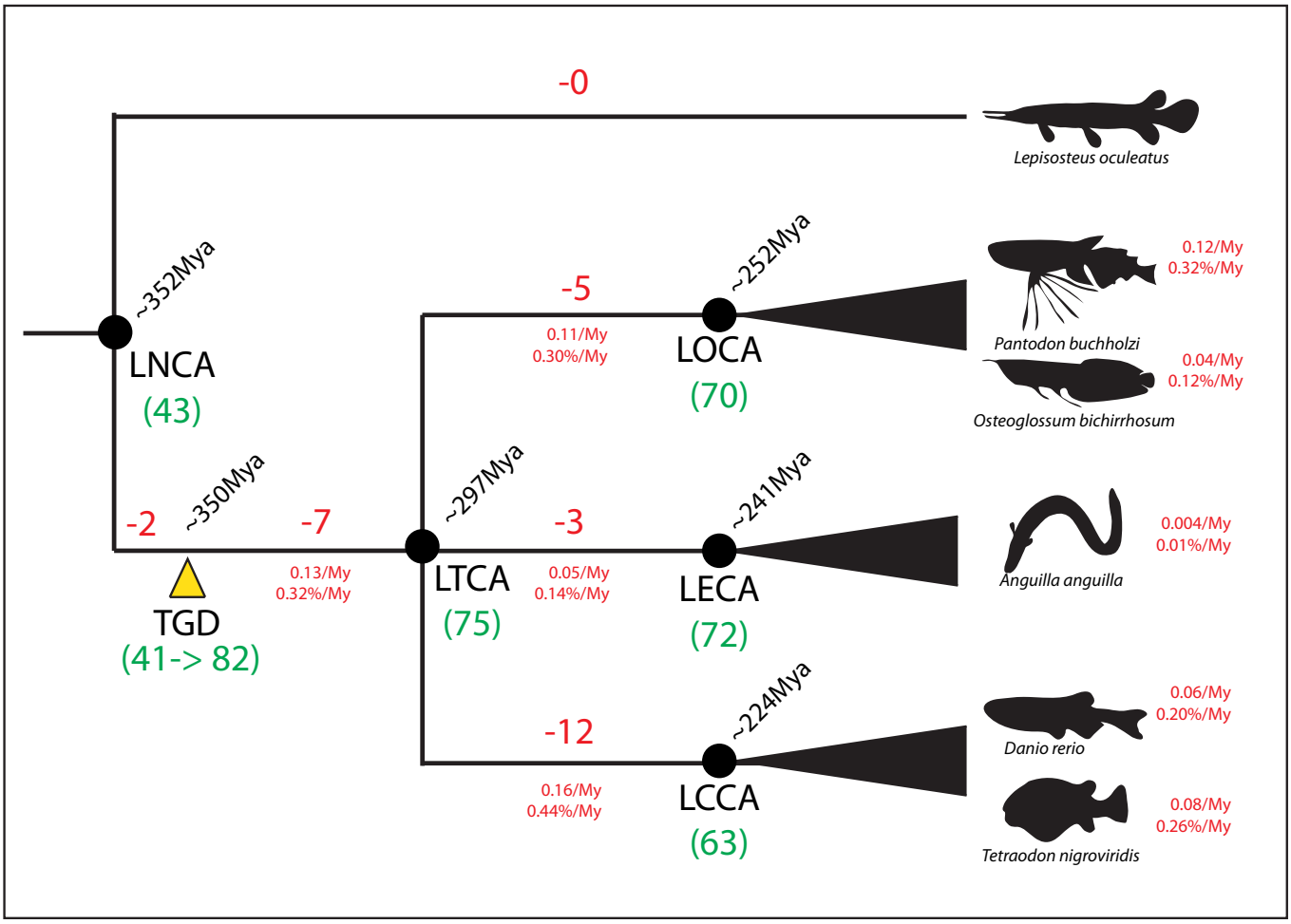



**Figure 5.3.** Rate of Hox gene losses mapped onto basal teleost phylogeny.

Number of hox gene losses is shown in red above each branch. Rate of hox gene losses in terms of number of hox genes, and % of the TGD-generated hox gene duplicates present at the preceding node, per My are shown in red below each branch. Divergence time estimates are based on the mean of the range of estimates from Hurley et al. 2007. Number of genes reconstructed at each ancestral node is shown in green. Overall, it can be seen that in the first ~50My following the TGD, very few hox genes were lost and that divergent resolution occurred at the level of the three major subdivisions following their speciation. LNCA = last neopterygian common ancestor, TGD = teleost genome duplication, LTCA = last (crown) teleost common ancestor, LOCA = last osteoglossomorph common ancestor, LECA = last elopomorph common ancestor, LCCA = last clupeocephalan common ancestor.

# Chapter 6

## General Discussion

Until now, research into the evolutionary patterns and processes associated with gene duplication has been largely focused on examples where lineages under investigation presently retain both members of a duplicated gene pair. This has been true for both duplicates generated by simple single gene duplication processes or large scale WGDs such as the VGD1/VGD2 events in vertebrates (reviewed in Dehal & Boore 2005) which occurred ~525 Mya and the more recent TGD which occurred ~350 Mya in the common ancestor of crown teleosts (reviewed in Glasauer & Neuhauss 2014). However, it is important to remember that no matter how long ago the WGD event and how distinct duplicates have become in the genome of extant species, all ancient palaeopolyploids including teleosts experienced a period of their evolutionary history as *de facto* polyploids when all genes were present as duplicates and this could have significantly influenced the evolution of duplicate gene pairs as highlighted throughout in this thesis. This calls into question the existence of so-called 'singleton' genes as every gene in palaeopolyploids was once a duplicate and has important theoretical consequences especially in the field of basal teleost phylogenomics where 'singleton' genes are frequently used as a proxy for true orthologs without careful phylogenetic assessment of 1:1 orthology as elaborated below.

The diploidization process in ancient palaeopolyploids has been considered most thoroughly in yeast and plants (Wolfe, 2001; Hufton and Panopoulou, 2009) but has largely been neglected in vertebrates. This may be about to change as recent whole

genome analyses in ‘pseudotetraploid’ salmonids including the rainbow trout (Berthelot et al., 2014) and Atlantic salmon (Lien et al., 2016), which experienced a shared salmonid specific genome duplication (SSGD) no later than 88 million years ago (Macqueen and Johnston, 2014) begin to dissect in depth the rate, proportion and biases in diploidization rates following the SSGD and have converged on the conclusion that diploidization can take a significantly longer period of time than previously thought as well as exhibiting different rates and patterns across the genome. The recent history of polyploidization in the salmonid lineage as well as their ecological and morphological diversity, value to industry, and experimental tractability has poised this lineage to lead the field in understanding the diploidization process following WGD. Interestingly, these recent analyses all agree with the model of delayed diploidization I have proposed following the more ancient TGD using deep comparative genome analysis in *P. buchholzi* and other osteoglossomorphs throughout this thesis. This concordance between the results in salmonids following the SSGD and those in osteoglossomorphs following the TGD suggests that further larger-scale whole genome analysis in *P. buchholzi* and other osteoglossomorphs may continue to yield novel insight into the rate, pattern and process of diploidization genome wide following the more ancient TGD.

Despite the abundant research opportunities into diploidization offered up by the salmonid system, further research into the diploidization process following more ancient WGD events including the TGD is still warranted. One area where the tetralogy concept outlined in Chapter 2 may have profound implications concerns the outstanding question of whether osteoglossomorphs, elopomorphs, or a clade containing both, is the

sister clade to clupeocephalan teleosts. Analysis using full mitochondrial genomes supports Osteoglossomorpha as the sister group to other teleosts (Inoue et al., 2001, 2003), while a more recent supertree analysis of 9 nuclear gene exons suggests Elopomorpha is the sister group (Near et al., 2012). Even some support for a combined (Osteoglossomorpha + Elopomorpha) sister clade to other teleosts has been found using ribosomal RNA genes (Le et al., 1993). In order for gene trees to have any hope of accurately reflecting real underlying evolutionary relationships between species it is absolutely necessary to limit phylogenetic analysis to 1:1 orthologs. The problem of hidden paralogy, the mistaken use of paralogs rather than orthologs in phylogenetic trees as a result of a gene loss, is a known source of bias in phylogenomic analysis (Kuraku 2010). The prospect of tetralogous genes, that is genes which fundamentally do not have an unbroken lineage of 1:1 orthology between all species sharing a WGD event, adds a further complication to species tree reconstruction. In principle, any genes which cannot be shown to share a clearly orthologous relationship between species, including tetralogous genes, should not be used in basal teleost phylogenomics. Although unduplicated mitochondrial genomes are not expected to be fraught with the same problems, all nuclear genes used for basal teleost phylogenomics should first be screened for potential hidden tetralogy, and only those loci which form well supported orthologous relationships should be used. The same cautious approach should be taken when attempting to reconstruct the phylogenetic relationships between any lineages which diverged following a WGD event. Hidden tetralogy will most likely only affect the phylogenetic relationships between species which diverged close to a WGD event and may not have experienced complete communal diploidization. My

results suggest that Osteoglossomorpha, Elopomorpha, and Clupeocephala may fit this criterion, and hidden tetralogy between phylogenetic markers of nuclear origin may be one reason why the relationship between these clades remains largely enigmatic.

A second outstanding controversy to which the concept of tetralogy may apply concerns the uncertainty surrounding the relative timing of the vertebrate VGD1/VGD2 events to the divergence of agnathans and gnathostomes (Kuraku, 2008; Shimeld and Donoghue, 2012). Extant agnathans, comprising hagfish and lampreys, have many duplicated genes in similar proportions to gnathostomes, but when phylogenetic trees are constructed clear 1:1 orthologies with gnathostome genes are often absent (Escriva et al., 2002; Irvine et al., 2002; Stadler et al., 2004; Qiu et al., 2011; Fujimoto et al., 2013). Explanations to account for this have included agnathans and gnathostomes sharing only one WGD (plus additional duplications in each lineage), or multiple independent duplications in each lineage. Analysis of the draft version of the sea lamprey (*Petromyzon marinus*) genome suggested that both VGD1 and VGD2 were likely shared between agnathans and gnathostomes, largely based on shared synteny (Smith et al., 2013). However the most thorough study of *Hox* genes in lamprey to date has found evidence of at least six Hox clusters implicating at least one independent round WGD after VGD1/VGD2 in the lamprey lineage, and casting doubt on whether even VGD1/VGD2 were shared based partially on absence of 1:1 orthology of any lamprey genes with gnathostomes (Mehta et al., 2013). My data suggest that, at least in the case of the TGD, gene trees do not necessarily resolve into clear orthology groups even following a single shared WGD event and that the expectation of 1:1 orthology following a shared WGD may be unwarranted. One might speculate that if the

ancestral vertebrate went through a prolonged tetraploid (or octoploid) stage, as previously suggested (Furlong and Holland, 2002), with some regions of the genome diploidizing before divergence of agnathans and gnathostomes, and others diploidizing independently in each lineage, it could lead to the same type of incongruent phylogenetic patterns which was observed between the Hox and other ANTP genes in osteoglossomorphs and clupeocephalans throughout this thesis. Of course, these explanations are not mutually exclusive and it is possible that a combination of independent and shared WGD and diploidization events occurred.

In the immediate future it will be interesting to survey other lineages which have experienced WGD either recently or in their distant ancestry and determine how widespread the phenomenon of independent diploidization truly is. As described previously, while plants and especially angiosperms have an evolutionary history rife with WGD events, the majority of plant polyploids are thought to have been formed via allotetraploidy, or the hybridization and merger genomes from two distinct lineages (Otto and Whitton, 2000). For example, in maize it is proposed that segmental allotetraploidy between two species which diverged ~20.5 Mya and that the allopolyploidization event occurred ~11.4 Mya (Gaut and Doebley, 1997). In cases where the allopolyploidization is a result of the merger of two highly diverged genomes, it is not expected that independent diploidization would be a significant influence in their evolution post WGD as the two diverged genomes would have been essentially 'pre-diploidized' and segregate as independent loci immediately upon formation of the polyploid hybrid (Doyle and Egan, 2010). However, if two recently diverged but still highly similar genomes hybridized to form an allopolyploid, a proportion of the resultant chromosomes

may exhibit tetrasomic inheritance for a period, effectively delaying their divergence and causing the coalescent of the gene pair to appear younger than the actual hybridization date (Doyle and Egan, 2010). It may therefore be that in a significant proportion of allopolyploids as well as autopolyploids, delayed diploidization of duplicates generated by WGD may be an important factor in the subsequent evolution of that lineage. In support of this view, some authors have noted the relatively high prevalence of asymmetric species distributions in plant lineages which experienced a WGD event and have proposed the WGD radiation lag-time model as a general pattern for lineage diversification post WGD (Schranz et al., 2012). Schranz et al. 2012 noted that many well-studied cultivated plant groups which experienced independent WGD events including Brassicaceae, Poaceae, Asteracea, Fabaceae and Solanaceae show huge assymetries in species richness such that highly diverse crown groups were paired with relatively species poor sister lineages and determined that WGD alone was not sufficient to stimulate the diversification of these lineages. Schranz et al. 2012 propose that WGD in plants at least may somehow 'prime' the lineage for subsequent diversification but only when stimulated by appropriate environmental conditions, in a similar vein to the proposed mechanism of climate-induced salmonid diversification tens of millions of years following the SSGD, suggested by Macqueen & Johnston 2014 and corroborated by whole genome scale analyses (Berthelot et al., 2014; Lien et al., 2016). The results presented in this thesis corroborate the applicability of the WGD radiation lag-time model in teleosts following the ancient TGD simply by observing the relatively poor diversity of osteoglossomorphs and elopomorphs compared with the extremely rich clupeocephalan crown and refute the contention that WGD was sufficient for driving the

teleost radiation, as previously proposed by Santini et. al 2009 who demonstrated that less than 10% of extant teleost diversity could be attributed to accelerated speciation rates on the branch of the teleost tree prior to the diversification of osteoglossomorphs, elopomorphs and clupeocephalans. The model of delayed and independent diploidization proposed here however goes further in suggesting a possible mechanistic explanation for such a lag between WGD and diversification. If ongoing homeologous recombination was actively preventing the phylogenetic and functional divergence of WGD-generated duplicates then even in response to environmental conditions permissive of radiation the genomes of recently polyploidized lineages may actively be impaired from adapting, as selection could not effectively mould the redundant duplicates present in the genome until they became disentangled one from the other. It is interesting to speculate therefore whether large-scale chromosomal recombination events such as the 8 major rearrangements predicted to have occurred in the ancestor of at least all clupeocephalans (Kasahara et al., 2007) were necessary to break the bonds of tetrasomy, expose independent duplicated genes to the scrutiny of selection, and permit their adaptive divergence. Under this scenario one could imagine that WGD is akin to storing potential evolutionary 'energy' of a lineage in a dam-like state which could not be released until large-scale genomic changes effectively open the adaptive 'sluice gate' facilitating speciation and adaptive divergence.

Unlike the analogy of the dam however, as evidence from this thesis suggests following the TGD and as the recently sequenced salmonid genomes corroborate in the case of the SSGD (Berthelot et al., 2014; Lien et al., 2016), not all regions of the genome are likely to have been released from the constraint of tetrasomic inheritance

following a WGD (whether allopolyploid or autopolyploid) simultaneously. Therefore, in addition to the accumulation of lineage specific gene losses, it is probable that in cases where diploidization only occurred subsequent to divergence of two lineages that independent subfunctionalization or neofunctionalization may have occurred. Therefore, for each locus following a WGD the potential for independent diploidization following speciation can offer more than one opportunity for evolutionarily adaptive changes to occur between the duplicates in each lineage where diploidization at that locus had not terminated prior to their divergence. While a duplicated gene pair may drift towards subfunctionalization in one lineage following the conclusion of diploidization, in another lineage (perhaps under different environmental pressures) one of the redundant duplicates may experience strong selection and take on new functions in a different context. It will be interesting in the future to test to what degree functional divergence of paralogs converges on a similar outcome or takes a different path between independently diploidized duplicates (confirmed by stringent phylogenetic and synteny analyses) in lineages which separated prior to the termination of diploidization including Osteoglossomorpha, Elopomorpha and Clupeocephala following the TGD as well as the deeply diverging salmonid lineages Coregoninae, Thymallinae and Salmoninae following the SSGD. Using independently diploidized, truly tetralogous loci, the repeatability of functional evolution by subfunctionalization and neofunctionalization following WGD may be tested.

By accounting for possible independent diploidization the number of potential outcomes from WGD therefore increases exponentially in proportion to the number of speciation events which occur prior to the conclusion of diploidization. Therefore, while

several lines of evidence now strongly refute the idea that the TGD was sufficient for causing the teleost radiation, and even to the contrary may actually have caused an initial delay in species diversification, when considering independent diploidization the number of opportunities for evolutionary innovation to occur between duplicates increases dramatically suggesting that the TGD may potentially be even more impactful as a long-term driver of diversification in teleosts than previously recognized.

# Appendix 1

## Related to Chapter 2: Enigmatic orthology relationships between Hox clusters of the African butterfly fish and other teleosts following ancient whole genome duplication

Published in Molecular Biology and Evolution (Martin and Holland 2014)

# Enigmatic Orthology Relationships between *Hox* Clusters of the African Butterfly Fish and Other Teleosts Following Ancient Whole-Genome Duplication

Kyle J. Martin[1] and Peter W.H. Holland*[,1]

[1]Department of Zoology, University of Oxford, Oxford, United Kingdom

***Corresponding author:** E-mail: peter.holland@zoo.ox.ac.uk.

**Associate editor:** Xun Gu

## Abstract

**Numerous ancient whole-genome duplications (WGD) have occurred during eukaryote evolution. In vertebrates, duplicated developmental genes and their functional divergence have had important consequences for morphological evolution. Although two vertebrate WGD events (1R/2R) occurred over 525 Ma, we have focused on the more recent 3R or TGD (teleost genome duplication) event which occurred approximately 350 Ma in a common ancestor of over 26,000 species of teleost fishes. Through a combination of whole genome and bacterial artificial chromosome clone sequencing we characterized all *Hox* gene clusters of *Pantodon buchholzi*, a member of the early branching teleost subdivision Osteoglossomorpha. We find 45 *Hox* genes organized in only five clusters indicating that *Pantodon* has suffered more *Hox* cluster loss than other known species. Despite strong evidence for homology of the five *Pantodon* clusters to the four canonical pre-TGD vertebrate clusters (one HoxA, two HoxB, one HoxC, and one HoxD), we were unable to confidently resolve 1:1 orthology relationships between four of the *Pantodon* clusters and the eight post-TGD clusters of other teleosts. Phylogenetic analysis revealed that many *Pantodon* genes segregate outside the conventional "a" and "b" post-TGD orthology groups, that extensive topological incongruence exists between genes physically linked on a single cluster, and that signal divergence causes ambivalence in assigning 1:1 orthology in concatenated *Hox* cluster analyses. Out of several possible explanations for this phenomenon we favor a model which keeps with the prevailing view of a single TGD prior to teleost radiation, but which also considers the timing of diploidization after duplication, relative to speciation events. We suggest that although the duplicated hoxa clusters diploidized prior to divergence of osteoglossomorphs, the duplicated hoxb, hoxc, and hoxd clusters concluded diploidization independently in osteoglossomorphs and other teleosts. We use the term "tetralogy" to describe the homology relationship which exists between duplicated sequences which originate through a shared WGD, but which diploidize into distinct paralogs from a common allelic pool independently in two lineages following speciation.**

***Key words:*** **tetraploidy, diploidization, homeobox, WGD, *Pantodon*, teleost.**

Article

## Introduction

The vast majority of protein-coding genes can be organized into multigene families whose history has been shaped by duplication, loss, and speciation events (Dayhoff 1976; Nei and Rooney 2005; Demuth and Hahn 2009). When comparing genes between species, sequences are classified by categories of molecular sequence homology first introduced by Fitch (1970) over 40 years ago: Orthology (the relationship between genes present in a common ancestor and separated through a speciation event) or paralogy (the relationship between sister genes generated through a gene duplication event). Establishing correct homology relationships is a prerequisite for a powerful range of comparative genomic analyses, but requires careful phylogenetic assessment (Thornton and DeSalle 2000). It is essential to distinguish between orthologous and paralogous genes because only orthologous genes related through direct vertical descent from a common ancestor can be used to infer species-level relationships using phylogenetic tools. The type of homology between two genes may also bear on their functional evolution. The classical view is that orthologs tend to be more similar in function than paralogs, which are predisposed to diverge following duplication (Koonin 2005; Dolinski and Botstein 2007). However, the generality of this phenomenon, recently dubbed the "orthology-conjecture," has been called into question (Studer and Robinson-Rechavi 2009; Nehrt et al. 2011; Gabaldón and Koonin 2013). With the explosion in the availability of genome sequence data, largely facilitated by the advent of "Next-Generation" sequencing technologies (Metzker 2009), the importance of accurate homology assessment has grown. This is especially true in lineages where whole-genome duplication (WGD) has occurred, resulting in the simultaneous duplication of all genes.

The high prevalence of extant polyploids suggests that WGD is both a pervasive and contemporary influence in genome evolution (Otto and Whitton 2000; Mable 2004; Otto 2007). Several clades, including vertebrates (Ohno 1970; Sidow 1996; Li et al. 2001; Putnam et al. 2008), have

© The Author 2014. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Open Access

also experienced ancient WGDs deep in their evolutionary past, and are today considered paleopolyploids (reviewed in Van de Peer et al. 2009). Despite their polyploid ancestry, living descendants from each of these ancient WGD events are now karyotypically diploid, having undergone the remarkable and yet still largely enigmatic evolutionary transition from tetrasomic back to disomic inheritance known as diploidization. The fundamental change that characterizes diploidization is the termination of recombination between duplicated homeologous chromosomes, and their preferential bivalent pairing such that four genetically interacting tetraploid alleles at a single locus is transformed into two pairs of diploid alleles at two independent (now paralogous) loci (Wolfe 2001). Only after recombination between duplicated loci has ceased can mutations become independently fixed and duplicates diverge; we refer to this as the allele to paralog transition. Two pathways to polyploid formation are recognized: Allopolyploidy where interspecific hybridization results in the fusion of two separate parental genomes, and autopolyploidy which involves intraspecific doubling of chromosomes without hybridization. The mode of polyploidization in most paleopolyploid lineages and factors potentially affecting the rate and pattern of the subsequent diploidization process is largely unknown. Addressing these outstanding questions will be necessary to fully describe the earliest stages of duplicate gene divergence following both recent polyploidization and ancient WGDs.

Within the vertebrates, two ancient WGD events are thought to have occurred in the stem lineage over 525 Ma, around the time of the divergence between the living jawless (agnathan) and jawed (gnathostome) vertebrates (fig. 1*A*; node 3); these are known as 1R/2R (first round/second round of WGD) (Holland et al. 1994; Li et al. 2001; Kuraku 2008; Putnam et al. 2008; Van de Peer et al. 2010; Smith et al. 2013). The vertebrate 1R/2R duplications were pivotal in the expansion of developmental gene families; following WGD there was a preferential retention of duplicates from this functional group, whereas most other gene families reverted to single copy (Freeling and Thomas 2006; Putnam et al. 2008; Edger and Pires 2009; Makino and McLysaght 2010). *Hox* genes encode transcription factors with many highly conserved roles during development including specification of anteroposterior positional identity across bilaterians. The expansion of the *Hox* gene repertoire by WGD in vertebrate evolution is perhaps the most widely cited example of developmental gene family expansion through duplication and has been credited with permitting a wide range of evolutionary innovations, including the establishment of fundamental components of the vertebrate body plan (Wada et al. 1999; Málaga-Trillo and Meyer 2001; Pick and Heffer 2012; Pascual-Anaya et al. 2013). The vertebrate 1R/2R events were responsible for generation of the four ancestral gnathostome *Hox* gene clusters (HoxA, HoxB, HoxC, and HoxD) from the single cluster shared with the last common ancestor of vertebrates and the cephalochordate *Branchiostoma* (Garcia-Fernàndez and Holland 1994; Putnam et al. 2008). Because of their highly conserved, colinear genomic organization, the vertebrate *Hox* gene clusters can serve as useful markers of the duplication status of entire genomes. It was partially on this basis that an additional independent WGD event was found in the teleost fish lineage (Amores et al. 1998). This event, known alternately as either 3R (third round of WGD) or the TGD (teleost genome duplication) (fig. 1*A*; node 13), occurred in the common ancestor of all approximately 26,000 species of teleost fish around 350 Ma and has been implicated by some authors in explosive speciation and morphological radiation of this group (which comprise ~50% of all vertebrate biodiversity) (Christoffels et al. 2004; Hoegg et al. 2004; Jaillon et al. 2004; Meyer and Van de Peer 2005; Ravi and Venkatesh 2008).

Progress toward understanding the TGD event has been made by studying the relatively few teleost species with good genomic resources available, such as the laboratory models zebrafish (*Danio rerio*) and medaka (*Oryzias latipes*), the genomic models Fugu (*Takifugu rubripes*) and green spotted pufferfish (*Tetraodon nigroviridis*), and evolutionary genetic model stickleback (*Gasterosteus aculeatus*). However, all of these species belong to the single largest teleost subdivision: The Clupeocephala (fig. 1*A*, node 16), which in turn comprised the superorders Ostarioclupeomorpha (Ostariophysii + Clupeomorpha) (fig. 1*A*, node 17) and Euteleostei (fig. 1*A*, node 18). The remaining two smaller but earlier branching teleost subdivisions, Osteoglossomorpha (fig. 1*A*, node 14) and Elopomorpha (fig. 1*A*, node 15) are studied less, but are poised to extend the range of comparative genomic analyses to a time closer to the TGD. Molecular phylogenetic studies have disagreed on whether osteoglossomorphs, elopomorphs, or a clade containing both is the sister group to clupeocephans (Le et al. 1993; Inoue et al. 2001, 2003; Hurley et al. 2007; Alfaro et al. 2009; Near et al. 2012). Genomic resources for elopomorphs in the genus *Anguilla* have recently been developed (Coppe et al. 2010; Henkel, Dirks, et al. 2012), leaving the osteoglossomorphs the only major teleost subdivision without a substantial amount of genomic data available. The Osteoglossomorpha (bony-tongues) consists of approximately 220 species (<1% of all teleosts) in five families: Hiodontidae, Osteoglossidae, Notopteridae, Mormyridae, and Pantodontidae (Hilton 2003; Sullivan 2004). Pantodontidae consists of only a single species, a small (~12 cm) tropical fish endemic to the fresh waters of Western Africa and common to the European aquarium trade since at least 1905 (White 1994): The African freshwater butterflyfish, *Pantodon buchholzi* (fig. 1*B*). The small size and widespread availability of *Pantodon*, in addition to its relatively small genome size (~753 Mb) (Hinegardner and Rosen 1972), make it an attractive model for genomic investigations.

Seven clusters of *Hox* genes have been found in clupeocephalan teleost genomes, resulting from the duplication of the four gnathostome clusters to eight during the TGD, followed by the loss of either a single hoxc or a single hoxd cluster in the percomorph (i.e., pufferfishes, medaka, and stickleback) and ostariophysiian (i.e., zebrafish) lineages, respectively (Woltering and Durston 2006). When reconstructing gene genealogies for multiple species the resolution of two reciprocally monophyletic orthology groups, conventionally

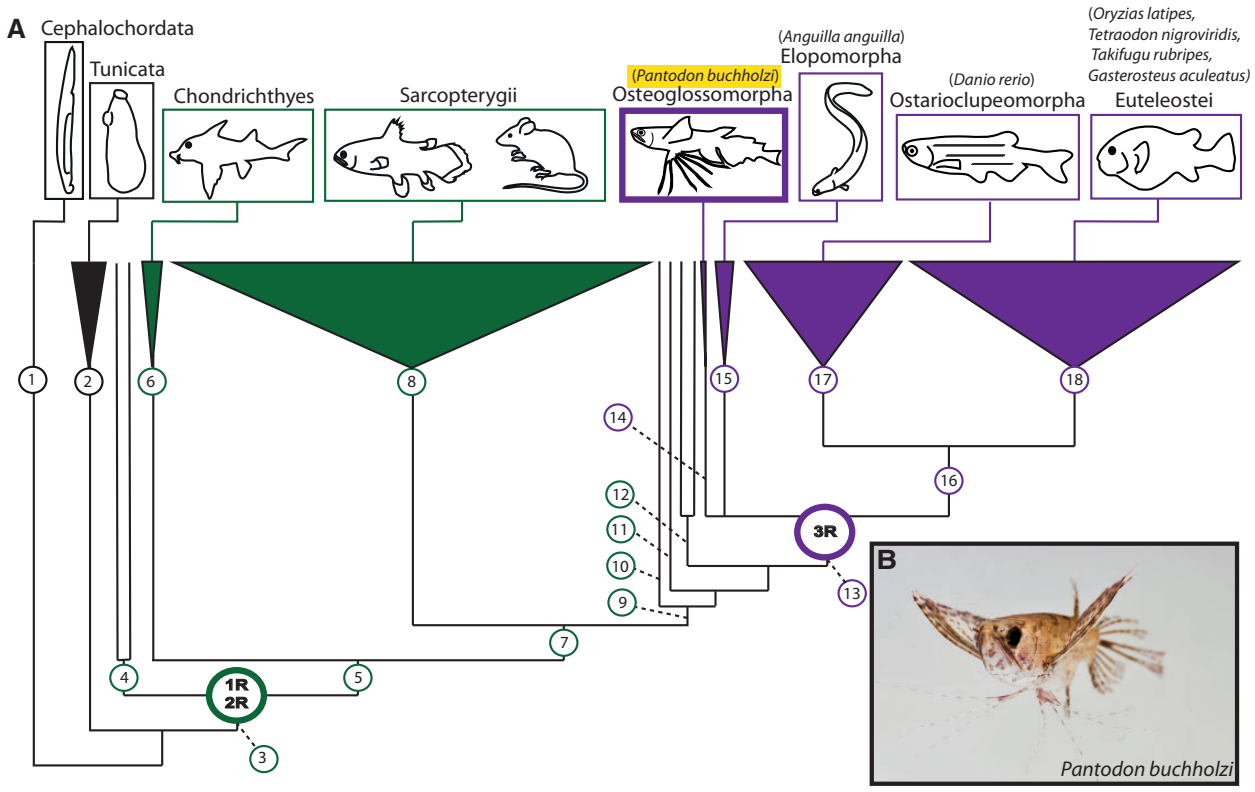



**Fig. 1.** WGDs and the key phylogenetic position of *Pantodon buchholzi*. (*A*) Cladogram depicting chordate phylogeny and biodiversity with the position of WGD events shown (1R/2R, TGD/3R). Osteoglossomorpha (Clade 14) branched from other teleosts near the root of the Teleostei (Clade 13) and is therefore a useful group for deep genomic comparisons of the TGD event with teleost species whose genomes have been well characterized within the Clupeocephala (Clade 16). Other major chordate clades are also illustrated for perspective. 1: Cephalochordata, 2: Tunicata, 3: Craniata, 4: Agnatha, 5: Gnathostomata, 6: Chondrichthyes, 7: Osteichthyes, 8: Sarcopterygii, 9: Actinopterygii, 10: Polypteriformes, 11: Acipenseriformes, 12: Holostei, 13: Teleostei, 14: Osteoglossomorpha, 15: Elopomorpha, 16: Clupeocephala, 17: Ostarioclupeomorpha, 18: Euteleostei. Representative species with available genomic data used in our analyses are listed. (*B*) Adult male specimen of the African freshwater butterflyfish, *P. buchholzi*.

labeled by the extensions "a" and "b" in teleosts, each containing one of a pair of duplicates is considered good evidence for a gene duplication event occurring prior to a speciation event. The frequency with which this phylogenetic tree topology is observed between duplicates located on chromosomal loci sharing synteny (paralogons) forms the backbone of the evidence for the TGD. The nomenclature of the eight clupeocephalan teleost *Hox* clusters was formed on this basis (hoxa*a*/hoxa*b*, hoxb*a*/hoxb*b*, hoxc*a*/hoxc*b*, hoxd*a*/hoxd*b*). The ability to reconstruct these orthology groups with phylogenetic methods relies on the shared derived mutations unique to members of the "a" and "b" clades and it therefore stands to reason that by the time the major clupeocephalan lineages diverged, the majority of their duplicates were already segregating as distinct paralogous loci. The timing of diploidization following the TGD relative to elopomorph and osteoglossomorph divergence, however, is less clear. Recent work in an elopomorph, the European eel (*Anguilla anguilla*), revealed eight *Hox* clusters and very few individual gene losses suggesting remarkable conservation of an ancestral post-TGD *Hox* complement arrangement, but noted difficulties in assigning 1:1 orthology relationships with clupeocephalan post-TGD "a" and "b" orthology groups (Henkel, Burgerhout, et al. 2012). A polymerase chain reaction (PCR) survey of *Hox* gene fragments in an osteoglossomorph, the Goldeye (*Hiodon alosoides*), has suggested that up to eight *Hox* clusters may be also be present in this species, but relied on isolated fragments as short as 82 bp to form hypotheses of homology (Chambers et al. 2009). Other studies have also found evidence for duplication of individual genes in osteoglossomorphs, which have been interpreted as originating with the TGD (Hoegg et al. 2004; Crow et al. 2006; Chambers et al. 2009); however in many of these cases, accurate homology relationships between the osteoglossomorphs genes and other teleosts were hindered by the low phylogenetic support for inclusion of the osteoglossomorph sequence within either the "a" or "b" post-TGD orthology groups, and the absence of any potentially corroborating physical linkage data.

We have generated the first extensive genomic data set from an osteoglossomorph fish for use in deep comparative analysis of the post-TGD diploidization process. Using bacterial artificial chromosome (BAC) pyrosequencing and Illumina whole-genome sequencing methods, we obtained all *Hox* gene sequences and their complete gene cluster organization in the African butterflyfish *P. buchholzi*. Our data reveal that in *Pantodon* extensive whole *Hox* cluster loss events are coupled with high individual gene retention rates, whereas phylogenetic analysis suggests that *Pantodon*

genes do not segregate as 1:1 orthologs with other teleost genes. We propose a model where following the TGD osteoglossomorphs diverged from other teleosts prior to complete genome diploidization and earlier than the fixation of the clupeocephalan teleost "a" and "b" orthology groups.

## Results

### Extensive *Hox* Cluster Loss in *Pantodon*

In order to investigate the evolutionary changes to genome structure in the immediate wake of the TGD we sought to characterize the *Hox* gene complement of the African butterflyfish *P. buchholzi*, a member of the early branching teleost subdivision Osteoglossomorpha. We hypothesized that if *Pantodon* shared the TGD with other teleosts it might have diverged before genomic changes ubiquitous to clupeocephalan teleosts arose and provide an earlier comparative genomic reference for inferring the condition of the ancestral teleost genome. We employed a three-tiered strategy involving degenerate PCR amplification, BAC clone pyrosequencing, and target-restricted assembly of Illumina whole-genome sequencing data to exhaustively screen the *Pantodon* genome for *Hox* genes.

PCR amplification from *Pantodon* genomic DNA with degenerate primers targeted to homeobox sequences of *Hox* and the physically linked *Evx/Eve* gene families resulted in the amplification of 34 unique *Hox* gene and 3 *Evx/Eve* gene homeobox fragments. These fragments were pooled and used to screen a custom-made *Pantodon* approximately 6× coverage BAC library with an average clone insert size of 138 kb. Ten *Hox*-positive BACs confirmed to contain all 37 PCR fragments were pooled to build one single-end library, which was fully sequenced by 454 to an estimated average depth of 30× (supplementary table S1, Supplementary Material online). A further 48 BACs isolated in our initial screens were pooled with the previous ten *Hox*-positive BACS to construct an additional 3-kb paired-end library, which was sequenced by 454 to an additional depth of approximately 8× (supplementary table S1, Supplementary Material online). Sequencing of both libraries yielded a sufficient amount of data for de novo BAC assembly which was performed by combining both data sets using Roche's newbler v2.6 software (supplementary table S2, Supplementary Material online). Careful manual annotation of the resulting scaffolds revealed that *P. buchholzi* possesses 45 intact full-length *Hox* gene coding sequences, interspersed with three mir-196 and four mir-10 family microRNAs in conserved positions (upstream *hox9* and *hox4* family genes, respectively), present on five separate scaffolds confirmed to encompass the ten original *Hox* BAC sequences by end sequencing (fig. 2). The five *Hox* gene clusters could be readily identified as one HoxA, two HoxB, one HoxC, and one HoxD cluster according to protein sequence similarity to the four ancestral gnathostome *Hox* gene clusters. On this basis, the five *Hox* clusters of *Pantodon* were tentatively named hoxa*x*, hoxb*x*, hoxb*y*, hoxc*x*, and hoxd*x* pending the phylogenetic evaluation of their homology with clupeocephalan teleost post-TGD "a" and "b" orthology groups.

Measured from the coding exons of the most 5′- and 3′-*hox* genes the approximate size of each *Pantodon* hox cluster is as follows: hoxax 55.2 kb, hoxbx 66.1 kb, hoxby 70 kb, hoxcx 93.2 kb, and hoxdx 45.4 kb. The size of the *Pantodon* genome has been estimated at 753 Mb (Hinegardner and Rosen 1972), intermediate between the genome sizes of zebrafish (1.5 Gb) and *Tetraodon* (342 Mb), the smallest vertebrate genome sequenced to date. The sizes of the hoxc and hoxd clusters of teleosts correlate well with the genome size and accordingly we find that the *Pantodon* hoxcx (93.2 kb) and hoxdx (45.4 kb) clusters are smaller than zebrafish hoxca (130.5 kb) and hoxda (52.6 kb) and larger than *Tetraodon* hoxca (69.3 kb) and hoxda (36 kb). The hoxcb cluster of zebrafish (72.2 kb) and the hoxdb cluster of *Tetraodon* (24.5 kb) are both reduced in size due to extensive gene loss not observed in *Pantodon* and are thus less relevant for comparison. The size of the hoxa and hoxb clusters of teleosts does not correlate as well with genome size and some *Tetraodon* clusters are larger than zebrafish. The *Pantodon* hoxax cluster (55.2 kb) is smaller than the hoxaa clusters of both zebrafish (58.2 kb) and *Tetraodon* (66.6 kb), but larger than their hoxab clusters (33.3 and 21.4 kb, respectively) which have experienced many gene losses. Both the *Pantodon* hoxbx (66.1 kb) and hoxby (70 kb) clusters are also smaller than the hoxba clusters of zebrafish (116.9 kb) and *Tetraodon* (148.8 kb) but larger than their hoxbb clusters (25.2 and 15.2 kb, respectively). Overall, the sizes of the hox clusters in *Pantodon* correlate with genome size in the case of hoxc and hoxd clusters, but not hoxa or hoxb clusters as observed in other teleosts.

Because five clusters is fewer than what has been found in any other teleost, we sought to guard against potential biases introduced by PCR amplification or BAC library construction and therefore sequenced the full genome of an individual male *Pantodon* specimen on one lane of Illumina using HiSeq2000 technology. This yielded 17.32 Gb of 100-bp paired-end (180 bp average insert size) reads corresponding to an estimated depth of 22.9× based on a genome size of 753 Mb (Hinegardner and Rosen 1972). We searched for *Hox* homeodomain sequences using the Target Restricted Assembly Method (TRAM), which allows targeted recovery and assembly of specific genomic loci from short-read sequencing data independent of whole-genome assembly (Johnson et al. 2013). The raw reads were used to build databases formatted for BLAST (Altschul et al. 1990) and HMMER (Söding 2005) searches which were queried using the collection of chordate homeodomain protein sequences contained in HomeoDB (Zhong et al. 2008) or the PFAM homeodomain hidden markov model profile (PF00046) using tBLASTn and hmmsearch algorithms, respectively, and reads were assembled using cap3 (Huang 1999). This method was specific enough to be able to independently assemble all 45 previously discovered *Hox* homeodomain-containing exon sequences at a read coverage level from 14× to 43×, and sensitive enough to recover homeodomain-encoding exons corresponding to members of all 11 currently recognized homeodomain classes. Two of these TRAM assembled exons appear to be fragments of *Hox* pseudogenes. A short fragment of a *hoxd2* gene, ψ*hoxd2x*, was subsequently found within a hoxdx BAC

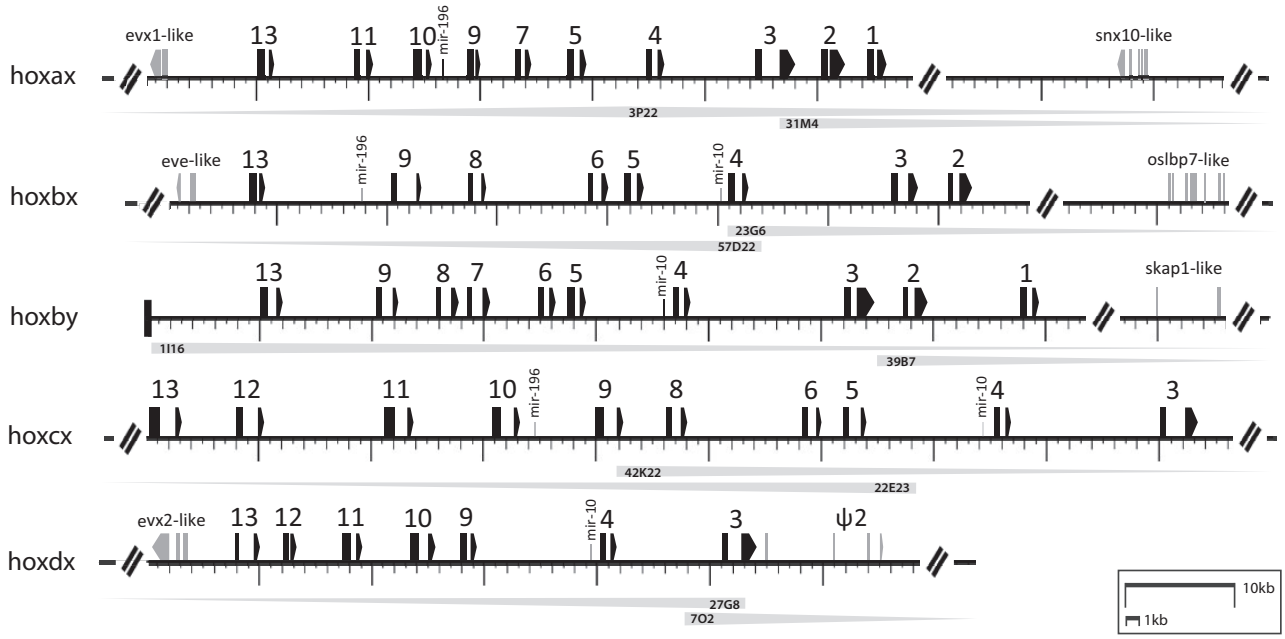



**Fig. 2.** The *Hox* gene complement of *Pantodon buchholzi*. With only five clusters (hoxax, hoxbx, hoxby, hoxcx, and hoxdx) containing a total of 45 full-length protein-coding *Hox* genes, *Pantodon* possess fewer *Hox* gene clusters than any other teleosts known, but retain a similar total number of individual genes. In black, exons of predicted *Hox* genes are depicted to scale alongside conserved microRNA genes in the mir-196 and mir-10 families, flanking *Evx* family genes and the most proximate nonhomeobox flanking genes or pseudogenes (gray boxes) we could identify in the region we sequenced. The IDs of BAC clones sequenced to assemble each of these scaffolds noted underlying each cluster. The total sizes of the assembled loci, and size of the region limited by the most 5′- and 3′-*Hox* gene-coding exon for each cluster, respectively, are: hoxax (199,116/55,231 bp), hoxbx (379,217/66,149 bp), hoxby (268,422/69,795 bp), hoxcx (246,971/93,161 bp), and hoxdx (254,542/45,374 bp).

(7O2) sequence, whereas the homeodomain-containing exon of a second *hoxa13* gene, ψ*hoxa13y*, could not be recovered from our BAC assemblies. No sequences with similarity to an additional *hoxa13* exon 1 sequence could be identified by TRAM. We infer that this is a remnant of a deleted cluster, Δhoxay. No further *Hox* homeobox sequences could be detected. Searches for additional members of the mir-10 and mir-196 families by TRAM also failed to recover any sequences in addition to those present within the five previously characterized clusters. The results from the *Pantodon* genome searches therefore strongly corroborate the conclusion that we had characterized the complete *Hox* complement of *P. buchholzi*. Hence, according to the prevailing pan-teleost TGD model, *Pantodon* has lost three complete gene clusters (one each of HoxA, HoxC, and HoxD) following the TGD. This is the most extensive case of *Hox* cluster loss following the TGD described to date, tripling the number of whole *Hox* cluster losses previously observed in any other teleost. It therefore seems that whatever constraint might exist to limit the number of *Hox* cluster losses in other teleosts does not apply to the *Pantodon* lineage.

Interestingly, despite the unprecedented number of whole *Hox* cluster losses, *Pantodon* has retained a total *Hox* gene complement (45 genes) similar to other teleosts (e.g., 49 in zebrafish, 46 in medaka). This is a consequence of the high proportion of duplicated genes retained in the hoxbx and hoxby clusters. Two members of the *hoxb13* family are retained in *Pantodon*, where at least one copy has been lost from the genomes of all other examined teleosts. Two members of the *hoxb2*, *hoxb4*, and *hoxb9* families, which have been reduced to only one member in all clupeocephalans examined but are retained in duplicate in the elopomorph *A. anguilla* (Henkel, Burgerhout, et al. 2012), are also kept in *Pantodon*. These additional genes, notable for their absence in the genomes of zebrafish, stickleback, medaka, and pufferfish were generally assumed to have been lost simultaneously with the TGD. We infer that they were lost specifically in clupeocephans following their divergence from osteoglossomorphs and elopomorphs, indicating a significant time lag between the TGD and many duplicated gene losses. Individual *hox* gene losses have been documented even several hundred million years following the TGD (Hoegg et al. 2007). The conservation of both the duplicated *hoxb2*, *hoxb4*, *hoxb6* and *hoxb13* gene family in osteoglossomorphs suggests that the ancestral teleost hoxb clusters were far more complete than previously thought from examining only clupeocephalans.

In addition to the three whole-cluster losses, *Pantodon* has experienced several individual *Hox* gene losses since diverging from the last teleost common ancestor, including *hoxb1x*, *hoxb7x*, *hoxb10x*, *hoxb10y*, *hoxc1x*, *hoxd8x*, and *hoxd1x*. We sought to evaluate to what extent these losses were shared with clupeocephalans suggesting rapid gene loss following the TGD, or independent suggesting a time delay between TGD and major changes to the teleost *Hox* gene complement. In order to ensure the same genes are compared in *Pantodon* and other teleosts, it is necessary to establish 1:1 orthology between the *Pantodon* genes and the either the

clupeocephalan counterparts belonging to either the post-TGD "a" or "b" orthology groups.

## Discordant *Hox* Gene Genealogies in *Pantodon*

We used phylogenetic methods to reconstruct unconstrained gene genealogies for each of the individual full-length *Hox* gene sequences of *P. buchholzi* to determine their orthology to clupeocephalan teleost *Hox* genes in the "a" and "b" orthology groups which formed following the TGD. The existence of more than four *Hox* gene clusters in *Pantodon* is consistent with a shared TGD model followed by the independent loss of three whole clusters in the *Pantodon* lineage. We therefore initially hypothesized that *Hox* genes in *Pantodon* would segregate with the canonical "a" or "b" post-TGD orthology groups recovered in clupeocephalan teleosts, whereas an unduplicated or independently duplicated sequence in *Pantodon* would segregate strictly outside of these clades. Because of their early divergence from other teleosts, osteoglossomorph sequences would by default be expected to segregate at the root of any teleost gene tree, potentially causing problems due to hidden paralogy where clupeocephalan teleosts retained only a single duplicate (Kuraku 2010). We therefore limited analyses to 26 orthology-informative *Pantodon* sequences (comprising 22 gene families) for which at least one clupeocephalan teleost with publicly available complete *Hox* gene complements retained both "a" and "b" paralogs. Two different phylogenetic reconstruction methods, maximum-likelihood (ML) and Bayesian Inference (BI), were used to evaluate the robustness of each gene genealogy and in order to minimize the potential impact of saturation due to the large amount of time since the TGD only amino acid sequences were used.

Contrary to our initial expectation, we found that many individual *Pantodon* gene tree topologies appeared to be inconsistent with our expectations under a shared TGD event (fig. 3; supplementary figs S1 and S2, Supplementary Material online). Rather than a single 1:1 orthology for each *Pantodon* gene with clupeocephalan genes belonging to either the TGD "a" or "b" orthology groups, several *Pantodon* genes segregated as an outgroup to, or in a polytomy with these clades. Furthermore, although we would expect genes which segregate within the "a" orthology group to be physically clustered together on a separate chromosome from those which segregate with the "b" orthology group, as found in other teleosts, we found that this pattern is disrupted in *Pantodon*. Even where strong phylogenetic support for 1:1 orthology of an individual *Pantodon* gene with members of either the clupeocephalan post-TGD "a" or "b" clades existed, adjacent genes yielded incongruent topologies such that within a single *Pantodon* cluster, genes grouping with the "a" and "b" orthology groups were interleaved, along with those which resolve best outside the combined TGD "a" and "b" clades entirely.

Examining the five orthology-informative genes on the *Pantodon* hoxax cluster revealed that four genes segregate with high support and good concordance between both ML and BI methods (bootstrap/posterior probability) with the post-TGD "a" orthology group: *hoxa13x* (76/1.0), *hoxa10x* (72/0.98), *hoxa9x* (98/0.58), and *hoxa2x* (99/1.0). Only *hoxa11x* segregated as an outgroup (64/0.82) to both the post-TGD "a" and "b" clades with the caveat that zebrafish *hoxa11b* broke the expected monophyly of the clupeocephalan "b" clade. We used the likelihood-based approximately unbiased (AU) test for alternative tree topologies (Shimodaira 2002) to evaluate three constrained topologies in which the *Pantodon* sequences were forcibly orthologized with either the clupeocephalan "a" (topology A) or "b" (topology B) post-TGD orthology groups or as an outgroup to the combined TGD "a" and "b" clade (topology O) (see fig. 4). All three constrained topologies enforced the respective monophyly of the post-TGD "a" and "b" orthology groups of clupeocephalans as well as the monophyly of the sequences used as unduplicated outgroups. Using this test, we were able to confidently ($P < 0.05$) rule out alternatives to orthology with the post-TGD "a" group (topology A) for three of five genes on the hoxax cluster: *hoxa13x*, *hoxa9x*, and *hoxa2x*. We were also able to rule out orthology of *hoxa11x* with the post-TGD "b" group (topology B), but not the possibility of an outgroup relationship (topology O). Only for *hoxa10x* was this test not able to rule out any of our three alternative hypotheses. Taking both the unconstrained phylogenies and constrained topology tests into account, we suggest that the hoxax cluster is a true 1:1 ortholog of the clupeocephalan hoxaa clusters. This is in line with our initial expectations for simple 1:1 orthology between *Pantodon* sequences and other teleosts due to a shared TGD event. The results from the remaining *Hox* genes, however, did not meet with our expectations under the prevailing model of the TGD.

Out of four orthology informative genes on the *Pantodon* hoxbx cluster two genes, *hoxb8x* (67/1.0) and *hoxb6x* (76/1.0) segregated with the post-TGD "a" orthology group, whereas *hoxb5x* grouped either weakly within a subset of clupeocephalan sequences belonging an atypically paraphyletic "a" clade (ML) or formed a polytomy with "a" and "b" clades (BI), and *hoxb3x* segregated best in an contradictory manner with the post-TGD "b" orthology group with both tree reconstruction methods, albeit with low support (24/0.52). Most alternative constrained topologies evaluated by the AU test were not significantly different from each other and only topology O for *hoxb8x* could be confidently rejected ($P < 0.05$).

Considering the hoxby cluster, with five informative genes, *hoxb6y* grouped reliably with the post-TGD "a" orthology group according to both methods (76/1.0), whereas both *hoxb8y* and *hoxb3y* segregated either as outgroups to or formed polytomies with the inclusive TGD "a" and "b" clades, and *hoxb1y* segregated with the post-TGD "b" orthology group (81/1.0). Curiously, *hoxb5y* grouped closely with *hoxb5x* among a clade of clupeocephalan sequences belonging an atypically paraphyletic clade comprising the nonpercomorph members of the post-TGD "a" orthology group (ML), or formed a polytomy with both "a" and "b" clades (BI). We also observe that both *hoxb6x* and *hoxb6y* also segregate best within the same clade comprising members of the clupeocephalan "a" orthology group. These irregularities are contrary to the default expectation under a shared TGD

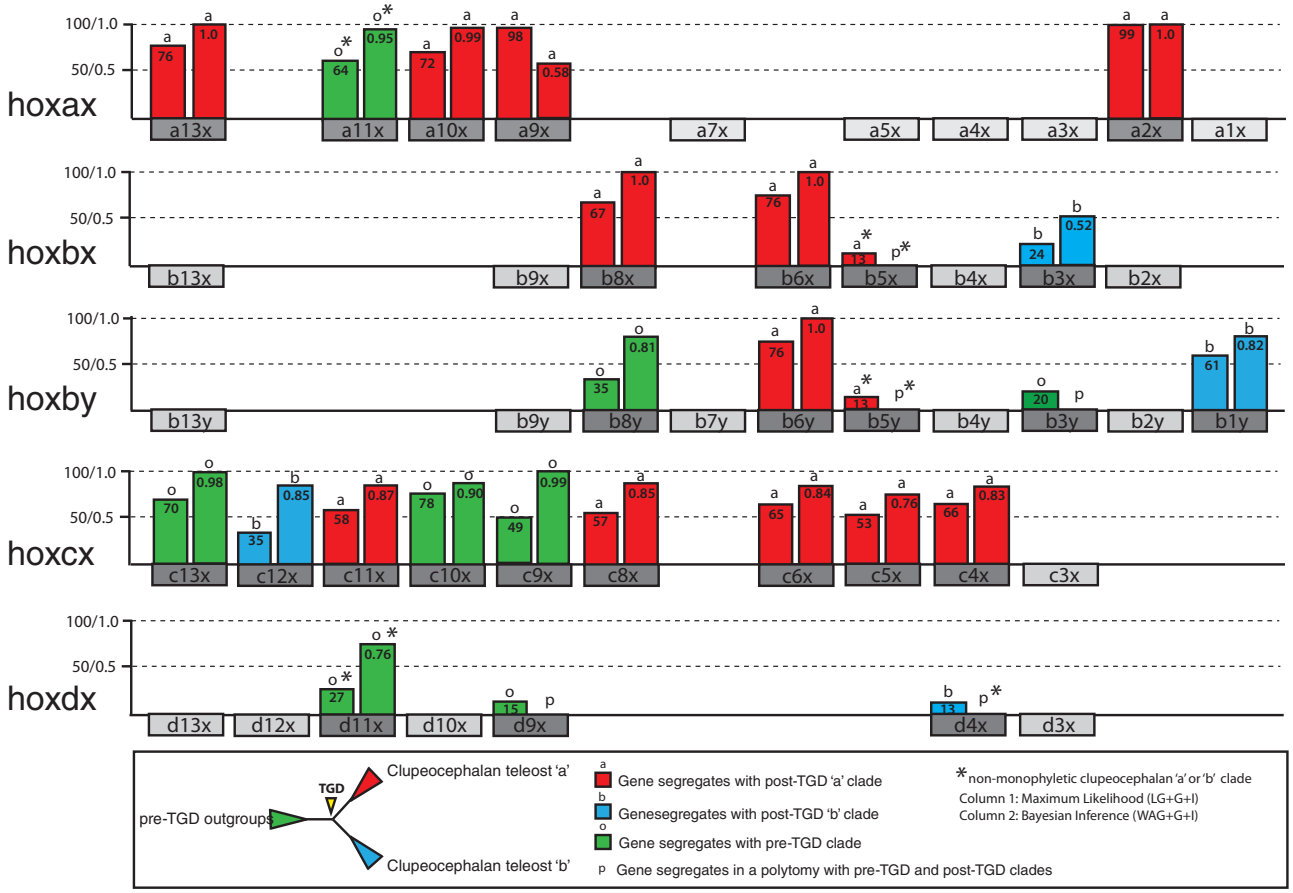



**Fig. 3.** The *Hox* genes of *Pantodon buchholzi* do not reliably segregate with the post-TGD "a" and "b" orthology groups of other teleosts in individual unconstrained vertebrate *Hox* gene trees. Phylogenetic trees were computed using ML and Bayesian methods with the *Hox* gene-coding sequences of *Pantodon*, other teleosts (eel, zebrafish, salmon, medaka, stickleback, *Tetraodon*, *Takifugu*, and *Astatotilapia*), sarcopterygians (coelacanth, *Xenopus*, *Anolis*, human, and mouse), and elephant shark. Support values corresponding to clades containing *Pantodon* sequences are plotted as two columns above each individual orthology-informative gene in the Hox cluster schematic. The height of each column corresponds to either the bootstrap support value (left column) or the posterior probability (right column). The column color corresponds to the clade containing the *Pantodon* sequence. We observe that across a single cluster, *Hox* genes where the *Pantodon* sequence clusters best with clupeocephalan post-TGD "a" orthologs (red), are interleaved with those which cluster best with the post-TGD "b" orthologs (blue), or outside the combined post-TGD "a" and "b" clades with nonteleost outgroups (green). Polytomies between the *Pantodon* sequence, the pre-TGD outgroups, and post-TGD "a" and "b" clades occurring in the Bayesian trees are indicated with a letter "p." Under a classic pan-teleost TGD model we would expect all *Pantodon* genes on each *Hox* cluster to segregate with either the "a" or the "b" orthology groups rather than a mixture, and no sequences which segregate best as an outgroup to the TGD node.

event which predicts that each *Pantodon* duplicate would belong to a single one of either the "a" or "b" orthology groups rather than both segregating together with only one. When we tested the statistical support for alternative constrained topologies we were able to eliminate topology O for *hoxb8x*, topology A for *hoxb8y*, and both topology A and O for *hoxb3y* which had appeared to segregate best as an outgroup to the inclusive TGD "a" and "b" clades in our unconstrained ML trees.

The hoxcx cluster has nine informative genes, five of which group with the post-TGD "a" orthology group: *hoxc11x* (58/0.87), *hoxc8x* (57/0.85), *hoxc6x* (65/0.84), *hoxc5x* (53/0.76), and *hoxc4x* (66/0.83). Three genes segregate as outgroups to the TGD "a" and "b" clades: *hoxc13x* (70/0.98), *hoxc10x* (78/0.90), and *hoxc9x* (49/0.99). Finally one gene segregated best with the post-TGD "b" clade: *hoxc12x* (35/0.85). The AU test was unable to rule out more than one of the three competing topologies for any gene of the hoxcx cluster, eliminating only topology B in the case of *hoxc10x*, *hoxc9x*, and *hoxc5x*.

Finally, within the hoxdx cluster, none of the three informative genes present segregated with high support in either the TGD "a" or "b" orthology groups. Although *hoxd4x* grouped best with clupeocephalan "b" genes in our ML analysis, bootstrap support was so exceedingly low (13) as to be considered unsupported. Of the other two informative genes, *hoxd11x* segregated as an outgroup to the TGD clade (27/0.76), whereas *hoxd9x* segregated either in an outgroup or in a polytomy inclusive of both TGD "a" and "b" clades. When the competing constrained topologies were evaluated through the AU test, none of the three topologies for any genes in the hoxdx could be rejected at a confidence level of $P < 0.05$.

The topological incongruences we observed between the individual *Pantodon* gene trees could be strictly interpreted as a contradiction of the prevailing hypothesis of a TGD event

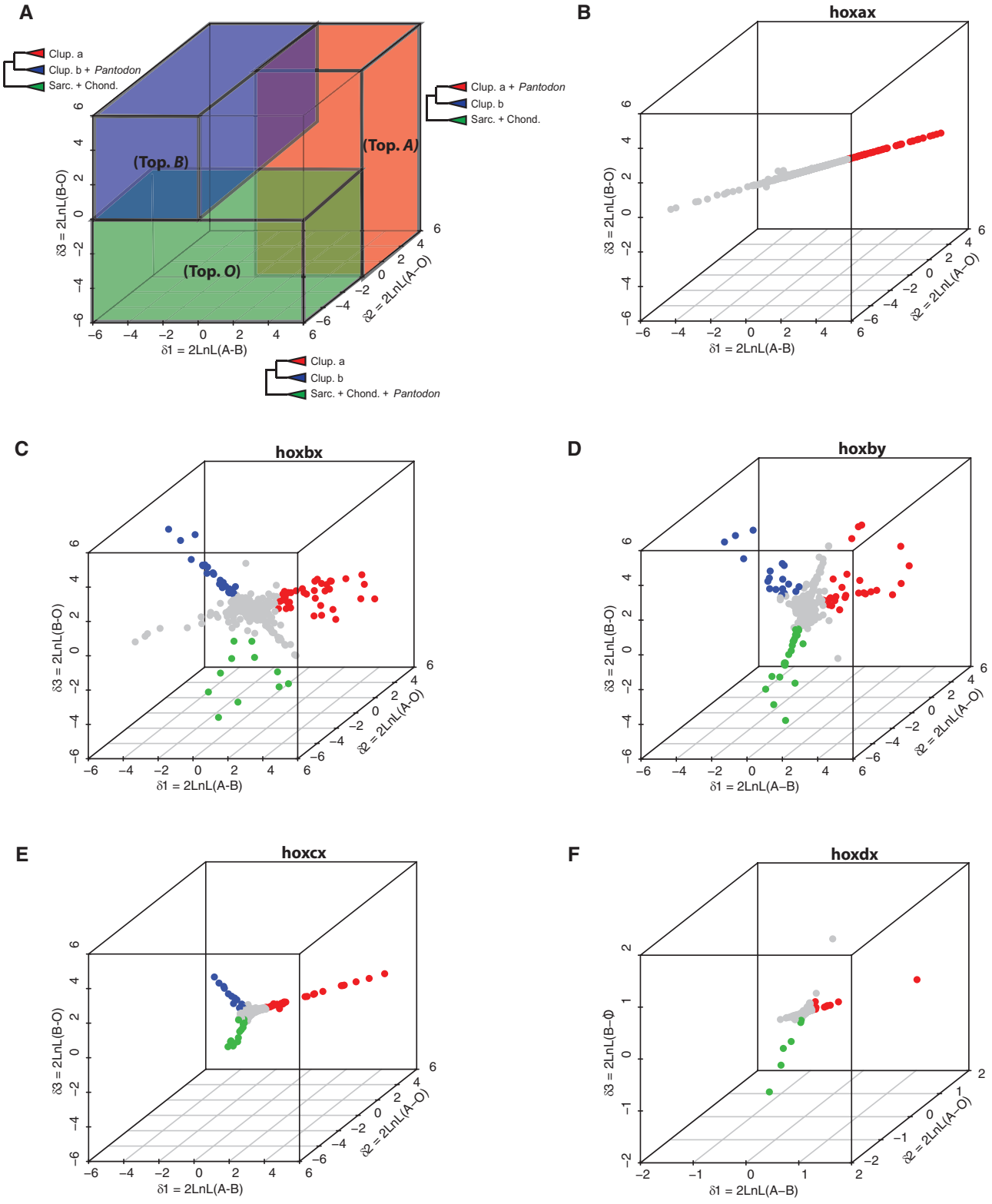



**Fig. 4.** 3D-SLRP using vertebrate whole Hox cluster concatenations reveals strong cluster-wide conflict between individual sites supporting each of three alternative hypotheses of *Pantodon* Hox cluster homology. Three different constrained tree topologies representing different hypotheses of *Pantodon* Hox cluster homology were compared. Topology A: ((post-TGD "a" + *Pantodon*)(post-TGD "b"))(pre-TGD outgroups), topology B:((post-TGD "a")(post-TGD "b" + *Pantodon*))(pre-TGD outgroups), and topology O:((post-TGD "a")(post-TGD "b"))(pre-TGD outgroups + *Pantodon*) were each compared in a pairwise fashion under a ML framework to model the support for each hypothesis of homology across each site in whole Hox cluster-concatenated alignments. (*A*) Schematized outline of a site-likelihood ratio plot showing regions in the graph which support each topology. Actual site-likelihood ratio plots for each *Pantodon* cluster are shown for hoxax (*B*), hoxbx (*C*), hoxby (*D*), hoxcx (*E*), and hoxdx (*F*). Each axis plots the site-wise likelihood ratio difference between one pair of competing topologies. The *x* axis plots the likelihood ratio between topology A and topology B (δ1), the *y* axis plots the likelihood ratio between topology A and topology O (δ2), and on the *z* axis the likelihood ratio between topology B and

(continued)

inclusive of osteoglossomorphs. However, both the presence of a duplicated cluster (hoxbx and hoxby) and the strong support for the inclusion of multiple individual *Pantodon* genes within the clupeocephalan teleost post-TGD "a" or "b" orthology groups (particularly those of the hoxax cluster) support the idea of a TGD event inclusive of *Pantodon*.

## Selection Is Not Sufficient to Explain Incongruent *Hox* Gene Genealogies

We sought an answer to the question of why the hoxbx, hoxby, hoxcx, and hoxdx clusters behaved atypically compared with the hoxax cluster in our phylogenetic analyses. It has been proposed that diversifying selection occurred in the immediate wake of the TGD between some gene duplicates, driving them apart, but not others (Crow et al. 2006). If this occurred between the genes in the duplicated hoxa clusters, but not in the hoxb, hoxc, or hoxd clusters this might cause stronger signal to accumulate between these duplicates and facilitate faithful reconstruction of their orthology. We tested for the presence of diversifying selection acting on the immediate post-TGD branches of each of the 22 orthology-informative *Hox* gene trees to compare the selective pressures on each of the duplicated *Hox* clusters under the assumption of an osteoglossomorph-inclusive TGD. Using constrained trees which enforced monophyly of the post-TGD "a" and "b" orthology groups of clupeocephalans but allowed the *Pantodon* sequence freedom, the MG94 codon substitution model (Muse and Gaut 1994) was used to estimate d*N*, d*S*, and $\omega$ ( = d*N*/d*S*) independently for each branch of the tree using a local model fit with the HyPhy software package (Pond et al. 2005). We found no evidence for different selective pressures acting immediately after the TGD between the hoxa gene trees and the hoxb, hoxc, or hoxd trees (supplementary table S3, Supplementary Material online). Only two branches arising directly from the TGD node (hoxa9a/hoxa9x and hoxc4b) had $\omega > 1$ ($\omega = 0.183/0.128 = 1.43$ and $\omega = 0.014/0.002 = 6.68$, respectively), consistent with positive selection. There were 15 additional proximate post-TGD branches in gene trees belonging to each of the hoxa, hoxb, hoxc, and hoxd clusters which returned infinite estimates for $\omega$ because d*S* was estimated at 0 under the topological constraints we imposed. In each of these cases, d*N* was estimated at $< 0.02$ making any inference of positive selection based on these results weak. Previous studies have also found evidence for constraint in the synonymous substitution rate of *Hox* genes which may inflate estimates of positive selection (Lin et al. 2008). Even if we were to take these results as acceptable evidence for positive selection, the prevalence of selection after the TGD node is no higher in the duplicated hoxa than in the hoxb, hoxc, or hoxd clusters.

Finally, because diversifying selection may be both episodic and affect only a small proportion of sites in an alignment, we tested for selection acting on only a small number of sites which might otherwise be lost in an overwhelmingly constrained sequence using the branch-site REL method implemented in HyPhy (Kosakovsky Pond et al. 2011). This method allows for three independent categories of $\omega$ estimates along each branch to accommodate site-wise variation in selective pressures, and does not require constraining the rest of the tree to a single d*N* and d*S* rate while testing for selection along one branch. Only one additional statistically significant instance of positive selection (Bonferroni corrected $P < 0.05$) could be found at the root of a post-TGD clade: A small proportion of sites ($q^+ = 4.44\%$) under positive selection could be found in the *hoxa2a/x* branch. We therefore believe that differences in selective pressure alone are therefore insufficient to explain the incongruent gene tree topologies in the *Pantodon* clusters.

## Heterogeneous Phylogenetic Signal within Concatenated *Hox* Clusters

We sought to explore additional explanations for the failure for the *Hox* gene genealogies of *Pantodon* to match the predictions of a shared TGD event. Given the relatively short sequence length of individual *Hox* genes and their high levels of conservation, we asked whether the incongruent topologies we recovered were the result of stochastic variation due to low signal intensity. The concatenation method has been useful in resolving phylogenetic problems where weak signal due to limited data has obscured underlying evolutionary relationships (Rokas et al. 2003). We built concatenated superalignments comprising all full-length protein-coding genes from each physically linked *Hox* cluster in *Pantodon* and a collection of other vertebrates. For each *Hox* cluster superalignment, we explicitly tested the collective orthology of the sequences within the *Pantodon* clusters to those of other teleosts by constructing trees for three competing constrained topologies and testing the statistical support for each using the AU test (Shimodaira 2002). As for individual genes, we constrained the *Pantodon* clusters explicitly within either the post-TGD "a" cluster (topology A) or "b" cluster (topology B) as would be expected under the canonical shared TGD, or outside the inclusive TGD "a" and "b" clades (topology O). The results of this test (summarized in supplementary table S4, Supplementary Material online) indicate that globally the *Pantodon* hoxax cluster is a 1:1 ortholog of the post-TGD "a" clusters of clupeocephalan teleosts, as

**Fig. 4.** Continued

topology O ($\delta$3) is plotted. Each point represents a single amino acid site. Sites are colored if the absolute magnitude of the corresponding site-likelihood ratio is more than 2 SD greater than the mean. Sites which support topology A are colored in red, sites which support topology B are blue, whereas sites which support topology O are green. Except for hoxax, which only contains sites which support topology A, there is conflicting phylogenetic signal in the *Hox* clusters of *Pantodon* which prevents unambiguous assignment to either the clupeocephalan teleost post-TGD "a" or "b" orthology groups.

expected for a shared TGD. Topologies B ($P = 2 \times 10^{-4}$) and O ($P = 2 \times 10^{-4}$) could be confidently rejected consistent with our results from individual gene genealogies. However, even when the signal of all informative *Hox* genes on a single cluster was concatenated, the results with the *Pantodon* hoxbx, hoxby, hoxcx, and hoxdx clusters still broke our assumptions of 1:1 orthology with other teleosts, as in our individual gene genealogies. For the *Pantodon* hoxby cluster and the hoxdx cluster none of the three competing topologies could be rejected, whereas for the hoxbx and the hoxcx clusters only topology O could be rejected ($P = 0.02$ and $P = 0.05$, respectively).

We employed a novel graphical method to visualize signal intensity and direction under three competing constrained tree topologies: 3D site-likelihood ratio plotting (3D-SLRP). We plotted the per-site likelihood ratios $\{\delta_s = 2 \ln[(\text{Site-likelihood Topology 1}) - (\text{Site-likelihood Topology 2})]\}$ which were calculated during the AU test between each pair of competing constrained topologies (A vs. B, A vs. O, and O vs. B) to obtain a 3D plot of the site-wise distribution of support for each of our three test topologies (fig. 4). Except for *Pantodon* hoxax, which shows a clear majority of sites uniquely supporting topology A (fig. 4*B*, red dots), the plots for the four remaining clusters show signal trifurcations in site-wise support for topologies A (red dots), B (blue dots), and O (green dots) (fig. 4*C*–*F*). Although the intensity of the trifurcating signals in hoxbx, hoxby, and hoxcx is as strong as the unidirectional signal in hoxax, in the case of hoxdx the signal is slightly weaker and this may have contributed to the difficulty in assigning orthology to this cluster. This heterogeneity in the signal for the full *Hox* cluster concatenations indicates that the failure of at least the *Pantodon* hoxbx, hoxby, hoxcx cluster concatenations to form clear 1:1 orthologous relationships with either the "a" or "b" post-TGD orthology groups in our AU tests is the result of conflicting signal between different sites in the *Pantodon* clusters, rather than an absence of signal. This outcome corroborates the results from the individual gene genealogies and supports the idea that the phylogenetic histories of individual sites/genes within the *Hox* clusters of *P. buchholzi* are heterogeneous.

### Ancient Conserved Noncoding Element Retention Patterns in *Pantodon* Hox Clusters

Identifying conserved noncoding elements (CNEs) is a useful method of identifying functional genomic elements, and many vertebrate CNEs found near developmental genes have been shown to play a role in gene regulation (Woolfe et al. 2005; Pennacchio et al. 2006). Genome-wide, ancient gnathostome CNEs (conserved between chondrichthyans and osteichthyans) appear to be evolving more quickly in teleosts than in other vertebrate lineages (Venkatesh et al. 2006; Lee et al. 2010). Earlier comparison of *Takifugu* and human Hox clusters with elephant shark *Callorhinchus milii* revealed that the duplicated *Takifugu* clusters retain fewer ancient CNEs than human and that all of the *Takifugu* post-TGD "b" clusters are depleted of ancient CNEs compared with the "a" clusters (Ravi et al. 2009). We performed VISTA alignments using the SLAGAN algorithm to identify ancient CNEs between elephant shark and the Hox clusters of *Pantodon* and five other teleosts (zebrafish, medaka, stickleback, *Takifugu*, and *Tetraodon*) (supplementary fig. S3 and table S5, Supplementary Material online). Our analysis reveals that although the post-TGD "b" clusters of all five clupeocephalan teleosts show a distinctive asymmetry in the retention of CNEs, the duplicated *Pantodon* hoxb clusters retain nearly the same number of CNEs (hoxbx:14, hoxby:15) (supplementary fig. S4*B*, Supplementary Material online). In clupeocephalans, the hoxbb cluster appears to have diverged more than the hoxba cluster having lost more CNEs (e.g., zebrafish hoxba: 26; hoxbb: 4, *Tetraodon* hoxba: 8, hoxbb: 3). The gene content of the clupeocephalan hoxbb clusters is also significantly lower than hoxba (e.g., zebrafish hoxba: 11; hoxbb: 4, *Tetraodon* hoxba: 9; hoxbb: 4). The symmetric pattern of CNE conservation in both *Pantodon* hoxb clusters is consistent with the higher degree of *Hox* gene retention (hoxbx: 8; hoxby: 10) and suggests a functional role for these CNEs, potentially in the regulation of *Hox* gene duplicates conserved in *Pantodon* but lost in other teleosts. This also suggests that the rapid structural evolution of the clupeocephalan hoxbb cluster only got underway after the speciation event separating *Pantodon* and other teleosts.

Ancient CNE retention patterns in the *Pantodon* hoxax, hoxcx and hoxdx clusters are more similar to clupeocephalan hoxaa, hoxca and hoxda clusters, respectively, than to hoxab, hoxcb or hoxdb which like hoxbb have experienced extensive *Hox* gene and CNE loss (supplementary fig. S4*A*, *C*, and *D*, Supplementary Material online). However, without corresponding duplicated clusters in *Pantodon* to compare with it is impossible to conclude whether the rapid structural evolution of the clupeocephalan hoxab, hoxcb and hoxdb clusters occurred before or after separation of the lineage leading to *Pantodon*. If asymmetric structural evolution only occurred in the clupeocephalan lineage after this speciation, as seems likely in the case of hoxbb, then presence or absence of CNEs is not phylogenetically informative.

## Discussion

We suggest a biological explanation for the unusual pattern of molecular evolution we observe that fits all of our data and is in line with our current understanding of genome diploidization dynamics in polyploids, but which has been largely overlooked in the study of ancient genome duplications. We propose that the four alleles of the duplicated hoxb, hoxc, and hoxd clusters did not complete diploidization into two pairs of alleles on separate paralogous gene clusters until after the very early divergence of osteoglossomorphs and clupeocephalans. In contrast, after the TGD the hoxa clusters resolved faster, prior to divergence of these lineages. This model implies that the last teleost common ancestor was a pseudotetraploid, like many present-day species which more recently underwent WGD, and had a composite genome consisting of both diploidized and tetraploid segments. Prolonged segregation as four alleles in a pseudotetraploid lineage would potentially subject the duplicated hoxb, hoxc,

and hoxd clusters to concerted evolution between duplicated loci, violating the assumptions of immediate genetic independence and simple 1:1 orthology of all teleost duplicates following the TGD.

One alternative model is that osteoglossomorphs experienced an independent WGD around the same time as the TGD, after diverging from other teleosts. Recently, an independent WGD event has been described in the American paddlefish (*Polyodon spathula*), a nonteleost actinopterygian fish, underlining the potential for undiscovered independent WGDs in other vertebrate lineages (Crow et al. 2012). However, this model does not explain the strong phylogenetic affinity for *Pantodon* hoxax with clupeocephalan hoxaa and, depending on the relative position of osteoglossomorphs and elopomorphs in the teleost tree, it might require yet another independent WGD in the elopomorph lineage. Likewise, independent duplication of the hoxb cluster in *Pantodon* is unlikely to have occurred as this would implicate additional independent duplications of the other clusters in the *Hiodon* lineage, which retains eight Hox clusters (Chambers et al. 2009), and does not fit our phylogenetic results. Another model might be that although osteoglossomorphs shared the TGD, the unprecedented loss of three whole Hox clusters in *Pantodon* constrained the evolution of the remaining clusters causing them to retain their similarity to their pre-TGD counterparts. While plausible, under this model we would expect that singular clusters in *Pantodon* (hoxax, hoxcx, hoxdx) would be more constrained and more difficult to homologize than clusters retained in duplicate (hoxbx/hoxby). However, what we observe is that the singular *Pantodon* hoxax cluster segregates well with the hoxaa clusters of other teleosts (which have never lost either of their hoxaa or hoxab clusters). Conversely, the pair of *Pantodon* hoxbx/hoxby clusters does not form clear 1:1 orthologies with other teleosts. Additionally, in clupeocephalan *Hox* genes we do not note any consistent difference in the ability of orphaned duplicates compared with replete pairs of genes in their ability to segregate in a single post-TGD orthology group. Finally, although our analyses suggest that selection on protein-coding genes was not significantly different between Hox clusters in the immediate wake of the TGD, and that signal trifurcation rather than absence was responsible for ambiguity in assigning 1:1 orthologies to the most of the clusters, it remains possible that forces operating at some as yet unidentified level are responsible for the phenomenon we observe here. The possible effect of long-branch attraction which can result in the erroneous grouping of quickly evolving branches in a phylogenetic tree by misinterpreting homoplastic characters as homologous changes should not be discounted. However, the generally high sequence conservation of *Pantodon* genes evident from selection analyses (supplementary table S3, Supplementary Material online) and the absence of strong statistical support for grouping the *Pantodon* sequences with quickly evolving branches such as the percomorph *hoxd4b*, *hoxd9b*, and *hoxd11b* (supplementary figs. S1*T*–*V* and S2*T*–*V*, Supplementary Material online) suggest that long-branch attraction is not a prominent factor in this case. The most plausible explanation we can give at this time remains the segmental and independent diploidization of the *Pantodon* genome following a shared TGD.

## Considering Diploidization Following the TGD

The finding that *P. buchholzi* has five *Hox* clusters including two HoxB loci is in agreement with the prevailing view that the TGD predates the divergence of the Osteoglossomorpha and hence is shared by all teleosts (Hoegg et al. 2004; Crow et al. 2006; Hurley et al. 2007). Based on observations in other teleosts, it is thought that the last teleost common ancestor possessed eight *Hox* clusters immediately after the TGD. This implies that three *Hox* clusters were lost in the *Pantodon* lineage (one HoxA, one HoxC, and one HoxD cluster). Teleost *Hox* cluster evolution appears to be significantly more dynamic than other gnathostome vertebrates which almost invariably retain the four ancestral *Hox* clusters with relatively few independent gene losses in different lineages (Liang et al. 2011). The only known exception to this stability are the elasmobranchs *Scyliorhinus canicula* and *Leucoraja erinacea* which appear to have lost their HoxC cluster after diverging from chimaerids >400 Ma (Ravi et al. 2009; Oulion et al. 2010; King et al. 2011). However, even among teleosts which have seen the zebrafish, and salmon lineages independently lose their hoxdb clusters (Woltering and Durston 2006; Mungpakdee et al. 2008), and percomorphs including medaka, pufferfish, and cichlids lose their hoxcb clusters (Amores et al. 2004; Tümpel et al. 2006; Hoegg et al. 2007), the loss of the hoxay, hoxcy, and hoxdy clusters in *Pantodon* stands out as the most extensive case *Hox* cluster loss currently known.

The presence of the duplicated HoxB cluster alone, however, is not strictly sufficient to conclude that *Pantodon* shared the TGD and lost three clusters. Our phylogenetic analyses revealed that most individual genes within the *Pantodon* hoxax cluster, and the whole cluster considered together in a concatenation of its component genes, groups with very good support with the hoxaa clusters of clupeocephalan teleosts exactly as one would expect if *Pantodon* shared the TGD with clupeocephalans and subsequently lost the hoxay cluster. This also implies that by the time *Pantodon* diverged from other teleosts, the duplicated HoxA clusters were already segregating as fully diploidized paralogous loci. However, the interleaving phylogenetic affinities of individual *Hox* genes within the *Pantodon* hoxbx, hoxby, hoxcx, and hoxdx clusters for incongruent topologies, considered together with the trifurcation in the support of individual sites within these clusters for the post-TGD "a" or "b" orthology groups, or outgroup topologies, contradict the predictions for a shared TGD event if we assume that all *Hox* clusters concluded diploidization simultaneously prior to the radiation of major teleost subdivisions.

These apparently paradoxical findings can be reconciled by postulating a protracted pseudotetraploid period in teleost evolution, following the ancient TGD polyploidization event and preceding the establishment of the fully diploid (paleopolyploid) karyotype of modern teleosts. If *Pantodon* diverged from other teleosts during this pseudotetraploid

period, prior to the conclusion of the diploidization process, it is likely that it shared the diploidization of some genomic loci prior to speciation, but not others. The effects of homeologous recombination and gene conversion between the highly similar tetraploid alleles of duplicated but undiploidized loci, as well as the potential for independent sorting of alleles during the speciation process, could have led to the divergent phylogenetic pattern which we observe if osteoglossomorphs diverged early enough that diploidization was concluded independently from other teleosts. If these genetic homogenizing forces were strong enough they would have the effect of slowing down or even stopping the molecular clock, preventing differences between duplicates from becoming fixed and effectively delaying their divergence until after diploidization had concluded. If this continued until after the deep divergence between osteoglossomorphs and clupeocephalans, duplicates in *Pantodon* would have separate coalescents from other teleosts marking the independent cessation of recombination, and would fall as outgroups to the clupeocephalan teleost "a" and "b" clades, as we observed in many cases (e.g., *hoxb8y*, *hoxc13x*). Likewise, a weaker degree of genetic exchange between homeologous loci might allow some mutations between duplicates to become fixed, but subsequently cause them to swap their physical position between homeologous chromosomes, resulting in the interleaving pattern we observe (e.g., *hoxb8x*, *hoxb3x*). Nonreciprocal homeologous recombination (gene conversion) may also occur resulting in the segregation of two duplicates within a single orthology group (e.g., *hoxb6x*, *hoxb6y*). It may be impossible to mark a definitive point in time when genetic exchange between duplicated homeologous loci ceases completely. It is likely that homeologous loci diverge progressively, undergoing genetic exchange less and less frequently with time until finally it becomes so infrequent as to be effectively nonexistent. Using lineages which speciated close to the TGD event, we were able to use phylogenetic data to date the relative order that duplicated genomic segments achieved complete independence. With this method we can infer that the duplicated hoxa clusters diploidized first, prior to the speciation event separating osteoglossomorphs and clupeocephalans, whereas the hoxb, hoxc and hoxd clusters most likely concluded diploidization independently after divergence of these two lineages. Our model of teleost *Hox* cluster evolution which gives an account of the relative timing of diploidization of the different clusters in addition to detailing individual gene and whole cluster loss patterns is outlined in figure 5.

## Diploidization Dynamics in Other Polyploids

The types of complex genetic interactions between duplicated homeologous loci which are possible after polyploidization are best documented in plants, where allopolyploidy in particular is a pervasive phenomenon (Otto and Whitton 2000; Wendel 2000; Doyle and Egan 2010). The idea that sequence similarity between duplicated homeologous chromosomes might linger in allopolyploid plants causing tetrasomic inheritance was first proposed over 50 years ago (Stebbins 1947). Recent work in plants including artificially resynthesized allopolyploid lines in the genus *Brassica* (Gaeta et al. 2007), a natural assemblage of very recently (<80 years) formed allopolyploid species in the genus *Tragopogon* (Chester et al. 2012) and older (1–2 My) allopolyploid lineages such as the cotton *Gossypium hirsutum* (Salmon et al. 2010) has revealed molecular evidence for extensive genetic exchange between duplicated homeologous loci. In contrast to autopolyploidy where intraspecific genome doubling provides ample opportunity for tetrasomic inheritance, allopolyploidy comes in two categories which differ in the extent of genetic exchange expected between homeologous chromosomes: "Genomic" and "segmental" allopolyploidy (Stebbins 1971). Genomic allopolyploids result from the interspecific hybridization of two highly divergent parental species genomes causing homeologous chromosomes to exhibit almost immediate bivalent segregation and disomic inheritance. In these allopolyploids the genome is effectively prediploidized and the coalescent time of duplicated genes predates the date of hybrid formation, and instead corresponds to the speciation date of the parental species. Segmental allopolyploids, however, result from hybridization where only a fraction of the parental genomes is divergent enough for immediate bivalent formation, and other regions continue to exhibit tetrasomic inheritance and eventually must diploidize. Maize (*Zea mays*) which experienced a polyploidization event approximately 11.4 Ma seems to be a segmental allotetraploid in origin (Gaut and Doebley 1997). Our results are consistent with the ancient TGD being either an autopolyploidy event where the duplicated HoxA clusters diploidized before the HoxB, HoxC, and HoxD clusters, or a segmental allopolyploidy event in which the HoxA clusters of the parental species were effectively prediploidized immediately upon hybrid formation, but strongly exclude genomic allopolyploidy as an option.

Much less is known about the process of diploidization after ancient WGDs in animal lineages, especially in vertebrates. Tetravalent formation during meiosis, which favors homeologous recombination and gene conversion, has been directly observed in recently polyploidized lineages including autotetraploid frogs (*Odontophrynus americanus*) (Becak et al. 1966) and allotetraploid loaches (*Misgurnus anguillicaudatus*) (Li et al. 2011). Interestingly, salmonid teleosts, which underwent an additional independent WGD (sometimes called 4R) after the TGD between 25 and 100 Ma (Allendorf and Thorgaard 1984), still do not seem to have completed the diploidization of their genomes, as the recent analysis of the complete genome of the rainbow trout has demonstrated (Berthelot et al. 2014). Furthermore, there is evidence in the salmonid lineage for ongoing recombination between duplicated loci (Johnson et al. 1987; Allendorf and Danzmann 1997; Gharbi et al. 2006) and for occasional tetravalent formation during meiosis (Lee and Wright 1981; Allendorf and Thorgaard 1984). If the divergence of osteoglossomorphs and clupeocephalans occurred relatively rapidly following the TGD, it would therefore be conceivable that the diploidization process was only partially complete, and concluded independently following their divergence. The earliest recognizable fossils of the major teleost subdivisions

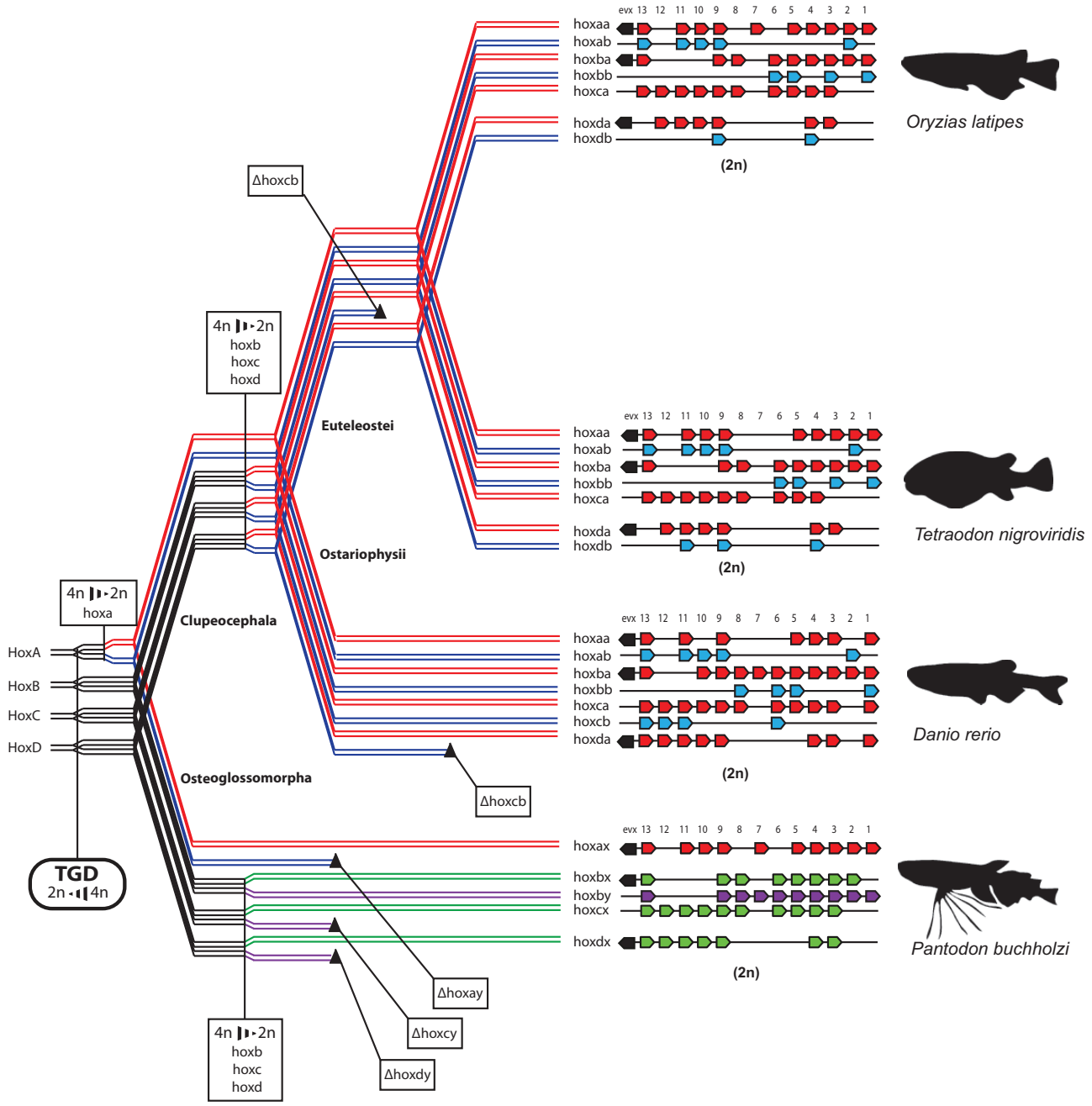



**Fig. 5.** The evolution of teleost *Hox* gene clusters following the TGD outlining the relative timing of diploidization events and the speciation of major teleost subdivisions. This model of Hox cluster evolution in teleosts illustrates the independent diploidization of *Hox* clusters following the TGD and the relative timing of cluster diploidization and speciation events. Each line in the phylogram represents an allele and the separation of pairs of lines accompanied with a change in color represents the completion of the diploidization of this locus. In this model, the duplicated hoxa clusters diploidize first, before the last common ancestor of all teleosts. The remaining clusters diploidize later, and independently in Clupeocephala and Osteoglossomorpha. Whole *Hox* cluster losses (black triangles) are also mapped highlighting the massive Hox cluster losses in the *Pantodon* lineage.

(Osteoglossmorpha, Elopomorpha, Clupeocephala) only appear near the Jurassic–Cretaceous boundary approximately 150 Ma (Arratia 1997). Molecular estimates for the divergence time of crown teleosts vary widely according to whether mitochondrial (268–326 Ma) or nuclear (181–265 Ma) data are used (Hurley et al. 2007). Estimates for the timing of the TGD based on dating the divergence of paralogous genes in clupeocephalan teleosts range between 320 and 350 Ma (Taylor et al. 2001; Christoffels et al. 2004; Vandepoele et al. 2004). Using the median divergence time estimates for the teleost radiation and the TGD, respectively, we arrive at a very rough approximation of 38 My (mitochondrial) to 112 My (nuclear) between the TGD and the divergence of the major teleost subdivisions. Because of the strong dependence of the molecular clock used to make these estimates on the diploidization process itself, the relative dating of the TGD and teleost diversification may become somewhat confounded. However, even with this caveat in mind most authors

concur it is likely that the period of time between the TGD and the teleost radiation was relatively brief. This brevity could conceivably have allowed for speciation to begin before diploidization had terminated.

## Tetralogy: A New Homology Subtype

We suggest that the classical concepts of orthology and paralogy are not adequate to describe the homology relationship between gene duplicates in two species which originated from a single duplication event, yet do not share a coalescent as a result of independent diploidization following speciation. We recommend the use of novel terminology for this subtype of molecular homology which will serve to quickly and accurately describe it: Tetralogy.

Under the classical model (single-WGD, shared diploidization), relationships between genes in two species which experienced a shared duplication (and shared diploidization) event prior to speciation resolve unambiguously into either 1:1 orthologs which descend from the same gene present in their last common ancestor, or paralogs which are descendants of two different duplicates in their last common ancestor (fig. 6*A*). If two separate duplication events of the same gene occurred independently following speciation (two-WGD model), 1:1 orthologs in the descendant species could not exist because the duplicates originated from two events and two distinct ancestral genes, and only paralogous relationships would exist (fig. 6*B*). In our model (single-WGD, independent diploidization) where two species share a duplication event but where duplicates diploidize into genetically independent loci from a shared allelic pool independently in the two lineages (fig. 6*C*), none of the resultant genes can be considered 1:1 orthologs because they do not share a single common ancestral sequence. However, they share an important property with the orthologous duplicates generated under the single-WGD shared diploidization model depicted in figure 6*A* in that they fundamentally originate from the exact same duplication event. This distinguishes them from the paralogous duplicates generated in the two-WGD model depicted in figure 6*B*. We use the term tetralogy to describe this unique type of 2:2 orthology relationship. Because the topology of individual gene trees resulting from independent diploidization is indistinguishable from those produced by independent duplication, only by examining multiple loci and syntenic genes can independent duplication be conclusively ruled out and tetralogy as a result of independent diploidization be inferred. When we compare phylogenetic trees of multiple genes across the *Hox* clusters of *Pantodon* and other teleosts, what we find is a combination of topologies. Some genes segregate in ways consistent with both the independent diploidization and independent duplication scenarios, whereas others segregate exactly as one would expect for a shared duplication and diploidization scenario. We can therefore infer that *Pantodon* did not experience a separate genome duplication event but rather that independent diploidization is likely to have occurred.

Because consistent *Hox* gene nomenclature which reflects both homology and physical position is needed, we advocate retaining the suffixes "*x*" and "*y*" for all *Hox* genes in *Pantodon*, including those individual genes which segregate within clupeocephalan "a" or "b" orthology groups but are part of a cluster which does not otherwise share unanimous 1:1 orthology with other teleost clusters. We also retain the naming system for members of the hoxax cluster mainly for consistency.

## Applications of the Tetralogy Concept to Outstanding Questions in Vertebrate Evolutionary Genomics

One area where the tetralogy concept may have profound implications concerns the outstanding question of whether osteoglossomorphs, elopomorphs, or a clade containing both is the sister clade to all other teleosts. Analysis using full mitochondrial genomes supports Osteoglossomorpha as the sister group to other teleosts (Inoue et al. 2001, 2003), whereas a more recent supertree analysis of nine nuclear gene exons suggests that Elopomorpha is the sister group (Near et al. 2012). Even some support for a combined (Osteoglossomorpha + Elopomorpha) sister clade to other teleosts has been found using ribosomal RNA genes (Le et al. 1993). In order for gene trees to have any hope of accurately reflecting real underlying evolutionary relationships between species, it is absolutely necessary to limit phylogenetic analysis to 1:1 orthologs. The problem of hidden paralogy, the mistaken use of paralogs rather than orthologs in phylogenetic trees as a result of a gene loss, is a known source of bias in phylogenomic analysis (Kuraku 2010). The prospect of tetralogous genes, that is genes which fundamentally do not have an unbroken lineage of 1:1 orthology between all species sharing a WGD event, adds a further complication to species tree reconstruction. In principle, any genes which cannot be shown to share a clearly orthologous relationship between species, including tetralogous genes, should not be used in phylogenomics. Although unduplicated mitochondrial genomes are not expected to be fraught with the same problems, all nuclear genes in teleosts used for phylogenomics should first be screened for potential hidden tetralogy, and only those loci which form well-supported orthologous relationships should be used. Hidden tetralogy will most likely only affect the phylogenetic relationships between species which diverged close to a WGD event and may not have experienced complete communal diploidization. Our results suggest that Osteoglossomorpha, Elopomorpha, and Clupeocephala may fit this criterion, and hidden tetralogy between phylogenetic markers of nuclear origin may be one reason why the relationship between these clades remains largely enigmatic.

A second outstanding controversy to which the concept of tetralogy may apply concerns the uncertainty surrounding the relative timing of the vertebrate 1R/2R WGD events to the divergence of agnathans and gnathostomes (Kuraku 2008; Shimeld and Donoghue 2012). Extant agnathans, comprising hagfish and lampreys, have many duplicated genes in similar proportions to gnathostomes, but when phylogenetic trees are constructed clear 1:1 orthologies with gnathostome genes are often absent (Escriva et al. 2002; Irvine et al. 2002; Stadler

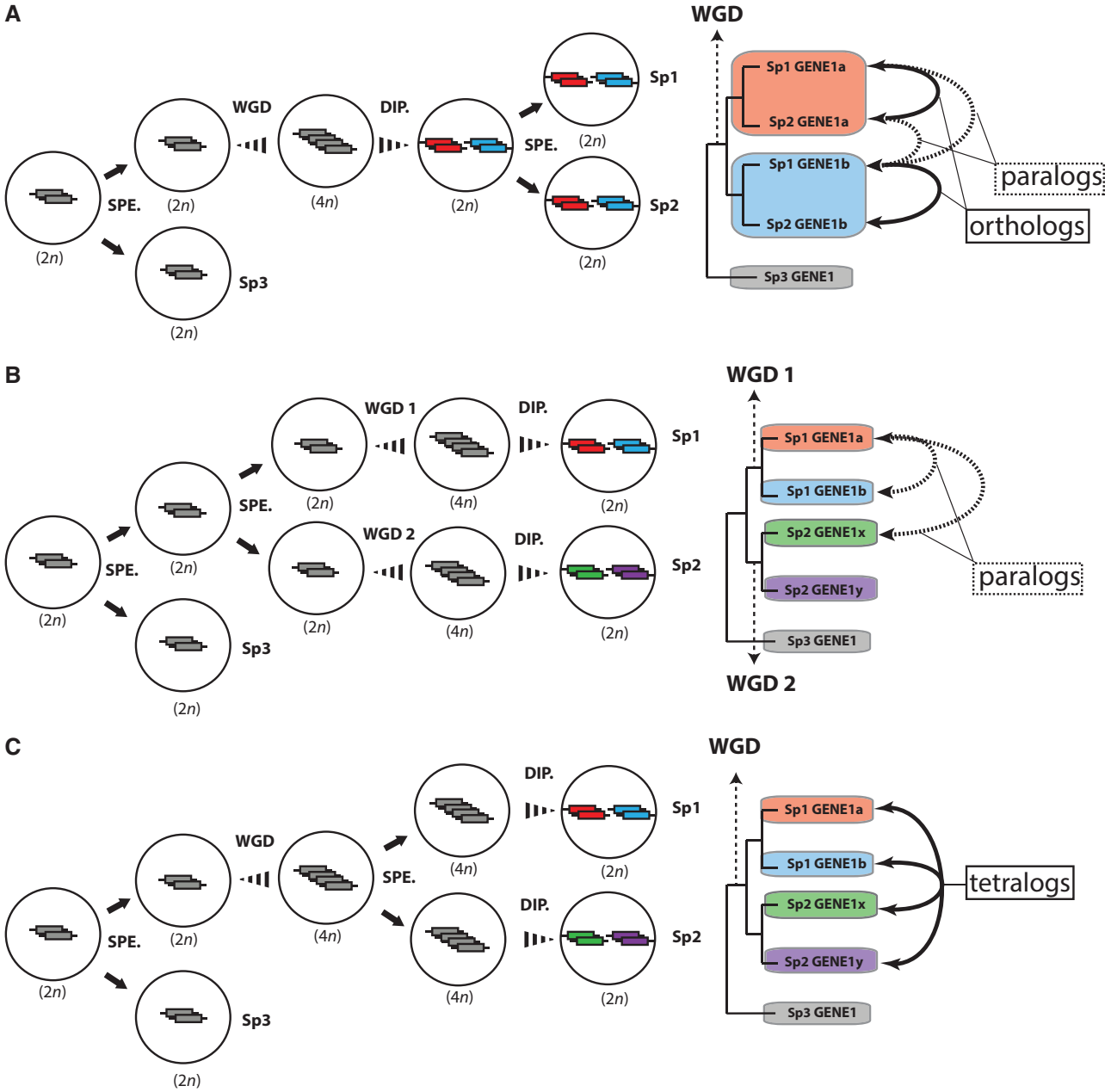



**Fig. 6.** Types of homology relationships following WGD. Schematic outlining the types of homology relationships which can exist between gene duplicates as a result of WGD, taking into account the relative timing of diploidization. (*A*) The classical model where both WGD and full diploidization (DIP.) occur before speciation (SPE.) results in two types of homology relationships: Orthology (e.g., Sp1 GENE1b and Sp2 GENE1b) and paralogy (e.g., Sp1 GENE1a and Sp2 GENE1b). (*B*) The classical alternative scenario to the model presented in (A), where speciation occurs before two independent WGD and subsequent diploidization, only results in the formation of paralogs (e.g., Sp1 GENE1a and Sp2 GENE1x). (*C*) A new scenario where a single WGD occurs prior to speciation but where diploidization occurs independently in each lineage can result in a novel type of homology relationship: Tetralogy (e.g., Sp1 GENE 1a and Sp1 GENE 1b are tetralogous with Sp2 GENE 1x and Sp2 GENE 1y). Following duplication in a diploid (2*n*) ancestor, and the resultant tetraploid (4*n*) speciates prior to full diploidization, recombination will cease independently and loci will make the transition from 4*n* alleles to 2*n* paralogs separately in each lineage. Shared derived mutations between will then be able to accumulate independently between each pair of duplicates in each lineage, and no single duplicate in one lineage can be considered orthologous with a single duplicate in the other.

et al. 2004; Qiu et al. 2011; Fujimoto et al. 2013). Explanations to account for this have included agnathans and gnathostomes sharing only one genome duplication (plus additional duplications in each lineage), or multiple independent duplications in each lineage. Analysis of the a draft version of the sea lamprey (*Petromyzon marinus*) genome suggested that both 1R and 2R were likely shared between agnathans and gnathostomes, largely based on shared synteny (Smith et al. 2013). However the most thorough study of *Hox* genes in lamprey to date has found evidence of at least six Hox clusters implicating at least one independent round of genome duplication after 1R/2R in the lamprey lineage, and casting doubt on whether even 1R/2R was shared based partially on absence of 1:1 orthology of any lamprey genes with

gnathostomes (Mehta et al. 2013). Our data suggest that, at least in the case of the TGD, gene trees do not necessarily resolve into clear orthology groups even following a single-shared genome duplication event and that the expectation of 1:1 orthology following a shared WGD may be unwarranted. One might speculate that if the ancestral vertebrate went through a prolonged tetraploid (or octoploid) stage with some regions of the genome diploidizing before divergence of agnathans and gnathostomes, and others diploidizing independently in each lineage, it could lead to the same type of incongruent phylogenetic patterns we observed in this study. Of course, these explanations are not mutually exclusive and it is possible that a combination of independent and shared genome duplications and diploidization events occurred.

## Conclusion

We have demonstrated that the *Hox* gene clusters of *P. buchholzi* represent a unique departure from the situation seen in other teleost fishes. *Pantodon* has fewer *Hox* gene clusters than any other teleost (five), but a similar total *Hox* gene number (45). We also observe that phylogenetic analysis of the *Hox* clusters of *P. buchholzi* fails to support a shared TGD hypothesis unless a model accounting for diploidization after duplication is considered. Under this model we propose that after a shared TGD, there was gradual and segmental diploidization of the genome which concluded independently in osteoglossomorphs and clupeocephalans, and that the last teleost common ancestor was therefore likely a pseudotetraploid. We argue that *Pantodon* shared TGD with other teleost fish, but that most of its *Hox* gene clusters (hoxbx/hoxby, hoxcx, hoxdx) loci violate assumptions of 1:1 orthology with clupeocephalan teleost *Hox* clusters as a result of independent diploidization. Therefore, instead of all duplicates which originated at the TGD showing clear 1:1 orthology across teleosts, some may share a novel type of 2:2 orthology relationship as a result of independent diploidization: Tetralogy. The prevalence of this phylogenetic pattern in other gene families across the genome of *Pantodon* and related osteoglossomorphs will be of significant interest and can be tested once additional genomic resources become available.

## Materials and Methods

### Genomic DNA Isolation

Adult specimens of *P. buchholzi* were obtained from retailers in the United Kingdom. Genomic DNA was prepared from muscle by standard proteinase K digestion and phenol-chloroform extraction.

### Isolation of Hox and Evx/Eve Probes by PCR

Degenerate primers (SO1 5′-garytngaraargartt-3′ and SO2 5′-cknckrttytgraacca-3′) were used to amplify *Hox* and related ANTP class homeobox-containing genes in *Pantodon*. Additional primers were designed to target-specific vertebrate *Hox* families PG1, PG5, PG9, and PG13: (Hox1SO1 5′-acvgarytngaraargartt-3′ and Hox1SO2 5′-catnckncкrttytgraacca-3′), (Hox5F1 5′-gcntayacbcgytaycagac-3′, used with SO2), (Hox9F1 5′-tgyccytayacnaaryayca-3′, used with SO2), (Hox13F1 5′–aarmgnrtnccntayasnaa-3′ and Hox13SO2 5′-acnckbckrttytgraacca-3′). Primers (EvxF 5′–yaycgngcnttcacvmgnga-3′ and EvxR 5′-ckctgnckyttgtccttcat-3′) were used to amplify *Evx* family genes. Amplified bands resulting from gradients of annealing temperature from 45 °C to 55 °C and 0.5–5.0 mM $MgCl_2$ were gel extracted and cloned in the pGEM-T vector. Sanger sequencing was performed using a standard BigDye protocol. In excess of 200 clones were sequenced identified by similarity to other vertebrate genes by BLAST against the NCBI nr database online.

### BAC Library Screening

An approximately 6 × BAC library for *P. buchholzi* containing 27,648 clones with an average insert size of 138 kb was prepared by Amplicon Express from genomic DNA extracted from muscle tissue of five mixed sex individuals. Digoxigenin (DIG)-labeled probes were synthesized by PCR from cloned *Hox* and *Evx* gene fragments. Five pools of *Hox* probes and three separate *Evx* probes were synthesized. Arrayed BAC library filters were hybridized under relaxed stringency conditions (35–42 °C in Easy Hyb buffer), and washed with 2× SSC, 0.1% SDS, then 0.2 × SSC, 0.1% SDS (25–42 °C), and finally in wash buffer (0.1 M maleic acid, 0.3% Tween, 25–42 °C). Detection used anti-DIG fab fragment and CDP Star. Two library screens, one with the 35 combined *Hox* probes and one with the three *Evx* probes, were performed and the results yielded 82 candidate hox cluster BACs which were then midi-prepped and screened by PCR for the presence of *Hox* or *Evx* genes.

### Isolation and Sequencing of BAC Clones

DNA from ten BACs comprising the full diversity of amplified probes (57D22, 23G6, 39B7, 1I16, 27G8, 7O2, 3P22, 31M4, 22E23, and 42K2) was pooled and used to prepare a single-end library for Roche 454 pyrosequencing; one-quarter plate gave targeted sequence coverage of approximately 30 × , and is referred to Library "E." The additional 48 of the original 72 candidate BACs isolated by hybridization, but which did not yield known *Hox* or other homeobox gene fragments by PCR, were prepared for sequencing as eight pools of separate cultures of six BACs each. These BAC pools consisted of pool A (1E1, 2H15, 3I22, 5P8, 10E24, 10M4), pool B (14D16, 16K5, 17M5, 19D4, 19N18, 21J10), pool C (22I13, 22A23, 26A20, 28C18, 29I4, 30I11), pool D (33B5, 33D19, 37H21, 39J24, 41C14, 41B1), pool E (42H17, 44N19, 44O19, 49E9, 49M13, 56F18), pool F (51H11, 51L3, 51N5, 52E9, 52H23, 53N20), pool G (57L10, 58N11, 59D23, 61N19, 64F9, 47B22), and pool H (65E11, 65I16, 66K2, 67O21, 67F12, 72O20). DNA from these pools, plus an aliquot of library "E," was mixed to give library "G," and used to prepare a 3-kb paired-end library for Roche 454 sequenced at sequencing approximately 8× coverage on one-quarter plate.

Sequenced libraries "E" and "G" were assembled together with the Roche Newbler assembler version 2.6. Assembly resulted in 251 scaffolds. Scaffolds comprising the full *Hox*

clusters of *Pantodon* for hoxax (200,509 bp), hoxbx (379,906 bp), hoxby (370,487 bp), hoxcx (247,835 bp), and hoxdx (254,542 bp) were identified by BLAST, and confirmed to contain the entire length of the original ten best BACs by end sequencing the clones using Sanger technology. Each cluster was spanned by two BACs from the original ten BACs sequenced in library "E."

## Whole-Genome Sequencing and the TRAM

Five micrograms of DNA from a single adult male specimen of *P. buchholzi* was used for Illumina HiSeq sequencing at the Wellcome Trust Center for Human Genetics, Oxford, using a 100-bp paired-end protocol and 180-bp insert size library. In total, 171 million reads, corresponding to approximately 22.9 × coverage of the approximately 753-Mb genome, were generated and used to build both local BLAST and HMMER search databases. To seed the searches against these databases, all vertebrate homeodomain protein sequences were downloaded from the homeodb web server (http://homeodb.zoo.ox.ac.uk/, last accessed July 4, 2014). Relaxed stringency tBLASTn search ($E = 10$) returned 14,297 sequences, which were extracted from the 171-million read database using a custom bash script, and were assembled using cap3 to produce 378 consensus sequences with a mean size of 227 bp and an N50 of 233 bp, corresponding to the approximately 180-bp homeobox-encoding exon. HMMER searches were performed using the PFAM homeobox.hmm profile and the hmmsearch algorithm to search the 6-frame translation of the reads. In total, 9,881 reads were extracted and assembled with cap3 (Huang 1999) into 338 consensus sequences. Cross-referencing the hmmsearch and tBLASTn results revealed 456 unique consensus sequences which represent putative homeobox sequences from the *Pantodon* genome. These were annotated using tBLASTn to the homeodb database and to a manually curated set of vertebrate *Hox* genes.

## Annotation of *Hox* Cluster Scaffolds

Initial ab initio predictions of all coding genes in the BAC assembly scaffolds were performed with genscan and verified by BLASTx searches against NCBI nr database. Only predicted proteins with high similarity ($E < 10^{-3}$) to known proteins were retained for further annotation. Predicted *Hox* gene sequences were mapped to the scaffolds and predicted coding sequences of these scaffolds were refined by BLASTn searches to the NCBI nr database and ClustalW multiple sequence alignment of predicted genes with the *Hox* gene-coding sequences of *D. rerio, Ta. rubripes, G. aculeatus, Te. nigroviridis, Homo sapiens, Xenopus tropicalis,* and *Gallus gallus* to manually refine predicted start codons, stop codons, and splice sites. The five *Hox* cluster scaffold sequences of *P. buchholzi* are deposited in GenBank under accession numbers KM102157–KM102161.

## Phylogenetic Analysis

Vertebrate species chosen for phylogenetic analysis were the clupeocephalan teleosts *D. rerio, Salmo salar, O. latipes, G. aculeatus, Ta. rubripes, Te. nigroviridis,* and *Astatotilapia burtoni*, the elopomorph teleost *A. anguilla* and sarcopterygian outgroups *Latimeria menadoensis, X. tropicalis, Anolis caroliensis, Mus musculus* and *H. sapiens* as well as the chondrichthyan *C. milii*. Full-length protein-coding Hox gene sequences from five complete teleost genomes *D. rerio* (zv9), *O. latipes* (MEDAKA1), *G. aculeatus* (BROADS1), *Ta. rubripes* (FUGU4), and *Te. nigroviridis* (TETRAODON8) as well as the sarcopterygian outgroups *X. tropicalis* (JGI_4.2), *A. caroliensis* (AnoCar2.0), *M. musculus* (GRCm38), and *H. sapiens* (GRCh37) were downloaded using the export function of the Ensembl genome browser (http://www.ensembl.org, last accessed July 4, 2014). *Salmo salar* sequences were obtained from NCBI accession numbers (EF695248–EF695353; EU02568–EU025719; EU221640–EU221655), *C. milli* from NCBI accession numbers (FJ824598–FJ824601), *L. menadoensis* from NCBI accession numbers (FJ497005–FJ497008), *A. burtoni* from NCBI accession numbers (EF594310–EF594316), and *A. anguilla* from NCBI accession numbers (JF891390–JF891400). Alignments for all individual Hox gene families were performed on amino acid sequences using ClustalW (Larkin et al. 2007) and trimmed with trimAl (Capella-Gutiérrez et al. 2009) using a gap threshold of 25%. Alignments used in the AU likelihood ratio tests were built using all available Hox gene family sequences, concatenated after separate alignment. ML trees were built using RAxML (Stamatakis 2006) with the LG amino acid substitution model with gamma and invariable parameters, and 1,000 bootstrap replicates. Bayesian trees were built using MrBayes (Huelsenbeck and Ronquist 2001) using the WAG amino acid substitution model and the invariable and gamma parameters across four chains for 500,000 generations and checked for convergence. Alternative topologies for likelihood ratio tests were constructed by building sets of three constrained trees containing three branches each—the clupeocephalan post-TGD "a" constrained branch, the clupeocephalan post-TGD "b" constrained branch, and an unduplicated outgroup branch—and forcing the individual Pantodon genes, or the whole concatenated cluster, to segregate alternatively within the "a" (topology A), "b" (topology B), or outgroup clades (topology O). Anguilla sequences were left unconstrained. The total tree and site-wise likelihood of each constrained topology was calculated using RAxML and the AU test (Shimodaira 2002) was performed with the program consel (Shimodaira and Hasegawa 2001). In each case, the unconstrained best topology was also built with RAxML and compared with the three constrained topologies. 3D-SLRPs were constructed using R.

## CNE Analysis

The mVISTA application of VISTA-tools (http://genome.lbl.gov/vista/index.shtml, last accessed July 4, 2014) using the SLAGAN algorithm was used to form VISTA plots of *Pantodon* and other vertebrate *hox* gene clusters. CNEs using the Elephant shark *C. millii* sequence as the baseline were predicted using the cutoff threshold 70% identity over 100 bp. CNEs were downloaded, and the total number,

length, and average size were counted using homemade PERL scripts.

### Selection Analysis

For selection analyses, individual *Hox* gene trees were constrained by assuming *Pantodon* shared the TGD and cluster diploidization with other teleosts and forced all *Pantodon* sequences in a single cluster within either the post-TGD "a" or "b" orthology groups as follows (hoxax + hoxaa, hoxbx + hoxba, hoxby + hoxbb, hoxcx + hoxca, hoxdx + hoxda) and built these constrained trees using RAxML with the same parameters as the unconstrained trees. We estimated selection on the most proximate post-TGD branches ("a" + "x," "b" + "y") using local branch models and branch-site models fit to the resultant ML trees using the software package HYPHY (Kosakovsky Pond et al. 2005). First, the MG94 codon substitution model (Muse and Gaut 1994) was fit locally to each individual branch to estimate d*N*, d*S*, and $\omega$ using the AnalyzeCodonData.bf function in HYPHY. Next, we used the BranchSiteREL.bf function of HYPHY to fit the MG94 model allowing three independent estimates of $\omega$ ($\omega^-$, $\omega^0$, $\omega^+$) along each branch to representing negative, neutral, and positive selection, respectively, quantified the proportion of sites ($p^+$) which appeared to evolve under positive selection.

## Supplementary Material

Supplementary figures S1–S4 and tables S1–S5 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

## Acknowledgments

The authors thank Jordi Paps, Ignacio Maeso, Ferdinand Marletaz, Jerome Hui, and Sebastian Shimeld for helpful discussions, John Mulley for initiating the BAC library construction, Christiane Hertz-Fowler and the University of Liverpool sequencing center for 454 sequencing, and the High-Throughput Genomics Group at the Wellcome Trust Centre for Human Genetics (funded by Wellcome Trust grant reference 090532/Z/09/Z and MRC Hub grant G0900747 91070) for the generation of Illumina Sequencing data. K.M. acknowledges support from the National Sciences and Engineering Research Council (NSERC) of Canada, the Oxford University Press Clarendon Fund, and the Elizabeth Hannah Jenkinson Fund of the University of Oxford, Department of Zoology. P.W.H.H. and K.M. acknowledge support from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013)/ERC grant (268513).

## References

Alfaro ME, Santini F, Brock C, Alamillo H, Dornburg A, Rabosky DL, Carnevale G, Harmon LJ. 2009. Nine exceptional radiations plus high turnover explain species diversity in jawed vertebrates. *Proc Natl Acad Sci U S A.* 106:13410–13414.

Allendorf FW, Danzmann RG. 1997. Secondary tetrasomic segregation of MDH-B and preferential pairing of homeologues in rainbow trout. *Genetics* 145:1083–1092.

Allendorf FW, Thorgaard GH. 1984. Tetraploidy and the evolution of salmonid fishes. In: Turner BJ, editor. Evolutionary genetics of fishes. New York: Plenum Press. p. 1–53.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic Local Alignment Search Tool. *J Mol Biol.* 215:403–410.

Amores A, Force A, Yan Y-L, Joly L, Amimiya C, Fritz A, Ho RK, Langeland J, Prince V, Wang Y-L, et al. 1998. Zebrafish hox clusters and vertebrate genome evolution. *Science* 282:1711–1714.

Amores A, Suzuki T, Yan Y-L, Pomeroy J, Singer A, Amemiya C, Postlethwait JH. 2004. Developmental roles of pufferfish Hox clusters and genome evolution in ray-fin fish. *Genome Res.* 14:1–10.

Arratia G. 1997. Basal teleosts and teleostean phylogeny. *Palaeo Ichthyol.* 7:5–168.

Becak M, Becak W, Rabello M. 1966. Cytological evidence of constant tetraploidy in the bisexual South American frog *Odontophrynus americanus*. *Chromosoma* 19:188–193.

Berthelot C, Brunet F, Chalopin D, Juanchich A, Bernard M, Noël B, Bento P, Da Silva C, Labadie K, Alberti A, et al. 2014. The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates. *Nat Commun.* 5:3657.

Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25:1972–1973.

Chambers KE, McDaniell 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.

Chester M, Gallagher JP, Symonds VV, Cruz da Silva AV, Mavrodiev EV, Leitch AR, Soltis PS, Soltis DE. 2012. Extensive chromosomal variation in a recently formed natural allopolyploid species, *Tragopogon miscellus* (Asteraceae). *Proc Natl Acad Sci U S A.* 109(4):1176–1181.

Christoffels A, Koh EGL, Chia J-M, Brenner S, Aparicio S, Venkatesh B. 2004. Fugu genome analysis provides evidence for a whole-genome duplication early during the evolution of ray-finned fishes. *Mol Biol Evol.* 21:1146–1151.

Coppe A, Pujolar JM, Maes GE, Larsen PF, Hansen MM, Bernatchez L, Zane L, Bortoluzzi S. 2010. Sequencing, de novo annotation and analysis of the first *Anguilla anguilla* transcriptome: EeelBase opens new perspectives for the study of the critically endangered European eel. *BMC Genomics* 11:635.

Crow KD, Smith CD, Cheng J-F, Wagner GP, Amemiya CT. 2012. An independent genome duplication inferred from Hox paralogs in the American paddlefish—a representative basal ray-finned fish and important comparative reference. *Genome Biol Evol.* 4:937–953.

Crow KD, Stadler PF, Lynch VJ, Amemiya C, Wagner GP. 2006. The "fish-specific" Hox cluster duplication is coincident with the origin of teleosts. *Mol Biol Evol.* 23:121–136.

Dayhoff M. 1976. Origin and evolution of protein superfamilies. *Fed Proc.* 35:2132–2138.

Demuth JP, Hahn MW. 2009. The life and death of gene families. *BioEssays* 31:29–39.

Dolinski K, Botstein D. 2007. Orthology and functional conservation in eukaryotes. *Annu Rev Genet.* 41:465–507.

Doyle JJ, Egan AN. 2010. Dating the origins of polyploidy events. *New Phytol.* 186:73–85.

Edger PP, Pires JC. 2009. Gene and genome duplications: the impact of dosage-sensitivity on the fate of nuclear genes. *Chromosome Res.* 17: 699–717.

Escriva H, Manzon L, Youson J, Laudet V. 2002. Analysis of lamprey and hagfish genes reveals a complex history of gene duplications during early vertebrate evolution. *Mol Biol Evol.* 19:1440–1450.

Fitch WM. 1970. Distinguishing homologous from analogous proteins. *Syst Zool.* 19:99–113.

Freeling M, Thomas BC. 2006. Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity. *Genome Res.* 16:805–814.

Fujimoto S, Oisi Y, Kuraku S, Ota KG, Kuratani S. 2013. Non-parsimonious evolution of hagfish Dlx genes. *BMC Evol Biol.* 13:15.

Gabaldón T, Koonin EV. 2013. Functional and evolutionary implications of gene orthology. *Nat Rev Genet.* 14:360–366.

Gaeta RT, Pires JC, Iniguez-Luy F, Leon E, Osborn TC. 2007. Genomic changes in resynthesized *Brassica napus* and their effect on gene expression and phenotype. *Plant Cell* 19:3403–3417.
Garcia-Fernàndez J, Holland PWH. 1994. Archetypal organization of the amphioxus Hox gene cluster. *Nature* 370:563–566.
Gaut BS, Doebley JF. 1997. DNA sequence evidence for the segmental allotetraploid origin of maize. *Proc Natl Acad Sci U S A.* 94: 6809–6814.
Gharbi K, Gautier A, Danzmann RG, Gharbi S, Sakamoto T, Hoyheim B, Taggart JB, Cairney M, Powell R, Krieg F, et al. 2006. A linkage map for brown trout (*Salmo trutta*): chromosome homeologies and comparative genome organization with other salmonid fish. *Genetics* 172:2405–2419.
Henkel CV, Burgerhout E, De Wijze DL, Dirks RP, Minegishi Y, Jansen HJ, Spaink HP, Dufour S, Weltzien F-A, Tsukamoto K, et al. 2012. Primitive duplicate Hox clusters in the European eel's genome. *PLoS One* 7:e32231.
Henkel CV, Dirks RP, De Wijze DL, Minegishi Y, Aoyama J, Jansen HJ, Turner B, Knudsen B, Bundgaard M, Hvam KL, et al. 2012. First draft genome sequence of the Japanese eel, *Anguilla japonica*. *Gene* 511: 195–201.
Hilton EJ. 2003. Comparative osteology and phylogenetic systematics of fossil and living bony-tongue fishes (Actinopterygii, Teleostei, Osteoglossomorpha). *Zool J Linn Soc.* 137:1–100.
Hinegardner R, Rosen DE. 1972. Cellular DNA content and the evolution of teleostean fishes. *Museum* 106:621–644.
Hoegg S, Boore JL, Kuehl JV, Meyer A. 2007. Comparative phylogenomic analyses of teleost fish Hox gene clusters: lessons from the cichlid fish *Astatotilapia burtoni*. *BMC Genomics* 8:317.
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.
Holland PW, Garcia-Fernàndez J, Williams NA, Sidow A. 1994. Gene duplications and the origins of vertebrate development. *Development* 1994:125–133.
Huang X. 1999. CAP3: A DNA sequence assembly program. *Genome Res.* 9:868–877.
Huelsenbeck JP, Ronquist F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* 17:754–755.
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 R Soc Lond B Biol Sci.* 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 Phylogenet Evol.* 20:275–285.
Inoue JG, Miya M, Tsukamoto K, Nishida M. 2003. Basal actinopterygian relationships: a mitogenomic perspective on the phylogeny of the "ancient fish." . *Mol Phylogenet Evol.* 26:110–120.
Irvine SQ, Carr JL, Bailey WJ, Kawasaki K, Shimizu N, Amemiya CT. 2002. Genomic analysis of Hox clusters in the sea lamprey *Petromyzon marinus*. *J Exp Zool.* 62:47–62.
Jaillon O, Aury J-M, Brunet F, Petit J-L, Strange-Thomann N, Mauceli E, Bouneau L, Fischer C, Ozouf-Costaz C, Bernot A, et al. 2004. Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. *Nature* 431:946–957.
Johnson KP, Walden KKO, Robertson HM. 2013. Next-generation phylogenomics using a Target Restricted Assembly Method. *Mol Phylogenet Evol.* 66:417–422.
Johnson KR, Wright JE, May B. 1987. Linkage relationships reflecting ancestral tetraploidy in salmonid fish. *Genetics* 116:579–591.
King BL, Gillis JA, Carlisle HR, Dahn RD. 2011. A natural deletion of the HoxC cluster in elasmobranch fishes. *Science* 334(6062):1517.
Koonin EV. 2005. Orthologs, paralogs, and evolutionary genomics. *Annu Rev Genet.* 39:309–338.
Kosakovsky Pond SL, Murrell B, Fourment M, Frost SD, Delport W, Scheffler K. 2011. A random effects branch-site model for detecting episodic diversifying selection. *Mol Biol Evol.* 24:1–13.
Kuraku S. 2008. Insights into cyclostome phylogenomics: pre-2R or post-2R. *Zool Sci.* 25:960–968.
Kuraku S. 2010. Palaeophylogenomics of the vertebrate ancestor—impact of hidden paralogy on hagfish and lamprey gene phylogeny. *Integr Comp Biol.* 50:124–129.
Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, et al. 2007. Clustal W and Clustal X version 2.0. *Bioinformatics* 23: 2947–2948.
Le HLV, Lecointre G, Perasso R. 1993. A 28S rRNA-based phylogeny of the gnathostomes—first steps in athe analysis of conflict and congruence with morphologically based cladeograms. *Mol Phylogenet Evol.* 2:31–51.
Lee AP, Kerk SY, Tan YY, Brenner S, Venkatesh B. 2010. Ancient vertebrate conserved noncoding elements have been evolving rapidly in teleost fishes. *Mol Biol Evol.* 28:1205–1215.
Lee GM, Wright JE. 1981. Mitotic and meiotic analyses of brook trout, *Salvelinus fontinalis*. *J Hered.* 72:321–327.
Li WH, Gu Z, Wang H, Nekrutenko A. 2001. Evolutionary analyses of the human genome. *Nature* 409:847–849.
Li Y-J, Yu Z, Zhang M-Z, Qian C, Abe S, Arai K. 2011. The origin of natural tetraploid loach *Misgurnus anguillicaudatus* (Teleostei: Cobitidae) inferred from meiotic chromosome configurations. *Genetica* 139:805–811.
Liang D, Wu R, Geng J, Wang C, Zhang P. 2011. A general scenario of Hox gene inventory variation among major sarcopterygian lineages. *BMC Evol Biol.* 11:25.
Lin Z, Ma H, Nei M. 2008. Ultraconserved coding regions outside the homeobox of mammalian Hox genes. *BMC Evol Biol.* 8:260.
Mable BK. 2004. "Why polyploidy is rarer in animals than in plants": myths and mechanisms. *Biol J Linn Soc Lond.* 82:453–466.
Makino T, McLysaght A. 2010. Ohnologs in the human genome are dosage balanced and frequently associated with disease. *Proc Natl Acad Sci U S A.* 107:9270–9274.
Málaga-Trillo E, Meyer A. 2001. Genome duplications and accelerated evolution of Hox genes and cluster architecture in teleost fishes. *Am Zool.* 41:676–686.
Mehta TK, Ravi V, Yamasaki S, Lee AP, Lian MM, Tay B-H, Tohari S, Yanai S, Tay A, Brenner S, et al. 2013. Evidence for at least six Hox clusters in the Japanese lamprey (*Lethenteron japonicum*). *Proc Natl Acad Sci U S A.* 110(40):16044–16049.
Metzker ML. 2009. Sequencing technologies—the next generation. *Nat Rev Genet.* 11:31–46.
Meyer A, Van de Peer Y. 2005. From 2R to 3R: evidence for a fish-specific genome duplication (FSGD). *BioEssays* 27:937–945.
Mungpakdee S, Seo H-C, Angotzi AR, Dong X, Akalin A, Chourrout D. 2008. Differential evolution of the 13 Atlantic salmon Hox clusters. *Mol Biol Evol.* 25:1333–1343.
Muse SV, Gaut BS. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Mol Biol Evol.* 11:715–724.
Near TJ, Eytan RI, Dornburg A, Kuhn KL, Moore JA, Davis MP, Wainwright PC, Friedman M, Smith WL. 2012. Resolution of ray-finned fish phylogeny and timing of diversification. *Proc Natl Acad Sci U S A.* 109:13698–13703.
Nehrt NL, Clark WT, Radivojac P, Hahn MW. 2011. Testing the ortholog conjecture with comparative functional genomic data from mammals. *PLoS Comput Biol.* 7:e1002073.
Nei M, Rooney AP. 2005. Concerted and birth-and-death evolution of multigene families. *Annu Rev Genet.* 39:121–152.
Ohno S. 1970. Evolution by gene duplication. New York: Springer-Verlag.
Otto SP. 2007. The evolutionary consequences of polyploidy. *Cell* 131: 452–462.
Otto SP, Whitton J. 2000. Polyploid incidence and evolution. *Annu Rev Genet.* 34:401–437.
Oulion S, Debiais-Thibaud M, d'Aubenton-Carafa Y, Thermes C, Da Silva C, Bernard-Samain S, Gavory F, Wincker P, Mazan S, Casane D. 2010. Evolution of Hox gene clusters in gnathostomes: insights from a

survey of a shark (*Scyliorhinus canicula*) transcriptome. *Mol Biol Evol.* 27:2829–2838.
Pascual-Anaya J, D'Aniello S, Kuratani S, Garcia-Fernàndez J. 2013. Evolution of Hox gene clusters in deuterostomes. *BMC Dev Biol.* 13:26.
Pennacchio LA, Ahituv N, Moses AM, Prabhakar S, Nobrega MA, Shoukry M, Minovitsky S, Dubchak I, Holt A, Lewis KD, et al. 2006. In vivo enhancer analysis of human conserved non-coding sequences. *Nature* 444:499–502.
Pick L, Heffer A. 2012. Hox gene evolution: multiple mechanisms contributing to evolutionary novelties. *Ann N Y Acad Sci.* 1256:15–32.
Pond SLK, Frost SDW, Muse SV. 2005. HyPhy: hypothesis testing using phylogenies. *Bioinformatics* 21:676–679.
Putnam NH, Butts T, Ferrier DEK, Furlong RF, Hellsten U, Kawashima T, Robinson-Rechavi M, Shoguchi E, Terry A, Yu J-K, et al. 2008. The amphioxus genome and the evolution of the chordate karyotype. *Nature* 453:1064–1071.
Qiu H, Hildebrand F, Kuraku S, Meyer A. 2011. Unresolved orthology and peculiar coding sequence properties of lamprey genes: the KCNA gene family as test case. *BMC Genomics* 12:325.
Ravi V, Lam K, Tay B-H, Tay A, Brenner S, Venkatesh B. 2009. Elephant shark (*Callorhinchus milii*) provides insights into the evolution of Hox gene clusters in gnathostomes. *Proc Natl Acad Sci U S A.* 106: 16327–16332.
Ravi V, Venkatesh B. 2008. Rapidly evolving fish genomes and teleost diversity. *Curr Opin Genet Dev.* 18:544–550.
Rokas A, Williams BL, King N, Carroll SB. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature* 425:798–804.
Salmon A, Flagel L, Ying B, Udall JA, Wendel JF. 2010. Homoeologous nonreciprocal recombination in polyploid cotton. *New Phytol.* 186: 123–134.
Shimeld SM, Donoghue PCJ. 2012. Evolutionary crossroads in developmental biology: cyclostomes (lamprey and hagfish). *Development* 139:2091–2099.
Shimodaira H. 2002. An approximately unbiased test of phylogenetic tree selection. *Syst Biol.* 51:492–508.
Shimodaira H, Hasegawa M. 2001. CONSEL: for assessing the confidence of phylogenetic tree selection. *Bioinformatics* 17:1246–1247.
Sidow A. 1996. Gen(om)e duplications in the evolution of early vertebrates. *Curr Opin Genet Dev.* 6:715–722.
Smith JJ, Kuraku S, Holt C, Sauka-Spengler T, Jiang N, Campbell M, Yandell MD, Manousaki T, Meyer A, Bloom OE, et al. 2013. Sequencing of the sea lamprey (*Petromyzon marinus*) genome provides insights into vertebrate evolution. *Nat Genet.* 45(4):415–421.
Söding J. 2005. Protein homology detection by HMM-HMM comparison. *Bioinformatics* 21:951–960.
Stadler PF, Fried C, Prohaska SJ, Bailey WJ, Misof BY, Ruddle FH, Wagner GP. 2004. Evidence for independent Hox gene duplications in the hagfish lineage: a PCR-based gene inventory of *Eptatretus stoutii*. *Mol Phylogenet Evol.* 32:686–694.
Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22:2688–2690.
Stebbins GL. 1947. Types of polyploids; their classification and significance. *Adv Genet.* 1:403–429.
Stebbins GL. 1971. Chromosomal evolution in higher plants. London: Edward Arnold Ltd.
Studer RA, Robinson-Rechavi M. 2009. How confident can we be that orthologs are similar, but paralogs differ? *Trends Genet.* 25(5):210–216.
Sullivan JP. 2004. Simultaneous analysis of five molecular markers provides a well-supported phylogenetic hypothesis for the living bonytongue fishes (Osteoglossomorpha: Teleostei). *Mol Phylogenet Evol.* 33:171–185.
Taylor JS, Van de Peer Y, Meyer A. 2001. Revisiting recent challenges to the ancient fish-specific genome duplication hypothesis. *Curr Biol.* 11:R1005–R1008.
Thornton JW, DeSalle R. 2000. Gene family evolution and homology: genomics meets phylogenetics. *Annu Rev Genomics Hum Genet.* 1: 41–73.
Tümpel S, Cambronero F, Wiedemann LM, Krumlauf R. 2006. Evolution of cis elements in the differential expression of two Hoxa2 coparalogous genes in pufferfish (*Takifugu rubripes*). *Proc Natl Acad Sci U S A.* 103:5419–5424.
Van de Peer Y, Maere S, Meyer A. 2009. The evolutionary significance of ancient genome duplications. *Nat Rev Genet.* 10:725–732.
Van de Peer Y, Maere S, Meyer A. 2010. 2R or not 2R is not the question anymore. *Nat Rev Genet.* 11:166.
Vandepoele K, De Vos W, Taylor JS, Meyer A, Van de Peer Y. 2004. Major events in the genome evolution of vertebrates: paranome age and size differ considerably between ray-finned fishes and land vertebrates. *Proc Natl Acad Sci U S A.* 101:1638–1643.
Venkatesh B, Kirkness EF, Loh Y-H, Halpern AL, Lee AP, Johnson J, Dandona N, Viswanathan LD, Tay A, Venter JC, et al. 2006. Ancient noncoding elements conserved in the human genome. *Science* 314:1892.
Wada H, Garcia-Fernàndez J, Holland PW. 1999. Colinear and segmental expression of amphioxus Hox genes. *Dev Biol.* 213:131–141.
Wendel JF. 2000. Genome evolution in polyploids. *Plant Mol Biol.* 42: 225–249.
White S. 1994. Out of deepest Africa. *Tropical Fish Hobbyist* 43:106–108.
Wolfe KH. 2001. Yesterday's polyploids and the mystery of diploidization. *Nat Rev Genet.* 2:333–341.
Woltering JM, Durston AJ. 2006. The zebrafish hoxDb cluster has been reduced to a single microRNA. *Nat Genet.* 38:601–602.
Woolfe A, Goodson M, Goode DK, Snell P, McEwan GK, Vavouri T, Smith SF, North P, Callaway H, Krys K, et al. 2005. Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol.* 3:e7.
Zhong Y-F, Butts T, Holland PWH. 2008. HomeoDB: a database of homeobox gene diversity. *Evol Dev.* 10:516–518.

## Appendix 2

## Analysis of Hox Genes in the spotted gar *Lepisosteus oculeatus*

Contribution to a publication in *Nature Genetics* (Braasch et al. 2016)

Title:
**Linking human biology to fish models: Sequence of the spotted gar genome**

Authors:
Ingo Braasch *, Andrew R. Gehrke , Jeramiah J. Smith , Kazuhiko Kawasaki , Tereza Manousaki , Jeremy Pasquier , Angel Amores , Thomas Desvignes , Peter Batzel , Julian Catchen , Aaron M. Berlin , Michael S. Campbell , Daniel Barrell,, Kyle J. Martin, John F. Mulley, Vydianathan Ravi, Alison P. Lee, Tetsuya Nakamura , Domitille Chalopin, Shaohua Fan, Dustin Wcisel, Cristian Cañestro, Jason Sydes , Felix E. G. Beaudry, Yi Sun,, Jana Hertel, Michael J. Beam , Federica Di Palma , Mario Fasold, Mikio Ishiyama, Jeremy Johnson , Steffi Kehr, Marcia Lara , John H. Letaw , Gary W. Litman, Ronda T. Litman, Masato Mikami, Tatsuya Ota, Nil Ratan Saha, Louise Williams , Peter F. Stadler, Han Wang,, John S. Taylor, Quenton Fontenot, Allyse Ferrara, Stephen M. J. Searle, Bronwen Aken,, Mark Yandell , Igor Schneider, Jeffrey A. Yoder, Jean-Nicolas Volff, Axel Meyer, Chris T. Amemiya, Byrappa Venkatesh, Peter W. H. Holland, Yann Guiguen , Julien Bobe , Neil H. Shubin , Jessica Alföldi , Kerstin Lindblad-Toh, & John H. Postlethwait *

Contributed text:

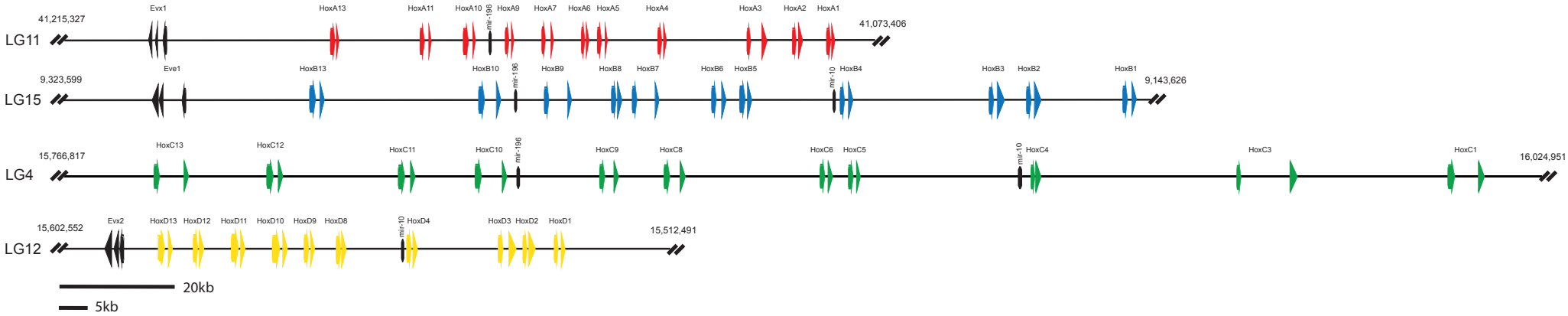
"With orthologs of 43 *hox* genes organized into four clusters (HoxA, HoxB, HoxC and HoxD), (Supplementary Figure 11) the protein coding complement of the gar *hox* clusters exhibits remarkable evolutionary stability (Supplementary Figure 12). Having suffered no gene losses (or possibly just *hoxD14*) since diverging from the last actinopterygian common ancestor of holosteans and teleosts, and only one gene (*hoxD14*) since the last rayfin (actinopterygian) common ancestor (LACA), the gar *hox* clusters serve as a useful outgroup for inferring the evolutionary consequences of *hox* cluster duplication associated with the TGD (Supplementary Figure 12).

The *hoxA6* and *hoxD2* genes appear to have been lost in the common ancestor of all teleosts following divergence from gar. The loss of *hoxA6* seems to be unique to teleosts, while

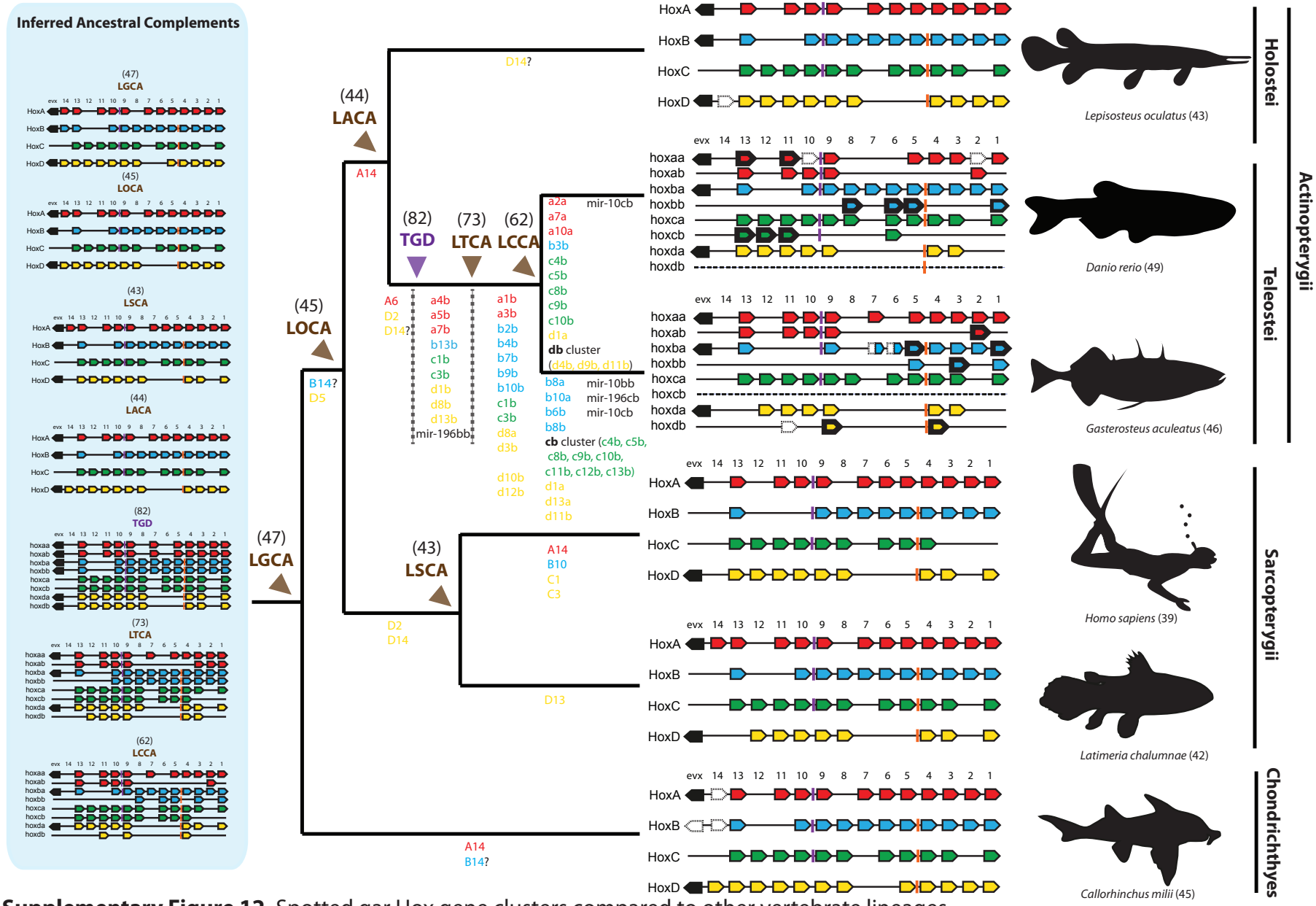
*hoxD2* has been lost at least twice – once in teleosts as well as in the last common ancestor of lobe-finned fish. Additionally, extensive lineage-specific gene losses of at least one *hox* gene duplicate characterizes the evolution of teleosts following the TGD with the result that most teleosts so far examined do not have significantly greater numbers of *hox* genes than gar (zebrafish 49, stickleback 46), and no teleosts have been found to maintain all 82 *hox* genes thought to be present in their last common ancestor immediately following the TGD. Reconstructed from genes observed in extant teleosts, the last teleost common ancestor (LTCA) had 74 *hox* genes (Supplementary Figure 12).

The protein-coding gene complement of the spotted gar *hox* clusters is also remarkably similar to the inferred complement of the last bony vertebrate (osteichthyan) common ancestor (LOCA). Only the *hoxA14* and *hoxD14* genes have been lost in the gar lineage since the ancient divergence between ray-finned (actinopterygian) and lobe-finned (sarcopterygian) fish ca. 450 million years ago. An intact *hoxA14* gene has to date been reported only in coelacanth and thus appears to have been lost at least three times since the last common ancestor of all jawed vertebrates – once in the lineage leading to cartilaginous fish (chondrichthyans), once in the lineage leading to tetrapods, and once in the lineage leading to ray-finned fishes. An intact *hoxD14* gene has been isolated in chondrichthyans including the elephant shark, horn shark, and lesser spotted catshark and has recently been reported in the paddlefish, a ray-finned fish diverging basally to spotted gar. No indication of *hoxD14* has been found in any teleost to date, but a discernible *hoxD14* pseudogene is found in the gar genome located in a conserved position between *hoxD13* and *evx2* (Supplementary Figure 13A), and with three regions of similarity corresponding to the three exons of other vertebrate h*oxD14* genes (Supplementary Figure 13B). The presence of a pseudogene in gar but not in teleosts suggests that this gene could have been inactivated independently in the two lineages, and more recently in the lineage leadings to gars. Alternatively, the presence of the *hoxD14* pseudogene presence might reflect the slower rate of evolution of gar hox clusters relative to teleosts. No *hoxD14* transcript was

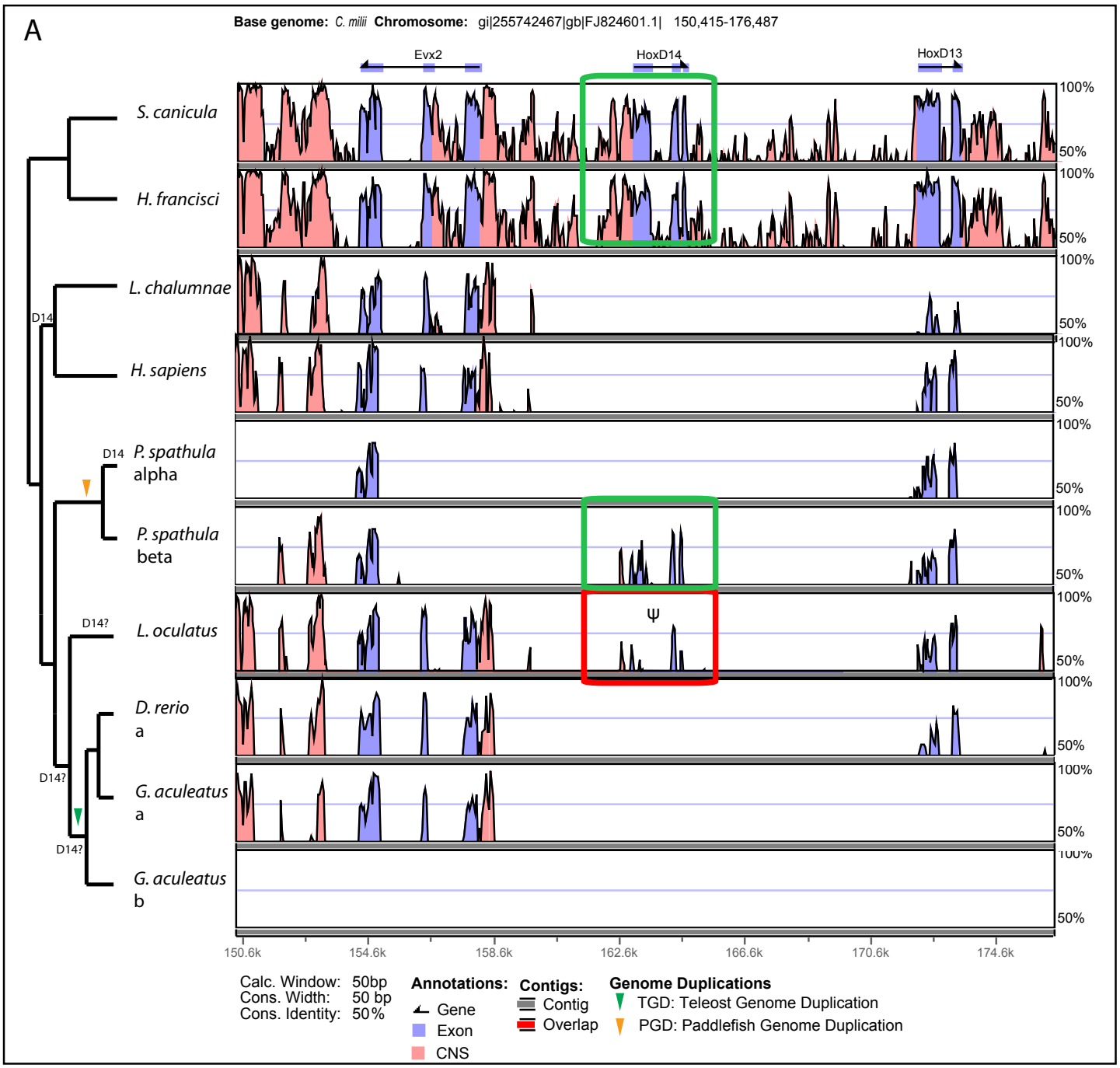
detected in either the gar or bowfin transcriptomes. Taken together, this suggests that the *hoxD14* gene was lost independently at least two times in evolution – once in the common ancestor of all lobe-finned vertebrates, and once in the last common ancestor of teleosts and holosteans (gars and bowfin) with slow sequence evolution in gars. Alternatively, *hoxD14* was lost three times, i.e. in lobe-finned vertebrates and independently in teleosts and in gars."




**Supplementary Figure 11.** Scale representation of gar **(***Lepisosteus oculeatus***)** Hox gene clusters**.**
HoxA is shown in red, HoxB in blue, HoxC in green, and HoxD in yellow. Gene names and miRNA names given above each element.




**Supplementary Figure 12.** Spotted gar Hox gene clusters compared to other vertebrate lineages.
Gene losses (colored gene names) are mapped onto the phylogeny. Inferred Hox gene complements of the last common ancestors of gnathostomes (LGCA), osteichthyans (LOCA),sarcopterygians (LSCA), actinopteryiigans (LACA), teleosts (LTCA), and clupeocephalan teleosts (LCCA), as well as gene content immediately following the teleost genome duplication (TGD) are shown to the left and total Hox gene counts are given in parentheses. Vertical bars on chromosomes: miR196 (purple) between Hox9 and Hox10 paralogy groups and miR10 (red) between Hox4 and Hox5 paralogy groups. In the teleosts *Danio rerio* and *Gasterosteus aculeatus* Hox protein-coding sequences which have experienced evolutionary rate acceleration



**Supplementary Figure 13.** The spotted gar hoxD14 pseudogene**.** A) VISTA plot of the Evx2-HoxD14- HoxD13 region in elephant shark (*C. milii*) against the (co)-orthologous regions from lesser spotted catshark (*S. canicula*), horn shark (*H. francisci*), coelacanth (*L. chalumnae*), human *(H. sapiens)*, paddlefish (*P. spathula*), gar (*L. oculatus*), zebrafish (*D. rerio*), and stickleback (*G. aculeatus*). Conserved sequences in exons are indicated in blue, conserved non-coding elements (CNEs) in red. A functional HoxD14 gene is found in cartilaginous fish and paddlefish (green boxes) 92, and a hoxD14 pseudogene (ψ) is found in gar (red box), including an adjacent conserved non-coding region. HoxD14 gene losses are mapped onto the phylogeny to the left. PGD, paddlefish genome duplication; TGD, teleost genome duplication. B) Alignment of elephant shark (Cmi), horn shark (Hfr) and paddlefish (Psp) HoxD14 nucleotide sequence with the gar pseudogene (Locψ) that translates into stop codons in all three exons.

## References


Abascal, F., M. L. Tress, and A. Valencia. 2015. Alternative splicing and co-option of transposable elements: The case of TMPO/LAP2α and ZNF451 in mammals. Bioinformatics 31:2257–2261.

Abouheif, E. 1997. Developmental genetics and homology: a hierarchical approach. Trends in Ecology & Evolution 12:7–10.

Alfaro, M. E., F. Santini, C. Brock, H. Alamillo, A. Dornburg, D. L. Rabosky, G. Carnevale, and L. J. Harmon. 2009. Nine exceptional radiations plus high turnover explain species diversity in jawed vertebrates. Proceedings of the National Academy of Sciences of the United States of America 106:13410–4.

Allendorf, F. W., and R. G. Danzmann. 1997. Secondary tetrasomic segregation of MDH-B and preferential pairing of homeologues in rainbow trout. Genetics 145:1083–92.

Allendorf, F. W., and G. H. Thorgaard. 1984. Tetraploidy and the evolution of salmonid fishes. Pages 1–53 *in* B. J. Turner, editor. Evolutionary Genetics of Fishes. Plenum Press.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic Local Alignment Search Tool. Journal of Molecular Biology 215:403–410.

Amemiya, C. T., J. Alföldi, A. P. Lee, S. Fan, H. Philippe, I. Maccallum, I. Braasch, T. Manousaki, I. Schneider, N. Rohner, C. Organ, D. Chalopin, J. J. Smith, M. Robinson, R. a. Dorrington, M. Gerdol, B. Aken, M. A. Biscotti, M. Barucca, D. Baurain, A. M. Berlin, G. L. Blatch, F. Buonocore, T. Burmester, M. S. Campbell, A. Canapa, J. P. Cannon, A. Christoffels, G. De Moro, A. L. Edkins, L. Fan, A. M. Fausto, N. Feiner, M. Forconi, J. Gamieldien, S. Gnerre, A. Gnirke, J. V. Goldstone, W. Haerty, M. E. Hahn, U. Hesse, S. Hoffmann, J. Johnson, S. I. Karchner, S. Kuraku, M. Lara, J. Z. Levin, G. W. Litman, E. Mauceli, T. Miyake, M. G. Mueller, D. R. Nelson, A. Nitsche, E. Olmo, T. Ota, A. Pallavicini, S. Panji, B. Picone, C. P. Ponting, S. J. Prohaska, D. Przybylski, N. R. Saha, V. Ravi, F. J. Ribeiro, T. Sauka-Spengler, G. Scapigliati, S. M. J. Searle, T. Sharpe, O. Simakov, P. F. Stadler, J. J. Stegeman, K. Sumiyama, D. Tabbaa, H. Tafer, J. Turner-Maier, P. van Heusden, S. White, L. Williams, M. Yandell, H. Brinkmann, J.-N. Volff, C. J. Tabin, N. Shubin, M. Schartl, D. B. Jaffe, J. H. Postlethwait, B. Venkatesh, F. Di Palma, E. S. Lander, A. Meyer, and K. Lindblad-Toh. 2013. The African coelacanth genome provides insights into tetrapod evolution. Nature 496:311–6.

Amemiya, C. T., T. P. Powers, S. J. Prohaska, J. Grimwood, J. Schmutz, M. Dickson, T. Miyake, M. a Schoenborn, R. M. Myers, F. H. Ruddle, and P. F. Stadler. 2010. Complete HOX cluster characterization of the coelacanth provides further evidence for slow evolution of its genome. Proceedings of the National Academy of Sciences of the United States of America 107:3622–7.

Amemiya, C. T., S. J. Prohaska, A. Hill-force, and P. F. Stadler. 2008. The Amphioxus

Hox Cluster : Characterization , Comparative Genomics , and Evolution. Journal of Experimental Zoology 477:465–477.

Amores, A., J. Catchen, A. Ferrara, Q. Fontenot, and J. H. Postlethwait. 2011. Genome Evolution and Meiotic Maps by Massively Parallel DNA Sequencing: Spotted Gar, an Outgroup for the Teleost Genome Duplication. Genetics 188:799–808.

Amores, A., A. Force, Y.-L. Yan, L. Joly, C. Amemiya, A. Fritz, R. K. Ho, J. Langeland, V. Prince, Y.-L. Wang, M. Westerfield, M. Ekker, and J. H. Postlethwait. 1998. Zebrafish hox Clusters and Vertebrate Genome Evolution. Science 282:1711–1714.

Amores, A., T. Suzuki, Y. Yan, J. Pomeroy, A. Singer, C. Amemiya, and J. H. Postlethwait. 2004. Developmental Roles of Pufferfish Hox Clusters and Genome Evolution in Ray-Fin Fish. Genome Research 14:1–10.

Aparicio, S., J. Chapman, E. Stupka, N. Putnam, J.-M. Chia, P. Dehal, A. Christoffels, S. Rash, S. Hoon, A. Smit, M. D. S. Gelpke, J. Roach, T. Oh, I. Y. Ho, M. Wong, C. Detter, F. Verhoef, P. Predki, A. Tay, S. Lucas, P. Richardson, S. F. Smith, M. S. Clark, Y. J. K. Edwards, N. Doggett, A. Zharkikh, S. V Tavtigian, D. Pruss, M. Barnstead, C. Evans, H. Baden, J. Powell, G. Glusman, L. Rowen, L. Hood, Y. H. Tan, G. Elgar, T. Hawkins, B. Venkatesh, D. Rokhsar, and S. Brenner. 2002. Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. Science 297:1301–10.

Arnegard, M. E., D. J. Zwickl, Y. Lu, and H. H. Zakon. 2010. Old gene duplication facilitates origin and diversification of an innovative communication system — twice. Proceedings of the National Academy of Sciences of the United States of America 107:22172–22177.

Arratia, G. 1997. Basal teleosts and teleostean phylogeny. Palaeo Ichthyologica 7:5–168.

Aury, J.-M., O. Jaillon, L. Duret, B. Noel, C. Jubin, B. M. Porcel, B. Ségurens, V. Daubin, V. Anthouard, N. Aiach, O. Arnaiz, A. Billaut, J. Beisson, I. Blanc, K. Bouhouche, F. Câmara, S. Duharcourt, R. Guigo, D. Gogendeau, M. Katinka, A.-M. Keller, R. Kissmehl, C. Klotz, F. Koll, A. Le Mouël, G. Lepère, S. Malinsky, M. Nowacki, J. K. Nowak, H. Plattner, J. Poulain, F. Ruiz, V. Serrano, M. Zagulski, P. Dessen, M. Bétermier, J. Weissenbach, C. Scarpelli, V. Schächter, L. Sperling, E. Meyer, J. Cohen, and P. Wincker. 2006. Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*. Nature 444:171–8.

Bao, W., K. K. Kojima, and O. Kohany. 2015. Repbase Update, a database of repetitive elements in eukaryotic genomes. Mobile DNA 6:1–6.

Becak, M., W. Becak, and M. Rabello. 1966. Cytological evidence of constant tetraploidy in the bisexual south american frog *Odontophrynus americanus*. Chromosoma 19:188–193.

Berthelot, C., F. Brunet, D. Chalopin, A. Juanchich, M. Bernard, B. Noël, P. Bento, C. Da Silva, K. Labadie, A. Alberti, J.-M. Aury, A. Louis, P. Dehais, P. Bardou, J. Montfort, C. Klopp, C. Cabau, C. Gaspin, G. H. Thorgaard, M. Boussaha, E. Quillet, R. Guyomard, D. Galiana, J. Bobe, J.-N. Volff, C. Genêt, P. Wincker, O. Jaillon, H. Roest Crollius, and Y. Guiguen. 2014. The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates. Nature Communications 5:1-10.

Böhne, A., Q. Zhou, A. Darras, C. Schmidt, M. Schartl, D. Galiana-Arnoux, and J.-N. Volff. 2012. Zisupton--a novel superfamily of DNA transposable elements recently active in fish. Molecular Biology and Evolution 29:631–45.

Braasch, I., F. Brunet, J. N. Volff, and M. Schartl. 2009. Pigmentation Pathway Evolution after Whole-Genome Duplication in Fish. Genome Biology and Evolution 1:479–493.

Braasch, I., A. R. Gehrke, J. J. Smith, K. Kawasaki, T. Manousaki, J. Pasquier, A. Amores, T. Desvignes, P. Batzel, J. Catchen, A. M. Berlin, M. S. Campbell, D. Barrell, K. J. Martin, J. F. Mulley, V. Ravi, A. P. Lee, T. Nakamura, D. Chalopin, S. Fan, D. Wcisel, C. Canestro, J. Sydes, F. E. G. Beaudry, Y. Sun, J. Hertel, M. J. Beam, M. Fasold, M. Ishiyama, J. Johnson, S. Kehr, M. Lara, J. H. Letaw, G. W. Litman, R. T. Litman, M. Mikami, T. Ota, N. R. Saha, L. Williams, P. F. Stadler, H. Wang, J. S. Taylor, Q. Fontenot, A. Ferrara, S. M. J. Searle, B. Aken, M. Yandell, I. Schneider, J. A. Yoder, J.-N. Volff, A. Meyer, C. T. Amemiya, B. Venkatesh, P. W. H. Holland, Y. Guiguen, J. Bobe, N. H. Shubin, F. Di Palma, J. Alfoldi, K. Lindblad-Toh, and J. H. Postlethwait. 2016. The spotted gar genome illuminates vertebrate evolution and facilitates human-teleost comparisons. Nature Genetics 48:427–437.

Braasch, I., and J. H. Postlethwait. 2012. Polyploidy in Fish and the Teleost Genome Duplication. Pages 341–383 *in* P. S. Soltis and D. E. Soltis, editors. Polyploidy and Genome Evolution. Springer Berlin Heidelberg, Berlin, Heidelberg.

Broughton, R. E. 2010. Origin and Phylogenetic Interrelationships of Teleosts. Pages 61–76 *in* J. S. Nelson, H. Schultze, and M. V. H. Wilson, editors. Origin and Phylogenetic Interrelationships of Teleosts. Verlag Dr. Friedrich Pfeil, München, Germany.

Broughton, R. E., R. Betancur-R, C. Li, G. Arratia, and G. Ortí. 2013. Multi-locus phylogenetic analysis reveals the pattern and tempo of bony fish evolution. PLOS Currents Tree of Life April 16:1–33.

Brunet, F. G., H. Roest Crollius, M. Paris, J.-M. Aury, P. Gibert, O. Jaillon, V. Laudet, and M. Robinson-Rechavi. 2006. Gene loss and evolutionary rates following whole-genome duplication in teleost fishes. Molecular Biology and Evolution 23:1808–16.

Bucciarelli, G., G. Bernardi, and G. Bernardi. 2002. An ultracentrifugation analysis of two hundred fish genomes. Gene 295:153–62.

Burge, C., and S. Karlin. 1997. Prediction of complete gene structures in human

genomic DNA. Journal of Molecular Biology 268:78–94.

Bürglin, T. R. 2011. Homeodomain Subtypes and Functional Diversity. Pages 95–122 *in* T. R. Hughes, editor. A Handbook of Transcription Factors. Springer Netherlands, Dordrecht.

Burns, K. H., and J. D. Boeke. 2008. Great exaptations. Journal of Biology 7:1–4.

Cantarel, B. L., I. Korf, S. M. C. Robb, G. Parra, E. Ross, B. Moore, C. Holt, A. Sánchez Alvarado, and M. Yandell. 2008. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Research 18:188–96.

Capella-Gutiérrez, S., J. M. Silla-Martínez, and T. Gabaldón. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25:1972–3.

Carroll, S. B. 2008. Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. Cell 134:25–36.

Chalopin, D., M. Naville, F. Plard, D. Galiana, and J.-N. Volff. 2015. Comparative Analysis of Transposable Elements Highlights Mobilome Diversity and Evolution in Vertebrates. Genome Biology and Evolution 7:567–580.

Chambers, K. E., R. McDaniell, J. D. Raincrow, M. Deshmukh, P. F. Stadler, and C. Chiu. 2009. Hox cluster duplication in the basal teleost *Hiodon alosoides* (Osteoglossomorpha). Theory in Biosciences 128:109–20.

Chester, M., J. P. Gallagher, V. V. Symonds, A. V. Cruz da Silva, E. V Mavrodiev, A. R. Leitch, P. S. Soltis, and D. E. Soltis. 2012. Extensive chromosomal variation in a recently formed natural allopolyploid species, *Tragopogon miscellus* (Asteraceae). Proceedings of the National Academy of Sciences of the United States of America 109:1176–1181.

Chiu, C.-H., K. Dewar, G. P. Wagner, K. Takahashi, F. Ruddle, C. Ledje, P. Bartsch, J.-L. Scemama, E. Stellwag, C. Fried, S. J. Prohaska, P. F. Stadler, and C. T. Amemiya. 2004. Bichir HoxA cluster sequence reveals surprising trends in ray-finned fish genomic evolution. Genome Research 14:11–7.

Christoffels, A., E. G. L. Koh, J.-M. Chia, S. Brenner, S. Aparicio, and B. Venkatesh. 2004. Fugu genome analysis provides evidence for a whole-genome duplication early during the evolution of ray-finned fishes. Molecular Biology and Evolution 21:1146–51.

Comai, L. 2005. The advantages and disadvantages of being polyploid. Nature Reviews Genetics 6:836–46.

Conant, G. C., and K. H. Wolfe. 2008. Turning a hobby into a job: how duplicated genes find new functions. Nature Reviews Genetics 9:938–50.

Coppe, A., J. M. Pujolar, G. E. Maes, P. F. Larsen, M. M. Hansen, L. Bernatchez, L.

Zane, and S. Bortoluzzi. 2010. Sequencing, de novo annotation and analysis of the first *Anguilla anguilla* transcriptome: EeelBase opens new perspectives for the study of the critically endangered European eel. BMC Genomics 11:1–13.

Crow, K. D., C. D. Smith, J.-F. Cheng, G. P. Wagner, and C. T. Amemiya. 2012. An independent genome duplication inferred from Hox paralogs in the American paddlefish--a representative basal ray-finned fish and important comparative reference. Genome Biology and Evolution 4:937–53.

Crow, K. D., P. F. Stadler, V. J. Lynch, C. Amemiya, and G. P. Wagner. 2006. The "fish-specific" Hox cluster duplication is coincident with the origin of teleosts. Molecular Biology and Evolution 23:121–36.

Danecek, P., A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, G. Lunter, G. T. Marth, S. T. Sherry, G. McVean, and R. Durbin. 2011. The variant call format and VCFtools. Bioinformatics 27:2156–8.

Dayhoff, M. 1976. Origin and Evolution of Protein Superfamilies. Federation Proceedings 35:2132–2138.

Dehal, P., and J. L. Boore. 2005. Two rounds of whole genome duplication in the ancestral vertebrate. PLoS Biology 3:e314.

Demuth, J. P., and M. W. Hahn. 2009. The life and death of gene families. BioEssays 31:29–39.

Denton, J. F., J. Lugo-Martinez, A. E. Tucker, D. R. Schrider, W. C. Warren, and M. W. Hahn. 2014. Extensive error in the number of genes inferred from draft genome assemblies. PLoS Computational Biology 10:e1003998.

Deschamps, J. 2007. Ancestral and recently recruited global control of the Hox genes in development. Current Opinion in Genetics & Development 17:422–7.

Di-Poï, N., J. I. Montoya-Burgos, and D. Duboule. 2009. Atypical relaxation of structural constraints in Hox gene clusters of the green anole lizard. Genome Research 19:602–10.

Di-Poï, N., J. I. Montoya-Burgos, H. Miller, O. Pourquié, M. C. Milinkovitch, and D. Duboule. 2010. Changes in Hox genes' structure and function during the evolution of the squamate body plan. Nature 464:99–103.

Dolinski, K., and D. Botstein. 2007. Orthology and functional conservation in eukaryotes. Annual Review of Genetics 41:465–507.

Donoghue, P. C. J., and M. A. Purnell. 2005. Genome duplication, extinction and vertebrate evolution. Trends in Ecology & Evolution 20:312–9.

Doyle, J. J., and A. N. Egan. 2010. Dating the origins of polyploidy events. The New Phytologist 186:73–85.

Eakin, G. S., and R. R. Behringer. 2003. Tetraploid development in the mouse. Developmental Dynamics 228:751–66.

Echave, J., S. J. Spielman, and C. O. Wilke. 2016. Causes of evolutionary rate variation among protein sites. Nature Reviews Genetics 17:109–121.

Edger, P. P., and J. C. Pires. 2009. Gene and genome duplications: the impact of dosage-sensitivity on the fate of nuclear genes. Chromosome Research 17:699–717.

Escriva, H., L. Manzon, J. Youson, and V. Laudet. 2002. Analysis of lamprey and hagfish genes reveals a complex history of gene duplications during early vertebrate evolution. Molecular Biology and Evolution 19:1440–50.

Feschotte, C. 2004. Merlin, a new superfamily of DNA transposons identified in diverse animal genomes and related to bacterial IS1016 insertion sequences. Molecular Biology and Evolution 21:1769–80.

Feschotte, C. 2008. The contribution of transposable elements ot the evolution of regulatory networks. Nature Reviews Genetics 9:397–405.

Fitch, W. M. 1970. Distinguishing homologous from analogous proteins. Systematic Zoology 19:99–113.

Fonseca, N. a, C. P. Vieira, P. W. H. Holland, and J. Vieira. 2008. Protein evolution of ANTP and PRD homeobox genes. BMC Evolutionary Biology 8:1–10.

Force, a, M. Lynch, F. B. Pickett, a Amores, Y. L. Yan, and J. Postlethwait. 1999. Preservation of duplicate genes by complementary, degenerative mutations. Genetics 151:1531–45.

Freeling, M., and B. C. Thomas. 2006. Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity. Genome Research 16:805–14.

Fried, C., S. J. Prohaska, and P. F. Stadler. 2004. Exclusion of repetitive DNA elements from gnathostome Hox clusters. Journal of Experimental Zoology. Part B, Molecular and Developmental Evolution 302:165–73.

Fu, L., B. Niu, Z. Zhu, S. Wu, and W. Li. 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 28:3150–2.

Fujimoto, S., Y. Oisi, S. Kuraku, K. G. Ota, and S. Kuratani. 2013. Non-parsimonious evolution of hagfish Dlx genes. BMC Evolutionary Biology 13:1–11.

Furlong, R. F., and P. W. H. Holland. 2002. Were vertebrates octoploid? Philosophical Transactions of the Royal Society of London B: Biological Sciences 357:531–44.

Gabaldón, T., and E. V Koonin. 2013. Functional and evolutionary implications of gene orthology. Nature Reviews Genetics 14:360–6.

Gaeta, R. T., J. C. Pires, F. Iniguez-Luy, E. Leon, and T. C. Osborn. 2007. Genomic changes in resynthesized Brassica napus and their effect on gene expression and phenotype. The Plant Cell 19:3403–17.

Gallardo, M. H., C. A. González, and I. Cebrián. 2006. Molecular cytogenetics and allotetraploidy in the red vizcacha rat, *Tympanoctomys barrerae* (Rodentia, Octodontidae). Genomics 88:214–221.

Garcia de la Serrana, D., E. a Mareco, and I. a Johnston. 2014. Systematic variation in the pattern of gene paralog retention between the teleost superorders Ostariophysi and Acanthopterygii. Genome Biology and Evolution 6:981–7.

Garcia-Fernàndez, J. 2005. The genesis and evolution of homeobox gene clusters. Nature Reviews Genetics 6:881–92.

Garcia-Fernàndez, J., and P. W. H. Holland. 1994. Archetypal organization of the amphioxus Hox gene cluster. Nature 370:563–566.

Gardner, C. A., and K. F. Barald. 1992. Expression Patterns of Engrailed-Like Proteins in the Chick-Embryo. Developmental Dynamics 193:370–388.

Gaut, B. S., and J. F. Doebley. 1997. DNA sequence evidence for the segmental allotetraploid origin of maize. Proceedings of the National Academy of Sciences of the United States of America 94:6809–14.

Ghanem, N., O. Jarinova, A. Amores, Q. Long, G. Hatch, B. K. Park, J. L. R. Rubenstein, and M. Ekker. 2003. Regulatory Roles of Conserved Intergenic Domains in Vertebrate Dlx Bigene Clusters. Genome Research:533–543.

Gharbi, K., A. Gautier, R. G. Danzmann, S. Gharbi, T. Sakamoto, B. Høyheim, J. B. Taggart, M. Cairney, R. Powell, F. Krieg, N. Okamoto, M. M. Ferguson, L.-E. Holm, and R. Guyomard. 2006. A linkage map for brown trout (*Salmo trutta*): chromosome homeologies and comparative genome organization with other salmonid fish. Genetics 172:2405–19.

Glasauer, S. M. K., and S. C. F. Neuhauss. 2014. Whole-genome duplication in teleost fishes and its evolutionary consequences. Molecular Genetics and Genomics 289:1045–60.

Gnerre, S., I. MacCallum, D. Przybylski, F. J. Ribeiro, J. N. Burton, B. J. Walker, T. Sharpe, G. Hall, T. P. Shea, S. Sykes, A. M. Berlin, D. Aird, M. Costello, R. Daza, L. Williams, R. Nicol, A. Gnirke, C. Nusbaum, E. S. Lander, and D. B. Jaffe. 2011. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proceedings of the National Academy of Sciences 108:1513–8.

Goodwin, S., J. D. McPherson, and W. R. McCombie. 2016. Coming of age: ten years of next-generation sequencing technologies. Nature Reviews Genetics 17:333–351.

Götz, S., J. M. García-Gómez, J. Terol, T. D. Williams, S. H. Nagaraj, M. J. Nueda, M.

Robles, M. Talón, J. Dopazo, and A. Conesa. 2008. High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic Acids Research 36:3420–35.

Grabherr, M. G., B. J. Haas, M. Yassour, J. Z. Levin, D. a Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, Z. Chen, E. Mauceli, N. Hacohen, A. Gnirke, N. Rhind, F. di Palma, B. W. Birren, C. Nusbaum, K. Lindblad-Toh, N. Friedman, and A. Regev. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature Biotechnology 29:644–52.

Graur, D., and W. Martin. 2004. Reading the entrails of chickens: molecular timescales of evolution and the illusion of precision. Trends in Genetics 20:80–86.

Gregory, T. R., J. a Nicol, H. Tamm, B. Kullman, K. Kullman, I. J. Leitch, B. G. Murray, D. F. Kapraun, J. Greilhuber, and M. D. Bennett. 2007. Eukaryotic genome size databases. Nucleic Acids Research 35:332–8.

Guo, B., X. Gan, and S. He. 2010. Hox genes of the Japanese eel *Anguilla japonica* and Hox cluster evolution in teleosts. Journal of Experimental Zoology Part B: Molecular and Developmental Evolution 314:135–47.

Hahn, M. W. 2009. Distinguishing among evolutionary models for the maintenance of gene duplicates. The Journal of Heredity 100:605–17.

Henkel, C. V, E. Burgerhout, D. L. de Wijze, R. P. Dirks, Y. Minegishi, H. J. Jansen, H. P. Spaink, S. Dufour, F.-A. Weltzien, K. Tsukamoto, and G. E. E. J. M. van den Thillart. 2012a. Primitive Duplicate Hox Clusters in the European Eel's Genome. PLoS One 7:e32231.

Henkel, C. V, R. P. Dirks, D. L. de Wijze, Y. Minegishi, J. Aoyama, H. J. Jansen, B. Turner, B. Knudsen, M. Bundgaard, K. L. Hvam, M. Boetzer, W. Pirovano, F.-A. Weltzien, S. Dufour, K. Tsukamoto, H. P. Spaink, and G. E. E. J. M. van den Thillart. 2012b. First draft genome sequence of the Japanese eel, *Anguilla japonica*. Gene 511:195–201.

Herpin, A., I. Braasch, M. Kraeussling, C. Schmidt, E. C. Thoma, S. Nakamura, M. Tanaka, and M. Schartl. 2010. Transcriptional rewiring of the sex determining dmrt1 gene duplicate by transposable elements. PLoS Genetics 6:e1000844.

Hillis, D. M. 1994. Homology in Molecular Biology. Pages 340–368 Homology: The Hierarchical Basis of Comparative Biology. Academic Press, Inc.

Hilton, E. J. 2003. Comparative osteology and phylogenetic systematics of fossil and living bony-tongue fishes (Actinopterygii, Teleostei, Osteoglossomorpha). Zoological Journal of the Linnean Society 137:1–100.

Hinegardner, R., and D. E. Rosen. 1972. Cellular DNA Content and the Evolution of Teleostean Fishes. The American Naturalist 106:621–644.

Ho, S. Y. W., and S. Duchêne. 2014. Molecular-clock methods for estimating

evolutionary rates and timescales. Molecular Ecology 23:5947–5965.

Hoegg, S., J. L. Boore, J. V Kuehl, and A. Meyer. 2007. Comparative phylogenomic analyses of teleost fish Hox gene clusters: lessons from the cichlid fish Astatotilapia burtoni. BMC Genomics 8:16.

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

Holland, P. W., J. Garcia-Fernàndez, N. A. Williams, and A. Sidow. 1994. Gene duplications and the origins of vertebrate development. Development 1994:125–33.

Holland, P. W. H. 2013. Evolution of homeobox genes. Wiley Interdisciplinary Reviews: Developmental Biology 2:31–45.

Holland, P. W. H. 2015. Did homeobox gene duplications contribute to the Cambrian explosion? Zoological Letters 1:1–8.

Holland, P. W. H., H. A. F. Booth, and E. A. Bruford. 2007. Classification and nomenclature of all human homeobox genes. BMC Biology 5:29.

Howe, K., M. D. Clark, C. F. Torroja, J. Torrance, C. Berthelot, M. Muffato, J. E. Collins, S. Humphray, K. McLaren, L. Matthews, S. McLaren, I. Sealy, M. Caccamo, C. Churcher, C. Scott, J. C. Barrett, R. Koch, G.-J. Rauch, S. White, W. Chow, B. Kilian, L. T. Quintais, J. a. Guerra-Assunção, Y. Zhou, Y. Gu, J. Yen, J.-H. Vogel, T. Eyre, S. Redmond, R. Banerjee, J. Chi, B. Fu, E. Langley, S. F. Maguire, G. K. Laird, D. Lloyd, E. Kenyon, S. Donaldson, H. Sehra, J. Almeida-King, J. Loveland, S. Trevanion, M. Jones, M. Quail, D. Willey, A. Hunt, J. Burton, S. Sims, K. McLay, B. Plumb, J. Davis, C. Clee, K. Oliver, R. Clark, C. Riddle, D. Eliott, G. Threadgold, G. Harden, D. Ware, B. Mortimer, G. Kerry, P. Heath, B. Phillimore, A. Tracey, N. Corby, M. Dunn, C. Johnson, J. Wood, S. Clark, S. Pelan, G. Griffiths, M. Smith, R. Glithero, P. Howden, N. Barker, C. Stevens, J. Harley, K. Holt, G. Panagiotidis, J. Lovell, H. Beasley, C. Henderson, D. Gordon, K. Auger, D. Wright, J. Collins, C. Raisen, L. Dyer, K. Leung, L. Robertson, K. Ambridge, D. Leongamornlert, S. McGuire, R. Gilderthorp, C. Griffiths, D. Manthravadi, S. Nichol, G. Barker, S. Whitehead, M. Kay, J. Brown, C. Murnane, E. Gray, M. Humphries, N. Sycamore, D. Barker, D. Saunders, J. Wallis, A. Babbage, S. Hammond, M. Mashreghi-Mohammadi, L. Barr, S. Martin, P. Wray, A. Ellington, N. Matthews, M. Ellwood, R. Woodmansey, G. Clark, J. Cooper, A. Tromans, D. Grafham, C. Skuce, R. Pandian, R. Andrews, E. Harrison, A. Kimberley, J. Garnett, N. Fosker, R. Hall, P. Garner, D. Kelly, C. Bird, S. Palmer, I. Gehring, A. Berger, C. M. Dooley, Z. Ersan-Ürün, C. Eser, H. Geiger, M. Geisler, L. Karotki, A. Kirn, J. Konantz, M. Konantz, M. Oberländer, S. Rudolph-Geiger, M. Teucke, K. Osoegawa, B. Zhu, A. Rapp, S. Widaa, C. Langford, F. Yang, N. P. Carter, J. Harrow, Z. Ning, J. Herrero, S. M. J. Searle, A. Enright, R. Geisler, R. H. a. Plasterk, C. Lee, M. Westerfield, P. J. de Jong, L. I. Zon, J. H. Postlethwait, C. Nüsslein-Volhard, T. J. P. Hubbard, H. R. Crollius, J. Rogers, and D. L. Stemple. 2013. The zebrafish reference genome

sequence and its relationship to the human genome. Nature 496:498–503.

Huang, X. 1999. CAP3: A DNA Sequence Assembly Program. Genome Research 9:868–877.

Hubbard, T. J. P., B. L. Aken, S. Ayling, B. Ballester, K. Beal, E. Bragin, S. Brent, Y. Chen, P. Clapham, L. Clarke, G. Coates, S. Fairley, S. Fitzgerald, J. Fernandez-Banet, L. Gordon, S. Graf, S. Haider, M. Hammond, R. Holland, K. Howe, A. Jenkinson, N. Johnson, A. Kahari, D. Keefe, S. Keenan, R. Kinsella, F. Kokocinski, E. Kulesha, D. Lawson, I. Longden, K. Megy, P. Meidl, B. Overduin, A. Parker, B. Pritchard, D. Rios, M. Schuster, G. Slater, D. Smedley, W. Spooner, G. Spudich, S. Trevanion, A. Vilella, J. Vogel, S. White, S. Wilder, A. Zadissa, E. Birney, F. Cunningham, V. Curwen, R. Durbin, X. M. Fernandez-Suarez, J. Herrero, A. Kasprzyk, G. Proctor, J. Smith, S. Searle, and P. Flicek. 2009. Ensembl 2009. Nucleic Acids Research 37:D690–7.

Huelsenbeck, J. P., and F. Ronquist. 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17:754–5.

Hufton, A. L., and G. Panopoulou. 2009. Polyploidy and genome restructuring: a variety of outcomes. Current Opinion in Genetics & Development 19:600–6.

Hurley, I. A., R. L. Mueller, K. A. Dunn, E. J. Schmidt, M. Friedman, R. K. Ho, V. E. Prince, Z. Yang, M. G. Thomas, and M. I. Coates. 2007. A new time-scale for ray-finned fish evolution. Proceedings of the Royal Society B: Biological Sciences 274:489–498.

Ikuta, T., N. Yoshida, N. Satoh, and H. Saiga. 2004. Ciona intestinalis Hox gene cluster: Its dispersed structure and residual colinear expression in development. Proceedings of the National Academy of Sciences of the United States of America 101:15118–23.

Innan, H., and F. Kondrashov. 2010. The evolution of gene duplications: classifying and distinguishing between models. Nature Reviews Genetics 11:97–108.

Inoue, J. G., M. Miya, K. Tsukamoto, and M. Nishida. 2001. A mitogenomic perspective on the basal teleostean phylogeny: resolving higher-level relationships with longer DNA sequences. Molecular Phylogenetics and Evolution 20:275–85.

Inoue, J. G., M. Miya, K. Tsukamoto, and M. Nishida. 2003. Basal actinopterygian relationships: a mitogenomic perspective on the phylogeny of the “ancient fish”. Molecular Phylogenetics and Evolution 26:110–20.

Inoue, J. G., M. Miya, K. Tsukamoto, and M. Nishida. 2004. Mitogenomic evidence for the monophyly of elopomorph fishes (Teleostei) and the evolutionary origin of the leptocephalus larva. Molecular Phylogenetics and Evolution 32:274–86.

Inoue, J., Y. Sato, R. Sinclair, K. Tsukamoto, and M. Nishida. 2015. Rapid genome reshaping by multiple-gene loss after whole-genome duplication in teleost fish

suggested by mathematical modeling. Proceedings of the National Academy of Sciences of the United States of America 112:14918–14923.

Irvine, S. Q., J. L. Carr, W. J. Bailey, K. Kawasaki, N. Shimizu, and C. T. Amemiya. 2002. Genomic Analysis of Hox Clusters in the Sea Lamprey *Petromyzon marinus*. Journal of Experimental Zoology Part B: Molecular and Developmental Evolution 294:47–62.

Jaillon, O., J.-M. Aury, F. Brunet, J.-L. Petit, N. Stange-Thomann, E. Mauceli, L. Bouneau, C. Fischer, C. Ozouf-Costaz, A. Bernot, S. Nicaud, D. Jaffe, S. Fisher, G. Lutfalla, C. Dossat, B. Segurens, C. Dasilva, M. Salanoubat, M. Levy, N. Boudet, S. Castellano, V. Anthouard, C. Jubin, V. Castelli, M. Katinka, B. Vacherie, C. Biémont, Z. Skalli, L. Cattolico, J. Poulain, V. De Berardinis, C. Cruaud, S. Duprat, P. Brottier, J.-P. Coutanceau, J. Gouzy, G. Parra, G. Lardier, C. Chapple, K. J. McKernan, P. McEwan, S. Bosak, M. Kellis, J.-N. Volff, R. Guigó, M. C. Zody, J. Mesirov, K. Lindblad-Toh, B. Birren, C. Nusbaum, D. Kahn, M. Robinson-Rechavi, V. Laudet, V. Schachter, F. Quétier, W. Saurin, C. Scarpelli, P. Wincker, E. S. Lander, J. Weissenbach, and H. Roest Crollius. 2004. Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. Nature 431:946–57.

Jaillon, O., J.-M. Aury, B. Noel, A. Policriti, C. Clepet, A. Casagrande, N. Choisne, S. Aubourg, N. Vitulo, C. Jubin, A. Vezzi, F. Legeai, P. Hugueney, C. Dasilva, D. Horner, E. Mica, D. Jublot, J. Poulain, C. Bruyère, A. Billault, B. Segurens, M. Gouyvenoux, E. Ugarte, F. Cattonaro, V. Anthouard, V. Vico, C. Del Fabbro, M. Alaux, G. Di Gaspero, V. Dumas, N. Felice, S. Paillard, I. Juman, M. Moroldo, S. Scalabrin, A. Canaguier, I. Le Clainche, G. Malacrida, E. Durand, G. Pesole, V. Laucou, P. Chatelet, D. Merdinoglu, M. Delledonne, M. Pezzotti, A. Lecharny, C. Scarpelli, F. Artiguenave, M. E. Pè, G. Valle, M. Morgante, M. Caboche, A.-F. Adam-Blondon, J. Weissenbach, F. Quétier, and P. Wincker. 2007. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature 449:463–7.

Johnson, K. P., K. K. O. Walden, and H. M. Robertson. 2010. A Target Restricted Assembly Method (TRAM) for Phylogenomics. Nature Precedings:1–3.

Johnson, K. P., K. K. O. Walden, and H. M. Robertson. 2013. Molecular Phylogenetics and Evolution Next-generation phylogenomics using a Target Restricted Assembly Method 66:417–422.

Johnson, K. R., J. E. Wright, and B. May. 1987. Linkage relationships reflecting ancestral tetraploidy in salmonid fish. Genetics 116:579–91.

Jones, F. C., M. G. Grabherr, Y. F. Chan, P. Russell, E. Mauceli, J. Johnson, R. Swofford, M. Pirun, M. C. Zody, S. White, E. Birney, S. Searle, J. Schmutz, J. Grimwood, M. C. Dickson, R. M. Myers, C. T. Miller, B. R. Summers, A. K. Knecht, S. D. Brady, H. Zhang, A. a Pollen, T. Howes, C. Amemiya, J. Baldwin, T. Bloom, D. B. Jaffe, R. Nicol, J. Wilkinson, E. S. Lander, F. Di Palma, K. Lindblad-Toh, and

D. M. Kingsley. 2012. The genomic basis of adaptive evolution in threespine sticklebacks. Nature 484:55–61.

Jones, P., D. Binns, H.-Y. Chang, M. Fraser, W. Li, C. McAnulla, H. McWilliam, J. Maslen, A. Mitchell, G. Nuka, S. Pesseat, A. F. Quinn, A. Sangrador-Vegas, M. Scheremetjew, S.-Y. Yong, R. Lopez, and S. Hunter. 2014. InterProScan 5: genome-scale protein function classification. Bioinformatics 30:1236–40.

Joyner, A. L., and G. R. Martin. 1987. En-1 and En-2, two mouse genes with sequence homology to the *Drosophila* engrailed gene: expression during embryogenesis. Genes & Development 1:29–38.

Kasahara, M., K. Naruse, S. Sasaki, Y. Nakatani, W. Qu, B. Ahsan, T. Yamada, Y. Nagayasu, K. Doi, Y. Kasai, T. Jindo, D. Kobayashi, A. Shimada, A. Toyoda, Y. Kuroki, A. Fujiyama, T. Sasaki, A. Shimizu, S. Asakawa, N. Shimizu, S.-I. Hashimoto, J. Yang, Y. Lee, K. Matsushima, S. Sugano, M. Sakaizumi, T. Narita, K. Ohishi, S. Haga, F. Ohta, H. Nomoto, K. Nogata, T. Morishita, T. Endo, T. Shin-I, H. Takeda, S. Morishita, and Y. Kohara. 2007. The medaka draft genome and insights into vertebrate genome evolution. Nature 447:714–9.

Kassahn, K. S., V. T. Dang, S. J. Wilkins, A. C. Perkins, and M. A. Ragan. 2009. Evolution of gene function and regulatory control after whole-genome duplication : Comparative analyses in vertebrates. Genome Research 19:1404–1418.

Katoh, K., K. Misawa, K. Kuma, and T. Miyata. 2002. MAFFT : a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Research 30:3059–3066.

Kim, D., G. Pertea, C. Trapnell, H. Pimentel, R. Kelley, and S. L. Salzberg. 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biology 14:1–13.

King, B. L., J. A. Gillis, H. R. Carlisle, and R. D. Dahn. 2011. A Natural Deletion of the HoxC Cluster in Elasmobranch Fishes. Science 334:1517.

Koh, E. G. L., K. Lam, A. Christoffels, M. V Erdmann, S. Brenner, and B. Venkatesh. 2003. Hox gene clusters in the Indonesian coelacanth, *Latimeria menadoensis*. Proceedings of the National Academy of Sciences of the United States of America 100:1084–8.

Koonin, E. V. 2005. Orthologs, paralogs, and evolutionary genomics. Annual Review of Genetics 39:309–38.

Korf, I. 2004. Gene finding in novel genomes. BMC Bioinformatics 5:1–9.

Kosakovsky Pond, S. L., B. Murrell, M. Fourment, S. D. Frost, W. Delport, and K. Scheffler. 2011. A random effects branch-site model for detecting episodic diversifying selection. Molecular Biology and Evolution 24:1–13.

Kramerov, D. a, and N. S. Vassetzky. 2011. Origin and evolution of SINEs in eukaryotic

genomes. Heredity 107:487–95.

Kuraku, S. 2008. Insights into cyclostome phylogenomics: pre-2R or post-2R. Zoological Science 25:960–8.

Kuraku, S. 2010. Palaeophylogenomics of the vertebrate ancestor--impact of hidden paralogy on hagfish and lamprey gene phylogeny. Integrative and Comparative Biology 50:124–9.

Kurosawa, G., N. Takamatsu, M. Takahashi, M. Sumitomo, E. Sanaka, K. Yamada, K. Nishii, M. Matsuda, S. Asakawa, H. Ishiguro, K. Miura, Y. Kurosawa, N. Shimizu, Y. Kohara, and H. Hori. 2006. Organization and structure of hox gene loci in medaka genome and comparison with those of pufferfish and zebrafish genomes. Gene 370:75–82.

Langmead, B., and S. L. Salzberg. 2012. Fast gapped-read alignment with Bowtie 2. Nature Methods 9:357–9.

Larhammar, D., and C. Risinger. 1994. Larhammar and Risinger. Molecular Phylogenetics and Evolution 3:59–68.

Larkin, M. A., G. Blackshields, N. P. Brown, R. Chenna, P. A. McGettigan, H. McWilliam, F. Valentin, I. M. Wallace, A. Wilm, R. Lopez, J. D. Thompson, T. J. Gibson, and D. G. Higgins. 2007. Clustal W and Clustal X version 2.0. Bioinformatics 23:2947–8.

Lavoué, S., M. Miya, M. E. Arnegard, P. B. McIntyre, V. Mamonekene, and M. Nishida. 2010. Remarkable morphological stasis in an extant vertebrate despite tens of millions of years of divergence. Proceedings of the Royal Society B: Biological Sciences 278:1003–8.

Le, H. L. Van, G. Lecointre, and R. Perasso. 1993. A 28S rRNA-based phylogeny of the gnathostomes - first steps in athe analysis of conflict and congruence with morphologically based cladeograms. Molecular Phylogenetics and Evolution 2:31–51.

Lee, A. P., S. Y. Kerk, Y. Y. Tan, S. Brenner, and B. Venkatesh. 2011. Ancient vertebrate conserved noncoding elements have been evolving rapidly in teleost fishes. Molecular Biology and Evolution 28:1205–15.

Lee, G. M., and J. E. Wright. 1981. Mitotic and meiotic analyses of brook trout, *Salvelinus fontinalis*. The Journal of Heredity 72:321–327.

Leitch, E. A. R., D. E. Soltis, P. S. Soltis, I. J. Leitch, J. C. Pires, S. C. L. E. Comber, C. Smith, and L. Le. 2004. Biological relevance of polyploidy : ecology to genomics Polyploidy in fishes : patterns and processes. Biological Journal of the Linnean Society 82:431–442.

Lerat, E. 2010. Identifying repeats and transposable elements in sequenced genomes:

how to find your way through the dense forest of programs. Heredity 104:520–33.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, G. P. Data, and T. Sam. 2009. The Sequence Alignment / Map format and SAMtools. Bioinformatics 25:2078–2079.

Li, R., W. Fan, G. Tian, H. Zhu, L. He, J. Cai, Q. Huang, Q. Cai, B. Li, Y. Bai, Z. Zhang, Y. Zhang, W. Wang, J. Li, F. Wei, H. Li, M. Jian, J. Li, Z. Zhang, R. Nielsen, D. Li, W. Gu, Z. Yang, Z. Xuan, O. a Ryder, F. C.-C. Leung, Y. Zhou, J. Cao, X. Sun, Y. Fu, X. Fang, X. Guo, B. Wang, R. Hou, F. Shen, B. Mu, P. Ni, R. Lin, W. Qian, G. Wang, C. Yu, W. Nie, J. Wang, Z. Wu, H. Liang, J. Min, Q. Wu, S. Cheng, J. Ruan, M. Wang, Z. Shi, M. Wen, B. Liu, X. Ren, H. Zheng, D. Dong, K. Cook, G. Shan, H. Zhang, C. Kosiol, X. Xie, Z. Lu, H. Zheng, Y. Li, C. C. Steiner, T. T.-Y. Lam, S. Lin, Q. Zhang, G. Li, J. Tian, T. Gong, H. Liu, D. Zhang, L. Fang, C. Ye, J. Zhang, W. Hu, A. Xu, Y. Ren, G. Zhang, M. W. Bruford, Q. Li, L. Ma, Y. Guo, N. An, Y. Hu, Y. Zheng, Y. Shi, Z. Li, Q. Liu, Y. Chen, J. Zhao, N. Qu, S. Zhao, F. Tian, X. Wang, H. Wang, L. Xu, X. Liu, T. Vinar, Y. Wang, T.-W. Lam, S.-M. Yiu, S. Liu, H. Zhang, D. Li, Y. Huang, X. Wang, G. Yang, Z. Jiang, J. Wang, N. Qin, L. Li, J. Li, L. Bolund, K. Kristiansen, G. K.-S. Wong, M. Olson, X. Zhang, S. Li, H. Yang, J. Wang, and J. Wang. 2010. The sequence and de novo assembly of the giant panda genome. Nature 463:311–7.

Li, W. H., Z. Gu, H. Wang, and A. Nekrutenko. 2001. Evolutionary analyses of the human genome. Nature 409:847–9.

Li, X., and M. S. Waterman. 2003. Estimating the Repeat Structure and Length of DNA Sequences Using ℓ-Tuples. Genome Research 13:1916–1922.

Li, Y.-J., Z. Yu, M.-Z. Zhang, C. Qian, S. Abe, and K. Arai. 2011. The origin of natural tetraploid loach *Misgurnus anguillicaudatus* (Teleostei: Cobitidae) inferred from meiotic chromosome configurations. Genetica 139:805–11.

Liang, D., R. Wu, J. Geng, C. Wang, and P. Zhang. 2011. A General Scenario of Hox Gene Inventory Variation among Major Sarcopterygian Lineages. BMC Evolutionary Biology 11:1–13.

Lien, S., B. F. Koop, S. R. Sandve, J. R. Miller, P. Matthew, J. S. Leong, D. R. Minkley, A. Zimin, F. Grammes, H. Grove, A. Gjuvsland, B. Walenz, R. A. Hermansen, K. Von Schalburg, E. B. Rondeau, A. Di Genova, J. K. A. Samy, and J. O. Vik. 2016. The Atlantic salmon genome provides insights into rediploidization. Nature 533:200–205.

Lieschke, G. J., and P. D. Currie. 2007. Animal models of human disease: zebrafish swim into view. Nature Reviews Genetics 8:353–67.

Liu, B., Y. Shi, J. Yuan, X. Hu, H. Zhang, N. Li, Z. Li, Y. Chen, D. Mu, W. Fan, and H. Kong. 2013. Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome projects. arXiv 1308.2012:1–47.

Lowe, C. B., and D. Haussler. 2012. 29 Mammalian Genomes Reveal Novel Exaptations of Mobile Elements for Likely Regulatory Functions in the Human Genome. PLoS One 7:e43128.

Luo, J., P. F. Stadler, S. He, and A. Meyer. 2007. PCR Survey of Hox Genes in the Goldfish *Carassius auratus auratus*. Journal of Experimental Zoology Part B: Molecular and Developmental Evolution 308B:250–258.

Lynch, M. 2000. The Evolutionary Fate and Consequences of Duplicate Genes. Science 290:1151–1155.

Lynch, M., and J. S. Conery. 2003a. The evolutionary demography of duplicate genes. Journal of Structural and Functional Genomics 3:35–44.

Lynch, M., and J. S. Conery. 2003b. The origins of genome complexity. Science 302:1401–4.

Lynch, M., and A. G. Force. 2000. The Origin of Interspecific Genomic Incompatibility via Gene Duplication. The American Naturalist 156:590–605.

Mable, B. K., M. a. Alexandrou, and M. I. Taylor. 2011. Genome duplication in amphibians and fish: an extended synthesis. Journal of Zoology 284:151–182.

MacDonald, R. B., M. Debiais-Thibaud, K. Martin, L. Poitras, B.-H. Tay, B. Venkatesh, and M. Ekker. 2010. Functional conservation of a forebrain enhancer from the elephant shark (*Callorhinchus milii*) in zebrafish and mice. BMC Evolutionary Biology 10:1–11.

Macqueen, D. J., and I. A. Johnston. 2014. A well-constrained estimate for the timing of the salmonid whole genome duplication reveals major decoupling from species diversification. Proceedings of the Royal Society B: Biological Sciences 281:1–8.

Maere, S., S. De Bodt, J. Raes, T. Casneuf, M. Van Montagu, M. Kuiper, and Y. Van de Peer. 2005. Modeling gene and genome duplications in eukaryotes. Proceedings of the National Academy of Sciences of the United States of America 102:5454–9.

Makino, T., and A. McLysaght. 2010. Ohnologs in the human genome are dosage balanced and frequently associated with disease. Proceedings of the National Academy of Sciences of the United States of America 107:9270–4.

Málaga-Trillo, E., and A. Meyer. 2001. Genome duplications and accelerated evolution of Hox genes and cluster architecture in teleost fishes. American Zoologist 41:676–686.

Mallo, M., and C. R. Alonso. 2013. The regulation of Hox gene expression during animal development. Development 140:3951–63.

Mallo, M., D. M. Wellik, and J. Deschamps. 2010. Hox genes and regional patterning of the vertebrate body plan. Developmental Biology 344:7–15.

Marçais, G., and C. Kingsford. 2011. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics 27:764–70.

Martin, K. J., and P. W. H. Holland. 2014. Enigmatic orthology relationships between Hox clusters of the African butterfly fish and other teleosts following ancient whole-genome duplication. Molecular Biology and Evolution 31:2592–611.

McClintock, J. M., R. Carlson, D. M. Mann, and V. E. Prince. 2001. Consequences of Hox gene duplication in the vertebrates: an investigation of the zebrafish Hox paralogue group 1 genes. Development 128:2471–84.

McClintock, J. M., M. a Kheirbek, and V. E. Prince. 2002. Knockdown of duplicated zebrafish hoxb1 genes reveals distinct roles in hindbrain patterning and a novel mechanism of duplicate gene retention. Development 129:2339–54.

McGaugh, S. E., J. B. Gross, B. Aken, M. Blin, R. Borowsky, D. Chalopin, H. Hinaux, W. R. Jeffery, A. Keene, L. Ma, P. Minx, D. Murphy, K. E. O'Quin, S. Rétaux, N. Rohner, S. M. J. Searle, B. a Stahl, C. Tabin, J.-N. Volff, M. Yoshizawa, and W. C. Warren. 2014. The cavefish genome reveals candidate genes for eye loss. Nature Communications 5:1–10.

Mehta, T. K., V. Ravi, S. Yamasaki, A. P. Lee, M. M. Lian, B.-H. Tay, S. Tohari, S. Yanai, A. Tay, S. Brenner, and B. Venkatesh. 2013. Evidence for at least six Hox clusters in the Japanese lamprey (*Lethenteron japonicum*). Proceedings of the National Academy of Sciences of the United States of America 110:16044–16049.

Metzker, M. L. 2009. Sequencing technologies — the next generation. Nature Reviews Genetics 11:31–46.

Meyer, a, and M. Schartl. 1999. Gene and genome duplications in vertebrates: the one-to-four (-to-eight in fish) rule and the evolution of novel gene functions. Current Opinion in Cell Biology 11:699–704.

Meyer, A., and Y. Van de Peer. 2005. From 2R to 3R: evidence for a fish-specific genome duplication (FSGD). BioEssays 27:937–45.

Moss, S. P., D. a Joyce, S. Humphries, K. J. Tindall, and D. H. Lunt. 2011. Comparative analysis of teleost genome sequences reveals an ancient intron size expansion in the zebrafish lineage. Genome Biology and Evolution 3:1187–96.

Mulley, J. F., C. Chiu, and P. W. H. Holland. 2006. Breakup of a homeobox cluster after genome duplication in teleosts. Proceedings of the National Academy of Sciences of the United States of America 103:10369–72.

Mulley, J. F., and P. W. H. Holland. 2010. Parallel retention of Pdx2 genes in cartilaginous fish and coelacanths. Molecular Biology and Evolution 27:2386–91.

Mungpakdee, S., H.-C. Seo, A. R. Angotzi, X. Dong, A. Akalin, and D. Chourrout. 2008. Differential evolution of the 13 Atlantic salmon Hox clusters. Molecular Biology and Evolution 25:1333–43.

Muse, S. V, and B. S. Gaut. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. Molecular Biology and Evolution 11:715–24.

Nadeau, J. H., and D. Sankoff. 1997. Comparable rates of gene loss and functional divergence after genome duplications early in vertebrate evolution. Genetics 147:1259–66.

Near, T. J., R. I. Eytan, A. Dornburg, K. L. Kuhn, J. A. Moore, M. P. Davis, P. C. Wainwright, M. Friedman, and W. L. Smith. 2012. Resolution of ray-finned fish phylogeny and timing of diversification. Proceedings of the National Academy of Sciences of the United States of America 109:13698–13703.

Nehrt, N. L., W. T. Clark, P. Radivojac, and M. W. Hahn. 2011. Testing the ortholog conjecture with comparative functional genomic data from mammals. PLoS Computational Biology 7:e1002073.

Nei, M., and A. P. Rooney. 2005. Concerted and birth-and-death evolution of multigene families. Annual Review of Genetics 39:121–52.

Nelson, J. S. 2006. Fishes of the World. Fourth Edi. John Wiley & Sons, Inc., Hoboken, New Jersey.

Nishihara, H., A. F. A. Smit, and N. Okada. 2006. Functional noncoding sequences derived from SINEs in the mammalian genome. Genome Research 16:864–874.

Ohno, S. 1970. Evolution by gene duplication. London: George Alien & Unwin Ltd. Berlin, Heidelberg and New York: Springer-Verlag.

Ohno, S., U. Wolf, and N. B. Atkin. 1968. Evolution from fish to mammals by gene duplication. Hereditas 59:169–87.

Oliveira, C., L. F. Almeida-Toledo, L. Mori, and S. A. Toledo-Filho. 1993. Cytogenetic and DNA content in six genera of the family Callichthyidae (Pisces, Siluriformes). Caryologia 46:171–188.

Otto, S. P. 2007. The evolutionary consequences of polyploidy. Cell 131:452–62.

Otto, S. P., and J. Whitton. 2000. Polyploid incidence and evolution. Annual Review of Genetics 34:401–437.

Oulion, S., M. Debiais-Thibaud, Y. D'Aubenton-Carafa, C. Thermes, C. Da Silva, S. Bernard-Samain, F. Gavory, P. Wincker, S. Mazan, and D. Casane. 2010. Evolution of Hox gene clusters in gnathostomes: insights from a survey of a shark (*Scyliorhinus canicula*) transcriptome. Molecular Biology and Evolution 27:2829–38.

Parra, G., K. Bradnam, and I. Korf. 2007. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics 23:1061–7.

Parra, G., K. Bradnam, Z. Ning, T. Keane, and I. Korf. 2009. Assessing the gene space

in draft genomes. Nucleic Acids Research 37:289–97.

Pascual-Anaya, J., S. D Aniello, S. Kuratani, and J. Garcia-Fernàndez. 2013. Evolution of Hox gene clusters in deuterostomes. BMC Developmental Biology 13:1–14.

Van de Peer, Y., S. Maere, and A. Meyer. 2009. The evolutionary significance of ancient genome duplications. Nature Reviews Genetics 10:725–32.

Van de Peer, Y., S. Maere, and A. Meyer. 2010. 2R or not 2R is not the question anymore. Nature Reviews Genetics 11:166.

Pennacchio, L. A., N. Ahituv, A. M. Moses, S. Prabhakar, M. A. Nobrega, M. Shoukry, S. Minovitsky, I. Dubchak, A. Holt, K. D. Lewis, I. Plajzer-Frick, J. Akiyama, S. De Val, V. Afzal, B. L. Black, O. Couronne, M. B. Eisen, A. Visel, and E. M. Rubin. 2006. In vivo enhancer analysis of human conserved non-coding sequences. Nature 444:499–502.

Phillips, R. B., K. A. Keatley, M. R. Morasch, A. B. Ventura, K. P. Lubieniecki, B. F. Koop, R. G. Danzmann, and W. S. Davidson. 2009. Assignment of Atlantic salmon (*Salmo salar*) linkage groups to specific chromosomes : Conservation of large syntenic blocks corresponding to whole chromosome arms in rainbow trout (*Oncorhynchus mykiss*). BMC Genetics 11:1–11.

Pick, L., and A. Heffer. 2012. Hox gene evolution: multiple mechanisms contributing to evolutionary novelties. Annals of the New York Academy of Sciences 1256:15–32.

Piferrer, F., A. Beaumont, J. Falguière, and M. Flajšhans. 2009. Polyploid fish and shellfish : Production , biology and applications to aquaculture for performance improvement and genetic containment. Aquaculture 293:125–156.

Pond, S. L. K., S. D. W. Frost, and S. V Muse. 2005. HyPhy: hypothesis testing using phylogenies. Bioinformatics 21:676–9.

Postlethwait, J., A. Amores, W. Cresko, A. Singer, and Y.-L. Yan. 2004. Subfunction partitioning, the teleost radiation and the annotation of the human genome. Trends in Genetics 20:481–90.

Postlethwait, J. H. 2000. Zebrafish Comparative Genomics and the Origins of Vertebrate Chromosomes. Genome Research 10:1890–1902.

Postlethwait, J. H. 2007. The Zebrafish Genome in Context : Ohnologs Gone Missing. Journal of Experimental Zoology Part B: Molecular and Developmental Evolution 308B:563–577.

Presgraves, D. C. 2010. The molecular evolutionary basis of species formation. Nature Reviews Genetics 11:175–80.

Prince, V. E., and F. B. Pickett. 2002. Splitting pairs: the diverging fates of duplicated genes. Nature Reviews Genetics 3:827–37.

Pruitt, K. D., G. R. Brown, S. M. Hiatt, F. Thibaud-Nissen, A. Astashyn, O. Ermolaeva, C. M. Farrell, J. Hart, M. J. Landrum, K. M. McGarvey, M. R. Murphy, N. a O'Leary, S. Pujar, B. Rajput, S. H. Rangwala, L. D. Riddick, A. Shkeda, H. Sun, P. Tamez, R. E. Tully, C. Wallin, D. Webb, J. Weber, W. Wu, M. DiCuccio, P. Kitts, D. R. Maglott, T. D. Murphy, and J. M. Ostell. 2014. RefSeq: an update on mammalian reference sequences. Nucleic Acids Research 42:D756–63.

Putnam, N. H., T. Butts, D. E. K. Ferrier, R. F. Furlong, U. Hellsten, T. Kawashima, M. Robinson-Rechavi, E. Shoguchi, A. Terry, J.-K. Yu, E. L. Benito-Gutiérrez, I. Dubchak, J. Garcia-Fernàndez, J. J. Gibson-Brown, I. V Grigoriev, A. C. Horton, P. J. de Jong, J. Jurka, V. V Kapitonov, Y. Kohara, Y. Kuroki, E. Lindquist, S. Lucas, K. Osoegawa, L. a Pennacchio, A. a Salamov, Y. Satou, T. Sauka-Spengler, J. Schmutz, T. Shin-I, A. Toyoda, M. Bronner-Fraser, A. Fujiyama, L. Z. Holland, P. W. H. Holland, N. Satoh, and D. S. Rokhsar. 2008. The amphioxus genome and the evolution of the chordate karyotype. Nature 453:1064–71.

Qiu, H., F. Hildebrand, S. Kuraku, and A. Meyer. 2011. Unresolved orthology and peculiar coding sequence properties of lamprey genes: the KCNA gene family as test case. BMC Genomics 12:1–11.

Raff, R. A. 2007. Written in stone: fossils, genes and evo-devo. Nature Reviews Genetics 8:911–20.

Raincrow, J. D., K. Dewar, C. Stocsits, S. J. Prohaska, C. T. Amemiya, P. F. Stadler, and C.-H. Chiu. 2011. Hox clusters of the bichir (Actinopterygii, *Polypterus senegalus*) highlight unique patterns of sequence evolution in gnathostome phylogeny. Journal of Experimental Zoology Part B: Molecular and Developmental Evolution 316:451–64.

Ravi, V., K. Lam, B.-H. Tay, A. Tay, S. Brenner, and B. Venkatesh. 2009. Elephant shark (*Callorhinchus milii*) provides insights into the evolution of Hox gene clusters in gnathostomes. Proceedings of the National Academy of Sciences of the United States of America 106:16327–32.

Ravi, V., and B. Venkatesh. 2008. Rapidly evolving fish genomes and teleost diversity. Current Opinion in Genetics & Development 18:544–50.

Rokas, A., and P. W. H. Holland. 2000. Rare genomic changes as a tool for phylogenetics. Trends in Ecology & Evolution 15:454–459.

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

Rondeau, E. B., D. R. Minkley, J. S. Leong, A. M. Messmer, J. R. Jantzen, K. R. von Schalburg, C. Lemon, N. H. Bird, and B. F. Koop. 2014. The genome and linkage map of the northern pike (*Esox lucius*): conserved synteny revealed between the salmonid sister group and the Neoteleostei. PLoS One 9:e102089.

Salmon, A., L. Flagel, B. Ying, J. A. Udall, and J. F. Wendel. 2010. Homoeologous

nonreciprocal recombination in polyploid cotton. The New Phytologist 186:123–34.

Sanger, F., S. Nicklen, and a R. Coulson. 1977. DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences of the United States of America 74:5463–7.

Sankoff, D., C. Zheng, and Q. Zhu. 2010. The collapse of gene complement following whole genome duplication. BMC Genomics 11:1–11.

Santini, F., L. J. Harmon, G. Carnevale, and M. E. Alfaro. 2009. Did genome duplication drive the origin of teleosts? A comparative study of diversification in ray-finned fishes. BMC Evolutionary Biology 9:1–15.

Sasaki, T., H. Nishihara, M. Hirakawa, K. Fujimura, M. Tanaka, N. Kokubo, C. Kimura-Yoshida, I. Matsuo, K. Sumiyama, N. Saitou, T. Shimogori, and N. Okada. 2008. Possible involvement of SINEs in mammalian-specific brain formation. Proceedings of the National Academy of Sciences of the United States of America 105:4220–5.

Sato, Y., Y. Hashiguchi, and M. Nishida. 2009. Temporal pattern of loss/persistence of duplicate genes involved in signal transduction and metabolic pathways after teleost-specific genome duplication. BMC Evolutionary Biology 9:1–14.

Schartl, M., R. B. Walter, Y. Shen, T. Garcia, J. Catchen, A. Amores, I. Braasch, D. Chalopin, J.-N. Volff, K.-P. Lesch, A. Bisazza, P. Minx, L. Hillier, R. K. Wilson, S. Fuerstenberg, J. Boore, S. Searle, J. H. Postlethwait, and W. C. Warren. 2013. The genome of the platyfish, *Xiphophorus maculatus*, provides insights into evolutionary adaptation and several complex traits. Nature Genetics 45:567–72.

Schmieder, R., and R. Edwards. 2011. Quality control and preprocessing of metagenomic datasets. Bioinformatics 27:863–4.

Schranz, E. M., S. Mohammadin, and P. P. Edger. 2012. Ancient whole genome duplications, novelty and diversification: The WGD Radiation Lag-Time Model. Current Opinion in Plant Biology 15:147–153.

Schulz, M. H., D. R. Zerbino, M. Vingron, and E. Birney. 2012. Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics 28:1086–92.

Sémon, M., and K. H. Wolfe. 2007. Reciprocal gene loss between *Tetraodon* and zebrafish after whole genome duplication in their ancestor. Trends in Genetics 23:108–12.

Shen, S., J. L. Crowley, Y. Wang, S. A. Bowring, D. H. Erwin, P. M. Sadler, C. Cao, D. H. Rothman, C. M. Henderson, J. Ramezani, H. Zhang, Y. Shen, X. Wang, W. Wang, L. Mu, W. Li, Y. Tang, X. Liu, L. Liu, Y. Zeng, Y. Jiang, and Y. Jin. 2011. Calibrating the End-Permian Mass Extinction. Science 334:1367–1372.

Shendure, J., and H. Ji. 2008. Next-generation DNA sequencing. Nature Biotechnology

26:1135–45.

Shimeld, S. M., and P. C. J. Donoghue. 2012. Evolutionary crossroads in developmental biology: cyclostomes (lamprey and hagfish). Development 139:2091–9.

Shimodaira, H. 2002. An approximately unbiased test of phylogenetic tree selection. Systematic Biology 51:492–508.

Shimodaira, H., and M. Hasegawa. 2001. CONSEL: for assessing the confidence of phylogenetic tree selection. Bioinformatics 17:1246–7.

Sidow, A. 1996. Gen(om)e duplications in the evolution of early vertebrates. Current Opinion in Genetics & Development 6:715–722.

Simillion, C., K. Vandepoele, M. C. E. Van Montagu, M. Zabeau, and Y. Van de Peer. 2002. The hidden duplication past of *Arabidopsis thaliana*. Proceedings of the National Academy of Sciences of the United States of America 99:13627–32.

Smit, A. F. A., R. Hubley, and P. Green. 1996. Repeatmasker: RepeatMasker Open-3.0. www.repeatmasker.org.

Smith, J. J., S. Kuraku, C. Holt, T. Sauka-Spengler, N. Jiang, M. S. Campbell, M. D. Yandell, T. Manousaki, A. Meyer, O. E. Bloom, J. R. Morgan, J. D. Buxbaum, R. Sachidanandam, C. Sims, A. S. Garruss, M. Cook, R. Krumlauf, L. M. Wiedemann, S. a Sower, W. A. Decatur, J. A. Hall, C. T. Amemiya, N. R. Saha, K. M. Buckley, J. P. Rast, S. Das, M. Hirano, N. McCurley, P. Guo, N. Rohner, C. J. Tabin, P. Piccinelli, G. Elgar, M. Ruffier, B. L. Aken, S. M. J. Searle, M. Muffato, M. Pignatelli, J. Herrero, M. Jones, C. T. Brown, Y.-W. Chung-Davidson, K. G. Nanlohy, S. V Libants, C. Yeh, D. W. McCauley, J. A. Langeland, Z. Pancer, B. Fritzsch, P. J. de Jong, B. Zhu, L. L. Fulton, B. Theising, P. Flicek, M. E. Bronner, W. C. Warren, S. W. Clifton, R. K. Wilson, W. Li, P. J. De Jong, B. Zhu, L. L. Fulton, B. Theising, P. Flicek, M. E. Bronner, W. C. Warren, S. W. Clifton, R. K. Wilson, and W. Li. 2013a. Sequencing of the sea lamprey (*Petromyzon marinus*) genome provides insights into vertebrate evolution. Nature Genetics 45:415–21.

Söding, J. 2005. Protein homology detection by HMM-HMM comparison. Bioinformatics 21:951–60.

Sonnhammer, E. L. L., and E. V. Koonin. 2002. Orthology, paralogy and proposed classification for paralog subtypes. Trends in Genetics 18:619–620.

de Souza, F. S. J., L. F. Franchini, M. Rubinstein, F. S. J. De Souza, L. F. Franchini, and M. Rubinstein. 2013. Exaptation of transposable elements into novel cis-regulatory elements: is the evidence always strong? Molecular Biology and Evolution 30:1239–51.

Stadler, P. F., C. Fried, S. J. Prohaska, W. J. Bailey, B. Y. Misof, F. H. Ruddle, and G. P. Wagner. 2004. Evidence for independent Hox gene duplications in the hagfish

lineage: a PCR-based gene inventory of *Eptatretus stoutii*. Molecular Phylogenetics and Evolution 32:686–94.

Stamatakis, A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22:2688–90.

Stanke, M., R. Steinkamp, S. Waack, and B. Morgenstern. 2004. AUGUSTUS: a web server for gene finding in eukaryotes. Nucleic Acids Research 32:W309–12.

Star, B., A. J. Nederbragt, S. Jentoft, U. Grimholt, M. Malmstrøm, T. F. Gregers, T. B. Rounge, J. Paulsen, M. H. Solbakken, A. Sharma, O. F. Wetten, A. Lanzén, R. Winer, J. Knight, J.-H. Vogel, B. Aken, Ø. Andersen, K. Lagesen, A. Tooming-Klunderud, R. B. Edvardsen, K. G. Tina, M. Espelund, C. Nepal, C. Previti, B. O. Karlsen, T. Moum, M. Skage, P. R. Berg, T. Gjøen, H. Kuhl, J. Thorsen, K. Malde, R. Reinhardt, L. Du, S. D. Johansen, S. Searle, S. Lien, F. Nilsen, I. Jonassen, S. W. Omholt, N. C. Stenseth, and K. S. Jakobsen. 2011. The genome sequence of Atlantic cod reveals a unique immune system. Nature 477:207–210.

Stebbins, G. L. 1947. Types of polyploids: Their classification and significance. Advances in Genetics 1:403–429.

Stebbins, G. L. 1971. Chromosomal evolution in higher plants. Edward Arnold Ltd., London.

Studer, R. A., and M. Robinson-Rechavi. 2009. How confident can we be that orthologs are similar, but paralogs differ? Trends in Genetics 25:210–6.

Sullivan, J. P. 2004. Simultaneous analysis of five molecular markers provides a well-supported phylogenetic hypothesis for the living bony-tongue fishes (Osteoglossomorpha : Teleostei). Molecular Phylogenetics and Evolution 33:171–185.

Sumiyama, K., S. Q. Irvine, D. W. Stock, K. M. Weiss, K. Kawasaki, N. Shimizu, C. S. Shashikant, W. Miller, and F. H. Ruddle. 2002. Genomic structure and functional control of the Dlx3-7 bigene cluster. Proceedings of the National Academy of Sciences of the United States of America 99:3–8.

Takatori, N., T. Butts, S. Candiani, M. Pestarino, D. E. K. Ferrier, H. Saiga, and P. W. H. Holland. 2008. Comprehensive survey and classification of homeobox genes in the genome of amphioxus, *Branchiostoma floridae*. Development Genes and Evolution 218:579–90.

Tatusov, R. L., M. Y. Galperin, D. A. Natale, and E. V Koonin. 2000. The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Research 28:33–36.

Taylor, J. S., I. Braasch, T. Frickey, A. Meyer, and Y. Van de Peer. 2003. Genome duplication, a trait shared by 22000 species of ray-finned fish. Genome Research 13:382–90.

Taylor, J. S., Y. Van De Peer, and A. Meyer. 2001a. Genome duplication , divergent resolution and speciation. Trends in Genetics 17:299–301.

Taylor, J. S., Y. Van de Peer, A. Meyer, M. Robinson-Rechavi, O. Marchand, H. Escriva, and V. Laudet. 2001b. Revisiting recent challenges to the ancient fish-specific genome duplication hypothesis. Current Biology 11:R1005–8.

Tekaia, F. 2016. Inferring orthologs : Open questions and perspectives. Genomics Insights 9:17–28.

Thomas-Chollier, M., and V. Ledent. 2008. Comparative phylogenomic analyses of teleost fish Hox gene clusters: lessons from the cichlid fish *Astatotilapia burtoni*: comment. BMC Genomics 9:35.

Thompson, A., D. Vo, C. Comfort, and H. H. Zakon. 2014. Expression evolution facilitated the convergent neofunctionalization of a sodium channel gene. Molecular Biology and Evolution 31:1941–55.

Thornton, J. W., and R. DeSalle. 2000. Gene family evolution and homology: genomics meets phylogenetics. Annual Review of Genomics and Human Genetics 1:41–73.

Tschopp, P., B. Tarchini, F. Spitz, J. Zakany, and D. Duboule. 2009. Uncoupling time and space in the collinear regulation of Hox genes. PLoS Genetics 5:e1000398.

Tümpel, S., F. Cambronero, L. M. Wiedemann, and R. Krumlauf. 2006. Evolution of cis elements in the differential expression of two Hoxa2 coparalogous genes in pufferfish (*Takifugu rubripes*). Proceedings of the National Academy of Sciences of the United States of America 103:5419–24.

Uyeno, T., and G. R. Smith. 1972. Tetraploid Origin of the Karyotype of Catostomid Fishes. Science 175:644–646.

Vandepoele, K., W. De Vos, J. S. Taylor, A. Meyer, and Y. Van de Peer. 2004. Major events in the genome evolution of vertebrates: paranome age and size differ considerably between ray-finned fishes and land vertebrates. Proceedings of the National Academy of Sciences of the United States of America 101:1638–43.

Venkatesh, B., E. F. Kirkness, Y. Loh, A. L. Halpern, A. P. Lee, J. Johnson, N. Dandona, L. D. Viswanathan, A. Tay, J. C. Venter, R. L. Strausberg, and S. Brenner. 2006. Ancient noncoding elements conserved in the human genome. Science 314:1892.

Volff, J.-N. 2005. Genome evolution and biodiversity in teleost fish. Heredity 94:280–94.

Voss, S. R., S. Putta, J. A. Walker, J. J. Smith, N. Maki, and P. A. Tsonis. 2013. Salamander Hox clusters contain repetitive DNA and expanded non-coding regions: a typical Hox structure for non-mammalian tetrapod vertebrates? Human Genomics 7:1–7.

Wada, H., J. Garcia-Fernàndez, and P. W. Holland. 1999. Colinear and segmental

expression of amphioxus Hox genes. Developmental Biology 213:131–41.

Wagner, G. P. 2007. The developmental genetics of homology. Nature Reviews Genetics 8:473–9.

Wang, Y., Y. Lu, Y. Zhang, Z. Ning, Y. Li, Q. Zhao, H. Lu, R. Huang, X. Xia, Q. Feng, X. Liang, K. Liu, L. Zhang, T. Lu, T. Huang, D. Fan, Q. Weng, C. Zhu, Y. Lu, W. Li, Z. Wen, C. Zhou, Q. Tian, X. Kang, M. Shi, W. Zhang, S. Jang, F. Du, S. He, L. Liao, Y. Li, B. Gui, H. He, Z. Ning, C. Yang, L. He, L. Luo, R. Yang, Q. Luo, X. Liu, S. Li, W. Huang, L. Xiao, H. Lin, B. Han, and Z. Zhu. 2015. The draft genome of the grass carp (*Ctenopharyngodon idellus*) provides insights into its evolution and vegetarian adaptation. Nature Genetics 47:625–31.

Wendel, J. F. 2000. Genome evolution in polyploids. Plant Molecular Biology 42:225–49.

White, S. 1994. Out of deepest Africa. Tropical Fish Hobbyist 43:106–108.

Wolfe, K. 2000. Robustness - it's not where you think it is. Nature Genetics 25:3–4.

Wolfe, K. H. 2001. Yesterday's polyploids and the mystery of diploidization. Nature Reviews Genetics 2:333–41.

Wolfe, K. H., and D. C. Shields. 1997. Molecular evidence for an ancient duplication of the entire yeast genome. Nature 387:708–13.

Woltering, J. M., and A. J. Durston. 2006. The zebrafish hoxDb cluster has been reduced to a single microRNA. Nature Genetics 38:601–2.

Woolfe, A., M. Goodson, D. K. Goode, P. Snell, G. K. McEwen, T. Vavouri, S. F. Smith, P. North, H. Callaway, K. Kelly, K. Walter, I. Abnizova, W. Gilks, Y. J. K. Edwards, J. E. Cooke, and G. Elgar. 2005. Highly conserved non-coding sequences are associated with vertebrate development. PLoS Biology 3:e7.

Wu, C., D. Zhang, M. Kan, Z. Lv, A. Zhu, Y. Su, D. Zhou, J. Zhang, Z. Zhang, M. Xu, L. Jiang, B. Guo, T. Wang, C. Chi, Y. Mao, J. Zhou, X. Yu, H. Wang, X. Weng, J. G. Jin, J. Ye, L. He, and Y. Liu. 2014. The draft genome of the large yellow croaker reveals well-developed innate immunity. Nature Communications 5:1–7.

Xu, P., X. Zhang, X. Wang, J. Li, G. Liu, Y. Kuang, J. Xu, X. Zheng, L. Ren, G. Wang, Y. Zhang, L. Huo, Z. Zhao, D. Cao, C. Lu, C. Li, Y. Zhou, Z. Liu, Z. Fan, G. Shan, X. Li, S. Wu, L. Song, G. Hou, Y. Jiang, Z. Jeney, D. Yu, L. Wang, C. Shao, L. Song, J. Sun, P. Ji, J. Wang, Q. Li, L. Xu, F. Sun, J. Feng, C. Wang, S. Wang, B. Wang, Y. Li, Y. Zhu, W. Xue, L. Zhao, J. Wang, Y. Gu, W. Lv, K. Wu, J. Xiao, J. Wu, Z. Zhang, J. Yu, and X. Sun. 2014. Genome sequence and genetic diversity of the common carp, *Cyprinus carpio*. Nature Genetics 46:1212–9.

Yang, C., L. Cao, W. Wang, Y. Yang, K. Abbas, B. Yan, H. Wang, L. Su, Y. Sun, and H. Wang. 2007. Comparative and evolutionary analysis in natural diploid and tetraploid weather loach *Misgurnus anguillicaudatus* based on cytochrome b

sequence data in central China. Environmental Biology of Fishes 86:145–153.

Yang, Z., and B. Rannala. 2012. Molecular phylogenetics: principles and practice. Nature Reviews Genetics 13:303–314.

Zakon, H. H. 2012. Adaptive evolution of voltage-gated sodium channels: The first 800 million years. Proceedings of the National Academy of Sciences of the United States of America 109:10619–25.

Zerbino, D. R. 2010. Using the Velvet de novo assembler for short-read sequencing technologies. Current Protocols in Bioinformatics Chapter 11:11.55.1–11.5.12.

Zerbino, D. R., and E. Birney. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Research 18:821–9.

Zhang, J. 2003. Evolution by gene duplication: an update. Trends in Ecology & Evolution 18:292–298.

Zhong, Y.-F., T. Butts, and P. W. H. Holland. 2008. HomeoDB: a database of homeobox gene diversity. Evolution & Development 10:516–8.

Zhou, Q., G. Zhang, Y. Zhang, S. Xu, R. Zhao, Z. Zhan, X. Li, Y. Ding, S. Yang, and W. Wang. 2008. On the origin of new genes in *Drosophila*. Genome Research 18:1446–1455.