Evolution of globin X genes

Original Research Article

Whole genome duplications and the diversification of the Globin-X genes of vertebrates

Federico G. Hoffmann\textsuperscript{1,2,*}, Jay F. Storz\textsuperscript{3}, Shigehiro Kuraku\textsuperscript{4}, Mike W. Vandewege\textsuperscript{5}, Juan C. Opazo\textsuperscript{6,7,8}

\textsuperscript{1}Department of Biochemistry, Molecular Biology, Entomology, and Plant Pathology, Mississippi State University, Starkville, Mississippi, USA.
\textsuperscript{2}Institute for Genomics, Biocomputing and Biotechnology, Mississippi State University, Starkville, Mississippi, USA.
\textsuperscript{3}School of Biological Sciences, University of Nebraska, Lincoln, NE, USA.
\textsuperscript{4}Laboratory for Phyloinformatics, RIKEN Center for Biosystems Dynamics Research, Kobe, Japan.
\textsuperscript{5}Department of Biology, Eastern New Mexico University, Portales, New Mexico, USA.
\textsuperscript{6}Integrative Biology Group, Universidad Austral de Chile, Valdivia, Chile.
\textsuperscript{7}Instituto de Ciencias Ambientales y Evolutivas, Facultad de Ciencias, Universidad Austral de Chile, Valdivia, Chile.
\textsuperscript{8}Millennium Nucleus of Ion Channel-Associated Diseases (MiNICAD), Valdivia, Chile.

*Correspondence:
Federico G. Hoffmann
federico.g.hoffmann@gmail.com, fgh19@msstate.edu

Keywords: GbX, Gene Family Evolution, Gene duplication,
Abstract
The globin superfamily of vertebrate genes is a textbook example of how the interplay between local gene duplications, whole-genome duplications, and regulatory changes can facilitate the evolution of novel protein functions. Almost every vertebrate examined possesses copies of hemoglobin and myoglobin in their genomes, and both cytoglobin and neuroglobin are present in the vast majority of vertebrate genomes surveyed as well. The phylogenetic distribution of globin-E, globin-Y and globin-X (GbX), however, is more spotty, suggesting multiple independent gene losses. Globin-X is an enigmatic globin with a wide phyletic distribution that spans protostomes and deuterostomes. Unlike canonical globins such as hemoglobins and myoglobins, functional data suggest that GbX does not have a primary respiratory function. Instead, available evidence suggests that GbX may play a role in protecting cells from oxidative damage and in reducing nitrite and it is predicted to be bound to the cell membrane. Recently released genomes from key vertebrate taxa provide an excellent opportunity to address questions about the early stages of evolution of these genes in vertebrates. In the current study, we integrate bioinformatic, syntenic and phylogenetic analyses to characterize the diversity of GbX genes in non-teleost ray-finned fishes, resolve relationships between the GbX genes of cartilaginous fish and the GbX genes of bony vertebrates, and demonstrate that the GbX genes of cyclostomes and gnathostomes have independent duplicative histories. Our study highlights the role that whole genome duplications (WGDs) have played in expanding the repertoire of genes in vertebrate genomes. Our results indicate that GbX paralogs have a remarkably high rate of retention following WGDs in comparison to other globin genes, and also provide an evolutionary framework for interpreting results of experiments that examine functional properties of GbX and patterns of tissue-specific expression. By identifying GbX genes products of different WGDs in the vertebrate tree of life, our results can guide the design of experimental work to explore whether gene duplicates that originate via WGDs have evolved novel functional properties or expression profiles relative to singleton or tandemly duplicated copies of GbX.
1 Introduction

Globins are small, oxygen-binding hemoproteins found in all domains of life (Vinogradov et al., 2005). The globin superfamily of vertebrates provides an excellent example of how the interplay between local gene duplications, whole-genome duplications, amino acid and regulatory changes can fuel the evolution of novel protein functions (Storz et al., 2011, 2013; Keppner et al., 2020). The different types of globins in vertebrate genomes can be classified into four groups 1) androglobin, 2) neuroglobin, 3) globin-X (GbX), and 4) vertebrate-specific globins. The fourth category includes hemoglobin and myoglobin genes of gnathostomes, in addition to cytoglobin, globin-E, globin-Y, and the independently evolved hemoglobin and myoglobin genes of cyclostomes (Hoffmann et al., 2012a).

Vertebrate-specific globins derive from a single ancestral gene present in the common ancestor of vertebrates, and their phylogenetic distribution among contemporary species reflects a complex history of lineage-specific duplications and deletions. Androglobin, the most recently discovered member of the vertebrate globins, is a chimeric protein that includes a rearranged globin domain that can be traced back to the common ancestor of choanoflagellates and animals (Hoogewijs et al., 2012). Neuroglobin also represents an ancient globin lineage that originated prior to the split between protostomes and deuterostomes (Burmester et al., 2000, 2002; Roesner et al., 2005; Dröge and Makałowski, 2011; Blank and Burmester, 2012; Hoffmann et al., 2012a), and in spite of intensive efforts, its physiological function remains a mystery (Fago et al., 2004; Ascenzi et al., 2014; Burmester and Hankeln, 2014; Keppner et al., 2020).

Almost every vertebrate examined possesses copies of hemoglobin and myoglobin in their genomes, and both cytoglobin and neuroglobin are present in the vast majority of vertebrate genomes surveyed as well (Hoffmann et al., 2011; Opazo et al., 2015). By contrast, the phylogenetic distribution of globin-E, globin-Y and GbX is more spotty, suggesting multiple independent gene losses. In the case of GbX, recently released genomes from key vertebrate taxa provide an excellent opportunity to address questions about its evolution. Globin-X is an especially enigmatic globin because it is predicted to be bound to the cell membrane (Blank et al., 2011), and because it has a very wide phyletic distribution that spans protostomes and deuterostomes (Blank and Burmester, 2012; Hoffmann et al., 2012a; Prothmann et al., 2020). There are GbX genes present in the genomes of insects, crustaceans, platyhelminthes, myriapods, spiders, hemichordates, and vertebrates among others, indicating that its origin predates the split between protostome and deuterostomes (Dröge and Makałowski, 2011). Unlike canonical globins such as hemoglobins and myoglobins, functional data suggest that GbX does not have a primary respiratory function. Instead, available evidence suggests that GbX may play a role in protecting cells from oxidative damage and in reducing nitrite (Corti et al., 2016; Koch and Burmester, 2016).

Whole-genome duplications have played a prominent role in the expansion and functional diversification of vertebrate-specific globins (Storz et al., 2011, 2013; Hoffmann et al., 2012b) and the hemoglobin gene repertoire of teleost fish (Opazo et al., 2013). Current evidence suggests that the repertoire of vertebrate GbX also expanded via WGDs. Initial genomic surveys of vertebrates revealed the presence of a single copy of GbX in a small number of distantly related vertebrate lineages that included some amphibians, some squamates, some teleost fish, elephant fish, and sea lamprey, which were all assumed to be 1-to-1 orthologs of each other (Roesner et al., 2005; Dröge and Makałowski, 2011; Hoffmann et al., 2012a). As the sample of vertebrate genomes increased, it became clear that there were different GbX genes in vertebrates (Opazo et al., 2015), and that apparent orthology among single copy GbX genes was the product of ‘hidden paralogy’, where genes are mistakenly identified as orthologs because of reciprocal, lineage-specific losses of alternative paralogs (Kuraku, 2010). Variation in the number of GbX paralogs among taxa and synteny comparisons suggest that whole-genome duplications (WGDs) were responsible for the expanded repertoire of GbX genes in
Evolution of globin X genes in vertebrates, and that subsequent lineage-specific WGDs also contributed to the increased GbX copy number in teleosts and salmonids (Opazo et al., 2015; Gallagher and Macqueen, 2017). However, questions remain regarding 1) the diversity of GbX genes in non-teleost ray-finned fishes, 2) the relationship of one of the GbX genes of cartilaginous fish and the GbX genes of bony vertebrates, and 3) the relationships between the GbX genes of cyclostomes and gnathostomes. Accordingly, the goal of this study is to unravel the duplicative history and diversification of GbX in vertebrates to address these questions. In addition, we analyze newly released genomes from vertebrate groups that experienced additional rounds of WGD to track the evolutionary fate of the GbX genes. Specifically, we integrate synteny and phylogenetic analyses to decipher the evolution of the GbX repertoire of cyclostomes and cartilaginous fish, and we examine the role of WGDs in the diversification of the GbX gene repertoire in several fish and amphibian taxa that experienced lineage-specific WGDs subsequent to the two rounds of WGD in the stem lineage of vertebrates. Our phylogenies also identify a highly divergent GbX paralog in several teleost fish, which might reflect the emergence of a functionally distinct GbX protein.

2 Materials and Methods
2.1 Bioinformatic searches
We combined bioinformatic searches for GbX-like sequences in vertebrate genomes in the National Center for Biotechnology Information (NCBI) (Sharma et al., 2018) and the Ensembl v.101 databases (Yates et al., 2020), some of them coming from the Vertebrate Genomes Project (Rhie et al., 2020). Our searches were seeded with known GbX paralogs identified in Opazo et al. (2015) from coelacanth, elephant fish (Callorhinchus milii, which is also referred to as elephant shark), spotted gar, and zebrafish. We first retrieved all putative GbX orthologs and paralogs of vertebrates from Ensembl v.101. We then extended our searches to include additional vertebrate genomes available in NCBI from lineages not well-represented in previous studies (Hoffmann et al., 2012a; Opazo et al., 2015; Gallagher and Macqueen, 2017). Importantly, our surveys include a much wider array of vertebrate lineages, allowing us to perform a much more comprehensive survey of the diversity of their GbX repertoires. We now include multiple cyclostomes, multiple cartilaginous fish, multiple amphibians, more squamates, more teleost fish, recently released genomes from non-teleost ray-finned fish (Du et al., 2020; Bi et al., 2021), plus the tuatara (the single extant representative of the order Rhynchocephalia). Because WGDs appear to have played an important role in the expansion of the vertebrate GbX repertoire, we purposely included genomes from representatives of vertebrate groups that have undergone additional lineage-specific WGSs. Such taxa include the sterlet (Du et al., 2020), salmonids (Berthelot et al., 2014; Lien et al., 2016), members of the subfamily Cyprininae (Xu et al., 2014, 2019; Chen et al., 2019), and the African clawed frog (Xenopus laevis) (Session et al., 2016). In the case of the Pacific lamprey, Entosphenus tridentatus (Hess et al., 2020), the pouched lamprey, Geotria australis, and the southern lamprey, Mordacia mordax, we annotated GbX genes by pairwise comparisons with the GbX genes of the sea lamprey, Petromyzon marinus, using BLAST (Altschul et al., 1990) and the ‘Blast 2 sequences’ tool (Tatusova and Madden, 1999). Similarly, we used the GbX genes from elephant fish to search for unannotated GbX paralogs in additional genomes from other cartilaginous fishes. Finally, as outgroup sequences, we included the full repertoire of globins from the acorn worm (Saccoglossus kowalevskii, Hemichordata), an invertebrate representative of deuterostomes that possesses the most diverse globin repertoire in the group (Hoffmann et al., 2012a).

We verified the identity of candidate GbX genes by reciprocal BLAST, comparing putative GbX sequences against the non-redundant protein sequence database (nr) of deuterostomes.

2.2 Sequence Alignment and Phylogenetic Analyses
We aligned amino acid sequences using the L-INS-i strategy from MAFFT v 7.471 (Katoh, 2005; Katoh et al., 2019) and estimated phylogenetic relationships using IQ-Tree v.2.0.6 (Minh et al., 2020). Support for the nodes was evaluated with the Shimodaira-Hasegawa approximate likelihood-ratio test and the aBayes tests (Anisimova et al., 2011) plus 10,000 pseudoreplicates of the ultrafast bootstrap procedure (Hoang et al., 2018). The best-fitting model of substitution was selected using the ModelFinder subroutine from IQ-Tree v.2.0.6 (Kalyaanamoorthy et al., 2017). Competing phylogenetic hypotheses were compared using the approximately unbiased test (Shimodaira, 2002a) as implemented in IQ-Tree v.2.0.6 (Minh et al., 2020).

2.3 Data curation
Because some of the sequences retrieved were annotated as neuroglobins or cytoglobins, and our searches potentially yielded redundant results, we first performed a phylogenetic analysis with all GbX candidates, to confirm the GbX identity of all retrieved sequences, to identify redundant sequences, and to detect potential annotation problems as evidenced by unusually long branches. Thus, for the second round of analyses, we removed all acorn worm globins other than 7, 8, 9, 10, and 16, which had already been shown to be the most closely related to GbX (Hoffmann et al., 2012a; Prothmann et al., 2020). We removed redundant records and retained a single representative species per genus, with the exception of the carp, where we kept two separate assemblies that include two separate duplications. Finally, we also removed truncated genes from the dataset. In the cases of African clawed frog, medaka, and zebrafish, the Ensembl and NCBI sequences are almost identical, so we only kept one. In the cases of elephant fish, sea lamprey, and coelacanth, we removed the Ensembl sequences and we only used records derived from the NCBI database due to better coverage and availability of synteny information. For example, the tree in supplementary figure S1 and synteny comparisons show that the truncated sea lamprey GbX paralog ENSPMAG00000007241 from Ensembl, which comes from an earlier assembly (Pmarinus_7.0), corresponds to the full-length gene LOC116943182 from NCBI, which comes from a more recent chromosome-level assembly (kPetMar1.pri) and includes better-resolved synteny. Finally, we discarded unusually short or long GbX candidates such as the ENSMALG00000006192 gene from the swamp eel, Monopterus albus, which is 376 amino acids long, and all of the putative GbX genes from the southern lamprey. Details on data curation are provided in supplementary table S1.

2.4 Synteny analyses
We explored the genomic context of the GbX genes in the Ensembl database v.100 (http://apr2020.archive.ensembl.org/index.html) by analyzing the presence of syntenic genes in vertebrate genomes with the help of the Genomicus browser v.100.01 (Nguyen et al., 2018). In the case of genomes not available in Ensembl, we checked synteny using the corresponding NCBI gene page, in combination with BLAST searches. Finally, because the presence of different PLEKHG paralogs has been used to define the genomic context of the different GbX paralogs of vertebrates (Opazo et al., 2015; Gallagher and Macqueen, 2017), we estimated phylogenetic relationships, following the same procedure mentioned above, for the PLEKHG1-3 paralogs of vertebrates to confirm our inferences of orthology. These analyses followed the same protocols used for the GbX phylogenies: we aligned amino acid sequences using the L-INS-i strategy from MAFFT v 7.471 (Katoh, 2005; Katoh et al., 2019) and estimated phylogenetic relationships using IQ-Tree v.2.0.6 (Minh et al., 2020) under the best-fitting model of substitution selected by the ModelFinder subroutine from IQ-Tree v.2.0.6 (Kalyaanamoorthy et al., 2017).

3 Results
3.1 Data description and nomenclature
We obtained 139 putative vertebrate GbX sequences from the Ensembl database, release 101, last accessed on September 28th, 2020, corresponding to the Ensembl gene tree ENSGT00730000111686, with representatives from jawless fish, ray-finned fish, squamates, testudines, tuatara, amphibians, and lobe-finned fish. We added an additional 25 sequences representing GbX candidates from testudines, amphibians, squamates, cartilaginous fish, lobe-finned fish, ray-finned fish, cyclostomes, plus the complete repertoire of globins from the acorn worm (Supplementary Table 1). As in previous studies, we did not find traces of GbX in the genomes of crocodilians, birds, or mammals despite the increased availability of genomes for these groups.

For the sake of consistency with previous studies, we followed the nomenclature from Gallagher and Macqueen (2017) in labeling orthologs of jawed vertebrates (gnathostomes). This nomenclature integrates information from phylogenetic and synteny analyses to infer orthology. Thus, orthologs of the spotted gar GbX1 gene ENSLOC0000014709, which is flanked by PLEHG2 and SUPT5, were labeled as GbX1 genes, and orthologs of the spotted gar GbX2 gene ENSLOC0000012798, which is flanked by PLEKHG3 and SRP14, were labeled as GbX2 genes. Within teleosts, which include duplicates of GbX2, orthologs of the GbX2a gene from Northern pike (Esox lucius, ENSELUG00000004427), which is flanked by a copy of PLEKHG3 and SRP14, were labeled as GbX2a genes, whereas orthologs of the GbX2b gene from the Northern pike (ENSELUG0000016373), which is flanked by PAPLNB and another copy of PLEKHG3, were labeled as GbX2b genes. Additional duplicates were identified by adding alternating letters and numbers to the name, as in the case of the sterlet GbX1a, GbX1b, GbX2a, and GbX2b genes, or the salmonid GbX2a1 and GbX2a2 genes. In these cases, similarities in the name of the paralogs do not reflect a shared duplicative history. In the case of cyclostomes, we labeled orthologs of the sea lamprey gene LOC116943182 (flanked by PLEKHG3 and SRP14) as GbX-C2, and orthologs of the sea lamprey gene LOC116948349 (flanked by another copy of PLEKHG3, TMEM180, and PYGL) as GbX-C1.

3.2 Phylogenetic analyses

We first estimated a maximum likelihood phylogenetic tree for the initial dataset of 190 sequences to ensure they were true GbX genes (available as Vert_GbX.190.fasta, Supplementary Online Material). The resulting tree placed all putative vertebrate GbX genes in a strongly supported monophyletic group that is sister to the clade that includes acorn worm globins 7, 8, 9,10, and 16, as in previous studies (Hoffmann et al., 2012a; Opazo et al., 2015). This tree confirmed the GbX identity of all the sequences we retrieved, and placed the GbX1 genes of gnathostomes and the GbX-C1 and GbX-C2 genes of cyclostomes in monophyletic groups, but the gnathostome GbX2 sequences were paraphyletic relative to the GbX1, GbX-C1, and GbX-C2 genes. These initial analyses also identified a duplication of GbX1 in the sterlet, duplications of the GbX2 gene in the Leishan spiny toad and the sterlet, and duplications of the teleost GbX2a gene in the subfamily Cyprininae in addition to the duplication of the GbX2a paralog in salmonids that was previously identified by Gallagher and Macqueen (Gallagher and Macqueen, 2017). In the case of the Leishan spiny toad and the common carp, the two paralogs are found on the same genomic fragment (Supplementary Table 1), suggesting that they derive from single-gene tandem duplications. A closer inspection of the tree revealed several unusually long branches that may have been caused by the inclusion of low-quality sequences, such as the GbX paralogs from the southern lamprey, which were excluded from further analyses. Importantly, this analysis demonstrated that species within a genus share a common GbX repertoire. The only exception was the absence of a GbX2b paralog in the orange clownfish, Amphiprion percula, relative to the clown anemonefish, Amphiprion ocellaris. The estimated phylogeny in combination with synteny comparisons revealed the presence of redundant paralogs in coelacanth, elephant fish, Western clawed frog, sea lamprey, medaka, and zebrafish.
In the second round of analyses, we removed all acorn worm globins other than 7, 8, 9, 10, and 16, we
removed redundant records, we retained a single representative species per genus (with the exception
of the carp, where we kept two separate assemblies that include two separate duplications), and we
removed truncated genes (available as Vert_GbX.134.fasta, Supplementary Online Materia). As in the
initial analysis, the resultant tree placed the GbX1 genes of gnathostomes and the GbX-C1 and GbX-
C2 genes of cyclostomes in monophyletic groups, but GbX2 sequences were paraphyletic relative to
the clade that included the GbX1, GbX-C1, and GbX-C2 genes, supplementary figure 2. In this tree,
relationships for the GbX2 sequences deviated quite strongly from the expected organismal
relationships. In particular, the GbX2a ohnolog (paralog derived from WGD) of teleost fish was split
into three separate clades, and the GbX2b ohnolog of teleosts was split into multiple lineages that were
placed as the deepest divergences of vertebrate GbXs. A strict reconciliation of this maximum
likelihood tree with the organismal tree would imply the presence of over ten GbX paralogs in the last
common ancestor of vertebrates with a large number of independent gene losses in descend
lineages, and would also imply that the duplication giving rise to the teleost GbX2a and 2b ohnologs
traces back to the vertebrate ancestor. However, a tree that minimizes these independent deletions,
where the GbX1 and GbX2 genes of gnathostomes were constrained to be monophyletic, and the
GbX2a and GbX2b ohnologs of teleosts were constrained to be monophyletic within GbX2s and sister
to each other, was not significantly different from the unconstrained tree (Table 1). Because this
constrained tree minimizes the inferred number of independent gene gains and losses, it agrees well
with assessments of conserved synteny among gnathostomes, and maps their duplication to the stem of
the teleosts on the organismal tree, we selected it as most plausible phylogenetic hypothesis and use it
as the basis of our evolutionary inferences. In this constrained tree (Fig 2, supplementary figure 3), the
GbX1 and GbX2 paralogs of gnathostomes were placed sister to each other, and the GbX-C1 and
GbX-C2 paralogs of cyclostomes were placed sister to each other as well.

Each of the four lamprey species in the final analyses includes single-copy representatives of the GbX-
C1 and GbX-C2 paralogs, and each of the paralog subtrees recovers the same species relationships.
This pattern indicates that each of the cyclostome GbX paralogs can be traced back to the last common
ancestor of these lamprey species. In addition, even though we only found a copy of GbX-C2 in
hagfish, its position on the trees as sister to the clade of GbX-C2 sequences of lampreys, also
supported by the available synteny data, indicates that the duplication that gave rise to GbX-C1 and C2
predates the split between hagfish and lampreys, and suggests that the hagfish secondarily lost the
GbX-C1 paralog or that the apparent absence of the gene represents an assembly artifact.

In the case of gnathostome GbX genes, the GbX1 sequences from 1- cartilaginous fish, 2- ray-finned
fishes, caecilians, 3- squamates, and 4- testudines were each placed in monophyletic groups; the single
GbX1 gene from coelacanth is placed sister to the GbX1 genes from ray-finned fishes; and the GbX1
gene from tuatara is placed sister to the GbX1 genes from squamates (supplementary figure 3). As in
previous studies (Opazo et al., 2015; Gallagher and Macqueen, 2017), we did not find traces of GbX1
in the genomes of teleost fishes in spite of our much denser sampling relative to earlier studies. The
sterlet genome represents the only case where we identified duplicate copies of GbX1, and these were
placed in a monophyletic group, sister to the GbX1 of the reedfish, with the spotted gar GbX1 as sister
to them. Relationships among the GbX1 genes of ray-finned fishes were not congruent with known
organismal relationships. The estimated gene tree placed the reedfish paralog sister to the 2 sterlet
paralogs instead of spotted gar GbX1, but the corresponding branches were very short, so we ignored
this discrepancy in our inferences. In this tree, caecilian GbX1 genes are placed sister to the clade that
includes cartilaginous fish, ray-finned fish and lobe-finned fish sequences, instead of being placed
sister to amniote GbX1s, but again, the corresponding branches were very short, so we did not attach
Evolution of globin X genes

importance to this apparent discrepancy. Relationships among the GbX1 sequences from amniotes matched the expected organismal relationships.

In the constrained tree (Fig 2, supplementary figure 3), relationships among the GbX2 genes matched organismal relationships at the order level. The putative elephant fish GbX2 ortholog is placed as sister to the clade that includes the GbX2 sequences from bony vertebrates, and relationships within the latter group were congruent with expected organismal relationships. The two tandem duplicates of the Leishan spiny toad are sister to each other in all analyses, and the two tandem duplicates of the carp are also close to each other in the tree, which suggests that they are very recent duplications (supplementary figures 1-3). The duplicate GbX2a paralogs of salmonids match organismal relationships (supplementary figures 1-3), consistent with the hypothesis that they trace their origins to the salmonid-specific WGD (Gallagher and Macqueen, 2017). The cyprinine-specific WGD appears to also have given rise to duplicate GbX2a paralogs in golden-line fish, carp, and goldfish. Here, however, the estimated tree groups the cyprinine paralogs by genus, with the exception of the tandem duplicate of the GbX2a paralog of the carp, ENSCCRG00000022746, which is grouped with the GbX2a paralogs of golden-line fishes (supplementary figures 1-3). A strict reconciliation of this subtree with the species tree would imply that golden-line fish (genus Sinocyclocheilus), goldfish (genus Carassius) and carp (genus Cyprinus) would each have undergone an independent WGD. However, our study involves a single gene family and we lack evidence for this from synteny comparisons.

Cyclostome genes are unusual with respect to nucleotide and amino acid composition and they also exhibit peculiar codon usage biases, which make it challenging to use phylogenetic approaches to resolve orthology between cyclostome and gnathostome genes (Qiu et al., 2011; Kuraku, 2013). In many instances, analyses of synteny have provided additional and independent information to resolve ambiguous gene phylogenies (Kuraku and Meyer, 2012; Campanini et al., 2015). In the case of GbX, the GbX2 genes of gnathostomes and the GbX-C2 genes of cyclostomes are flanked by copies of the PLEKHG3 and SRP14 genes, strongly suggesting they are 1-to-1 orthologs. However, a phylogenetic scenario where gnathostome GbX2 was constrained to be sister to the cyclostome GbX-C2 (as would be expected if the two genes were 1-to1 orthologs) was statistically rejected in our tree topology tests (Table 1).

Synteny analyses in Opazo et al. (2015) suggest that the GbX1 and GbX2 genes of gnathostomes derived from one of the two rounds of WGD early in vertebrate evolution. If the GbX2 of gnathostomes and the GbX-C2 of cyclostomes derived from the same duplication, the syntenic genes that co-duplicated with them would be expected to reflect the same duplicative history, so the PLEKHG3 gene of gnathostomes that is adjacent to GbX2 would be sister to the PLEKHG3 gene of cyclostomes that is adjacent to GbX-C2. We explored this issue by estimating phylogenetic relationships among the gnathostome and cyclostome PLEKHG2 and PLEKHG3 genes, which are used to define the genomic context of the vertebrate GbX genes and validate inferences of orthology (Opazo et al., 2015; Gallagher and Macqueen, 2017). The estimated PLEKH phylogeny recapitulates the results of the GbX phylogeny, placing the two PLEKHG3 paralogs of cyclostomes in a monophyletic group, as their neighboring GbX-C1 and GbX-C2 genes in the GbX tree, and placing the PLEKHG2 and PLEKHG3 genes of gnathostomes in a monophyletic group as their neighboring GbX1 and GbX2 genes in the GbX phylogeny. Thus, both synteny and phylogenetic analyses indicate that the GbX-C1 and C2 genes of cyclostomes and the GbX1 and GbX2 genes of gnathostomes are products of independent duplication events.

3.3 Expression of the different GbX paralogs
The available evidence indicates that the GbX paralogs of elephant fish, spotted gar, and salmon are differentially expressed (Opazo et al., 2015; Gallagher and Macqueen, 2017). In the elephant fish, RNA-seq data indicates that the GbX1 and GbX2 paralogs are mostly expressed in the spleen, and GbX1 (AKU74647), which is labeled as GbX2 in the original study of Opazo et al. (2015), is also expressed in brain, spleen and testis, whereas GbX2 (XP 007891388), which is labeled a GbX1 in Opazo et al. (2015), is expressed in brain, gills, intestine, kidney and liver (Opazo et al., 2015). In the case of spotted gar, quantitative PCR data indicate that GbX1 is most highly expressed in the brain, and expression is also detected in heart, gill, liver, intestine and spleen, whereas expression of GbX2 is only detected in the brain (Gallagher and Macqueen, 2017). In the case of salmonids, quantitative PCR data revealed the following: 1) expression of the GbX2a1 paralog (ENSSSAG00000003360) is highest in the brain and it is also expressed in intestine and eye; 2) expression of the GbX2a2 paralog (ENSSSAG00000007165) was not detected; and 3) expression of the GbX2b paralog (ENSSSAG000000047904) is highest in the intestine and is also detected in the brain, stomach, and eye (Gallagher and Macqueen, 2017). These data suggest that these genes are expressed in a variety of tissues, but that patterns of expression vary in a lineage-specific manner that is not easy to put into the evolutionary framework of the organismal tree at this stage.

4 Discussion

After performing an exhaustive homolog search using genome-wide sequence resources, we integrated phylogenetic and synteny analyses to infer the duplicative history of GbX paralogs in vertebrates. Because these analyses included highly contiguous cyclostome genomes plus newly released genomes from cartilaginous fish as well as representatives of the deepest-branching lineages of ray-finned fishes (Du et al., 2020; Bi et al., 2021), we were able to resolve unanswered questions regarding the early stages of evolution of GbX genes in vertebrates. Since the time of its initial discovery (Roesner et al., 2005), GbX went from being an obscure gene found in a very limited sample of vertebrates to becoming a credible candidate to provide clues about the functional role of the ancestor of all animal globins (Blank et al., 2011; Song et al., 2020). This change in paradigm came with an increased interest in deciphering its evolutionary history and its still elusive functional role (Burmester and Hankeln, 2014; Keppner et al., 2020). We have recently documented the presence of GbX paralogs in arthropods (Prothmann et al., 2020), confirming phylogenetic predictions that indicate that the origin of GbX predates the split between deuterostomes and protostomes (Burmester et al., 2002; Roesner et al., 2005; Dröge and Makalowski, 2011; Blank and Burmester, 2012; Hoffmann et al., 2012a)(Opazo et al., 2015), which is estimated to have occurred ~ 730 million years ago (Kumar et al., 2017).

4.1 Evolution of GbX in early vertebrates

Our reconstructions shed light on the early stages of evolution of the GbX genes in vertebrates. Our increased sampling allows us to resolve orthology for the two different GbX genes of cartilaginous fish relative to the rest of the gnathostomes. In turn, this implies that the two GbX paralogs of gnathostomes, GbX1 and GbX2, can be traced back to the last common ancestor of cartilaginous fish and bony vertebrates, and synteny analyses suggest that these two paralogs are actually ohnologs that derive from one of the two possible rounds of WGD early in vertebrate evolution, in agreement with Opazo et al. (2015). Similarly, the two GbX paralogs of cyclostomes, GbX-C1 and GbX-C2, can be traced back to the last common ancestor of hagfishes and lampreys. Synteny analyses of the GbX genes in the sea lamprey and the pouched lamprey reveal the shared presence of PLEKHG3 genes next to GbX-C1 and GbX-C2. The conserved synteny and phylogenetic analyses suggest that the PLEKHG3 and GbX genes of cyclostomes coduplicated, which would indicate that they derive from a segmental duplication or even a WGD. In this regard, our results are consistent with previous studies that indicate that lamprey genomes underwent extensive segmental duplications (Smith and Keinath,
Our results provide strong evidence that the gnathostome GbX paralogs derive from one of the two vertebrate-specific WGDs, either 1R or 2R, confirming inferences from Opazo et al. (2015) (Fig 3). In addition, the GbX-C1 and C2 paralogs appear to derive from segmental duplications involving additional genes, and this segmental duplication could correspond to a WGD. Reconciling the observed relationships between GbX1, GbX2, GbX-C1 and GbX-C2, with the organismal phylogeny is not trivial, especially given uncertainty about the timing and number of WGDs that occurred early in vertebrate evolution. There is agreement that early vertebrates underwent 2 rounds of WGD (Meyer and Schartl, 1999; McLysaght et al., 2002; Dehal and Boore, 2005), 1R and 2R, and there is also agreement that cyclostomes and gnathostomes share 1R. There is less agreement about the placement of 2R on the vertebrate tree (Kuraku et al., 2009). Some authors place 2R in the common ancestor of cyclostomes and gnathostomes (Sacerdot et al., 2018), whereas others place 2R in the last common ancestor of gnathostomes (Simakov et al., 2020), and there is also the suggestion that cyclostomes might have undergone an additional and independent WGD early in their evolution (Mehta et al., 2013). Our results do not fit any of these scenarios better than the alternatives, and would require multiple gene losses to fit in any of them.

4.2 Evolution of the GbX paralogs of gnathostomes and cyclostomes

The GbX paralogs of gnathostomes and cyclostomes have followed contrasting evolutionary trajectories. Both the GbX1 and GbX2 paralogs of gnathostomes and the GbX-C1 and GbX-C2 paralogs of cyclostomes can be traced back to the common ancestor of each group, but whereas the GbX1 and GbX2 paralogs of gnathostomes have been lost independently multiple times and have undergone additional duplications, the GbX-C1 and GbX-C2 paralogs of cyclostomes have been retained by all lampreys (Fig 3). It appears that hagfishes may have secondarily lost GbX-C1, but a more contiguous assembly is needed for confirmation. Among gnathostomes, spotted gar, elephant fish, sterlet, and coelacanth have retained both GbX1 and GbX2 paralogs, whereas mammals and archosaurs (birds + crocodilians) have lost both. GbX1 was independently lost in teleost fish, mammals, anurans, and archosaurs, whereas GbX2 was independently lost in caecilians, amniotes, and reedfish. Transcriptomic data indicate that only GbX1 has been retained in salamanders (Queiroz et al., 2020), which would imply an additional independent loss of GbX2.

Among ray-finned fish, the sterlet possesses a total of 4 GbX copies - the most of any vertebrate examined to date - as this species has retained both of the GbX1 and GbX2 duplicates derived from the WGD specific to this lineage (Du et al., 2020). Reedfish has only retained a copy of GbX1, gar has retained copies of both GbX1 and GbX2, and teleosts have only retained copies of GbX2 (Fig 1, supplementary figures 1-3). The GbX2a and GbX2b paralogs of teleost fish appear to derive from the teleost-specific WGD (Gallagher and Macqueen, 2017), and they have also been differentially retained among species. The GbX2 paralog has been retained in 83 out of 84 teleost fish in ensembl v101. The only exception is the Chinese medaka (Oryzias sinensis), and because of the overall low gene coverage of this assembly, we suspect this to be an artefact. By contrast, the GbX2b paralog has been retained in 25 out of 84 teleost fish genomes examined — representing 8 different higher-level lineages and implying multiple independent losses (Supplementary Table S1). The GbX2a and GbX2b paralogs of salmonid fishes also exhibit a highly asymmetric rates of gene retention: all six examined salmonid genomes retained the two GbX2a ohnologs derived from the salmonid-specific WGD (Gallagher and Macqueen, 2017), but only a single copy of GbX2b, suggesting that the latter gene reverted to a diploid state shortly after the salmonid-specific WGD (supplementary figures 1-3). Similarly, the duplicate GbX2 paralogs of the subfamily Cyprininae had contrasting fates. The GbX2a paralog duplicated in
the cyprinine-specific WGD and has been retained in most cases, whereas the GbX2b paralog was apparently lost earlier, in the common ancestor of cyprinines and zebrafish.

In addition to showing contrasting levels of synteny conservation and retention rates, the GbX2a and GbX2b paralogs of teleost fish also seem to follow diverging evolutionary pathways in terms of genomic context and amino acid changes. The genomic context of the teleost fish GbX2a gene is highly conserved, and in the vast majority of the cases, the GbX2a gene is flanked by SRP14 and PLEKHG3, as in the ancestral GbX2 gene. In the case of GbX2b, however, the genomic context is more variable, although in most cases there is a copy of the CEP170 gene in the vicinity, and sometimes there is a copy PLEKHG3 as well. The GbX2a and GbX2b genes also exhibit different rates of amino acid substitution rates, as evidenced by the longer branches in the GbX2b portion of the phylogenetic trees. Asymmetric rates of evolution are often associated with differences in evolutionary constraints among sister paralogs, and the GbX2a and 2b paralogs of teleost fish appear to fit this pattern well (Pál et al., 2006).

Our synteny and phylogenetic analyses indicate that WGDs have played a major role in the diversification of GbX paralogs, just as they did in the vertebrate-specific globins (Hoffmann et al., 2012b; Opazo et al., 2013, 2015; Storz et al., 2013). Also like in vertebrate-specific globins, the repertoire of GbX in extant lineages has been impacted by lineage-specific losses and the differential retention of relatively old duplicates. Specifically, our results indicate that WGDs have given rise to at least 6 pairs of duplicates, that GbX1 was lost at least 4 independent times, and that GbX2 was lost at least 8 times independently (Fig 3), plus multiple additional losses of the GbX2b gene in teleost fishes. Starting with the oldest, the first pair of WGD-derived duplicates corresponds to the GbX1 and GbX2 ohnologs of gnathostomes, which derive from either the 1R or 2R WGDs early in vertebrate evolution. The second pair corresponds to the GbX2a and GbX2b ohnologs of teleost fish, which derive from the teleost WGD, and the third corresponds to the duplicate copies of GbX2a in salmonids, which derive from the salmonid-specific WGD, both identified by Gallagher and Macqueen (Gallagher and Macqueen, 2017). The fourth pair corresponds to the duplicate copies of GbX2a in cyprinins, which derive from the cyprinin-specific WGD, and finally, the fifth and sixth pair correspond to duplicate copies of both GbX1 and GbX2 that derive from the sterlet-specific WGD. The Xenopus-specific WGD is the only vertebrate WGD that does not appear to have produced an expansion of the GbX gene repertoire, and the possibility that the GbX-C1 and GbX-C2 duplicates of cyclostomes derive from an WGD as well.

Our study highlights the role that WGDs have played in expanding the repertoire of genes in vertebrate genomes. Our results indicate that GbX paralogs have a remarkably high rate of retention following WGDs in comparison to other globin genes. Our results also provide an evolutionary framework for interpreting results of experiments that examine functional properties of GbX and patterns of tissue-specific expression. By identifying GbX ohnologs that are products of different WGDs during the radiation of vertebrates, our results can guide the design of experimental work to explore whether gene duplicates that originate via WGDs have evolved novel functional properties or expression profiles relative to singleton or tandemly duplicated copies of GbX in the same species.

5 Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

6 Author contributions
FGH designed the study. FGH, SK, JCO, and MWW collected and/or analyzed data. FGH, SK, JCO, and JFS wrote the manuscript. FGH, SK, JCO, and JFS reviewed and edited the manuscript. All authors contributed to the article and approved the submitted version.

7 Funding

FGH acknowledges support from the National Science Foundation (OIA-1736026), the Mississippi Agricultural and Forestry Experiment Station, and the National Institute of Food and Agriculture, U.S. Department of Agriculture, Hatch project under accession number MIS-399150. JCO acknowledges support from Fondo Nacional de Desarrollo Científico y Tecnológico from Chile (FONDECYT 1210471) and Millennium Nucleus of Ion Channels Associated Diseases (MiNICAD), Iniciativa Científica Milenio, Ministry of Economy, Development and Tourism. JFS acknowledges support from the National Science Foundation (OIA-1736249) and National Institute of Health (HL087216). SK acknowledges support from the JSPS KAKENHI (20H03269).

8 Acknowledgements

FGH thanks Amanda Coward Black for editorial assistance. JCO wants to acknowledge the members of the Integrative Biology Group, Universidad Austral de Chile for their constant support, scientific enthusiasm, and creative feedback.

Table 1. Results of topology tests

| Tree                                      | logL    | ΔL | bp-RELL | p-KH | p-SH | c-ELW | p-AU |
|-------------------------------------------|---------|----|---------|------|------|-------|------|
| Unconstrained                             | -20600.3| 0  | 0.428   | 0.617| 1    | 0.428 | 0.629|
| GbX2 monophyletic                         | -20604.5| 4.2| 0.199   | 0.383| 0.789| 0.199 | 0.454|
| GbX2, GbX2a, & GbX2b monophyletic         | -20604.9| 4.6| 0.359   | 0.412| 0.6  | 0.357 | 0.441|
| GbX2 sister to GbX-C2                     | -20640.4|40.1|0.0148   |0.0466|0.121|0.015 |0.036|

bp-RELL : bootstrap proportion using RELL method (Kishino et al., 1990).
p-KH : p-value of one sided Kishino-Hasegawa test (Kishino and Hasegawa, 1989).
p-SH : p-value of Shimodaira-Hasegawa test (Shimodaira and Hasegawa, 1999).
c-ELW : Expected Likelihood Weight (Strimmer and Rambaut, 2002)).
p-AU : p-value of approximately unbiased (AU) test (Shimodaira, 2002b).
Figure legends

Figure 1. Maximum likelihood phylogram describing evolutionary relationships among the vertebrate globin-X candidates identified in our study. The tree was rooted with the full set of Acorn worm globins. The tree with the terminal labels is available as Supplementary Figure 1, and corresponding alignment is available as Supplementary Data File 1.

Figure 2. Maximum likelihood phylogram describing evolutionary relationships among the curated set of vertebrate globin-X candidates in our study, where the the GbX1 and GbX2 genes of gnathostomes were constrained to be monophyletic, and the GbX2a and GbX2b ohnologs of teleosts were constrained to be monophyletic within GbX2s and sister to each other. This tree was not statistically different from an unconstrained tree, which is available as Supplementary Figure 2, and minimizes the number of independent gene gains and losses. The tree was rooted with Acorn worm globins 7, 8, 9, 10, and 16. The tree with the terminal labels is available as Supplementary Figure 3, and corresponding alignment is available as Supplementary Data File 2.

Figure 3. Graphical summary of the role of WGDs in the expansion of the vertebrate GbX repertoire. Organismal relationships on the right, and the number of GbX paralogs per lineage on the left. WGDs, segmental duplications (SD), and gene losses are mapped to their corresponding branch. We placed the 1R and 2R WGDs following Sacerdot et al. (2018). Symbols on a branch are arranged according to their relative order. Asterisks on top of the vertical bars denote that these duplications gave rise to GbX paralogs present in extant species. Because we cannot determine which of the 1R or 2R WGD in the vertebrate ancestor gave rise to GbX1 and GbX2, we placed the asterisk between them. Derive from either 1R or 2R, and our analyses cannot determine which of the two. In the case of cyclostomes, our analyses indicate that GbX-C1 and GbX-C2 probably derive from a segmental duplication or a WGD. Note that a full tree of all species examined would include 2 additional tandem duplications and multiple additional gene losses.

Supplementary Figure 1. Unconstrained Tree with 190 sequences with labels
Supplementary Figure 2. Unconstrained Tree with 134 sequences with labels
Supplementary Figure 3. Constrained Tree with 134 sequences with labels
Supplementary Figure 4. PLEKHG tree. We might need to move this figure to the main text.
Evolution of globin X genes

References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410.

Anisimova, M., Gil, M., Dufayard, J.-F., Dessimoz, C., and Gascuel, O. (2011). Survey of branch support methods demonstrates accuracy, power, and robustness of fast likelihood-based approximation schemes. *Syst. Biol.* 60, 685–699.

Ascenzi, P., Gustincich, S., and Marino, M. (2014). Mammalian nerve globins in search of functions. *IUBMB Life* 66, 268–276.

Berthelot, C., Brunet, F., Chalopin, D., Juanchich, A., Bernard, M., Noël, B., et al. (2014). The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates. *Nat. Commun.* 5, 3657.

Bi, X., Wang, K., Yang, L., Pan, H., Jiang, H., Wei, Q., et al. (2021). Tracing the genetic footprints of vertebrate landing in non-teleost ray-finned fishes. *Cell*. doi:10.1016/j.cell.2021.01.046.

Blank, M., and Burmester, T. (2012). Widespread occurrence of N-terminal acylation in animal globins and possible origin of respiratory globins from a membrane-bound ancestor. *Mol. Biol. Evol.* 29, 3553–3561.

Blank, M., Wollberg, J., Gerlach, F., Reimann, K., Roesner, A., Hankeln, T., et al. (2011). A membrane-bound vertebrate globin. *PLoS One* 6, e25292.

Burmester, T., Ebner, B., Weich, B., and Hankeln, T. (2002). Cytoglobin: A Novel Globin Type Ubiquitously Expressed in Vertebrate Tissues. *Mol. Biol. Évol.* 19, 416–421.

Burmester, T., and Hankeln, T. (2014). Function and evolution of vertebrate globins. *Acta Physiol.* 211, 501–514.

Burmester, T., Weich, B., Reinhardt, S., and Hankeln, T. (2000). A vertebrate globin expressed in the brain. *Nature* 407, 520–523.

Campanini, E. B., Vandewege, M. W., Pillai, N. E., Tay, B.-H., Jones, J. L., Venkatesh, B., et al. (2015). Early Evolution of Vertebrate Mybs: An Integrative Perspective Combining Synteny, Phylogenetic, and Gene Expression Analyses. *Genome Biol. Évol.* 7, 3009–3021.

Chen, Z., Omori, Y., Koren, S., Shirokiya, T., Kuroda, T., Miyamoto, A., et al. (2019). De novo assembly of the goldfish (Carassius auratus) genome and the evolution of genes after whole-genome duplication. *Sci Adv* 5, eaav0547.

Corti, P., Xue, J., Tejero, J., Wajih, N., Sun, M., Stolz, D. B., et al. (2016). Globin X is a six-coordinate globin that reduces nitrite to nitric oxide in fish red blood cells. *Proc. Natl. Acad. Sci. U. S. A.* 113, 8538–8543.

Dehal, P., and Boore, J. L. (2005). Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biol.* 3, e314.

Dröge, J., and Makalowski, W. (2011). Phylogenetic analysis reveals wide distribution of globin X.
Evolution of globin X genes

Biol. Direct 6, 54.

Du, K., Stöck, M., Kneitz, S., Klopp, C., Woltering, J. M., Adolphi, M. C., et al. (2020). The sterlet sturgeon genome sequence and the mechanisms of segmental rediploidization. Nat Ecol Evol 4, 841–852.

Fago, A., Hundahl, C., Malte, H., and Weber, R. E. (2004). Functional properties of neuroglobin and cytoglobin. Insights into the ancestral physiological roles of globins. IUBMB Life 56, 689–696.

Gallagher, M. D., and Macqueen, D. J. (2017). Evolution and Expression of Tissue Globins in Ray-Finned Fishes. Genome Biol. Evol. 9, 32–47.

Hess, J. E., Smith, J. J., Timoshevskaya, N., Baker, C., Caudill, C. C., Graves, D., et al. (2020). Genomic islands of divergence infer a phenotypic landscape in Pacific lamprey. Mol. Ecol. 29, 3841–3856.

Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q., and Vinh, L. S. (2018). UFBoot2: Improving the Ultrafast Bootstrap Approximation. Mol. Biol. Evol. 35, 518–522.

Hoffmann, F. G., Opazo, J. C., Hoogewijs, D., Hankeln, T., Ebner, B., Vinogradov, S. N., et al. (2012a). Evolution of the globin gene family in deuterostomes: lineage-specific patterns of diversification and attrition. Mol. Biol. Evol. 29, 1735–1745.

Hoffmann, F. G., Opazo, J. C., and Storz, J. F. (2011). Differential loss and retention of cytoglobin, myoglobin, and globin-E during the radiation of vertebrates. Genome Biol. Evol. 3, 588–600.

Hoffmann, F. G., Opazo, J. C., and Storz, J. F. (2012b). Whole-genome duplications spurred the functional diversification of the globin gene superfamily in vertebrates. Mol. Biol. Evol. 29, 303–312.

Hoogewijs, D., Ebner, B., Germani, F., Hoffmann, F. G., Fabrizius, A., Moens, L., et al. (2012). Androglobin: a chimeric globin in metazoans that is preferentially expressed in Mammalian testes. Mol. Biol. Evol. 29, 1105–1114.

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A., and Jermiin, L. S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. Nat. Methods 14, 587–589.

Katoh, K. (2005). MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Research 33, 511–518. doi:10.1093/nar/gki198.

Katoh, K., Rozewicki, J., and Yamada, K. D. (2019). MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. Brief. Bioinform. 20, 1160–1166.

Keppner, A., Marie, D., Correia, M., Koay, T. W., Orlando, I. M. C., Vinogradov, S. N., et al. (2020). Lessons from the post-genomic era: Globin diversity beyond oxygen binding and transport. Redox Biol 37, 101687.

Kishino, H., and Hasegawa, M. (1989). Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea. J.
Evolution of globin X genes

Mol. Evol. 29, 170–179.

Kishino, H., Miyata, T., and Hasegawa, M. (1990). Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. J. Mol. Evol. 31, 151–160.

Koch, J., and Burmester, T. (2016). Membrane-bound globin X protects the cell from reactive oxygen species. Biochem. Biophys. Res. Commun. 469, 275–280.

Kumar, S., Stecher, G., Suleski, M., and Hedges, S. B. (2017). TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. Mol. Biol. Evol. 34, 1812–1819.

Kuraku, S. (2010). Palaeophylogenomics of the vertebrate ancestor--impact of hidden paralogy on hagfish and lamprey gene phylogeny. Integr. Comp. Biol. 50, 124–129.

Kuraku, S. (2013). Impact of asymmetric gene repertoire between cyclostomes and gnathostomes. Semin. Cell Dev. Biol. 24, 119–127.

Kuraku, S., and Meyer, A. (2012). Detection and phylogenetic assessment of conserved synteny derived from whole genome duplications. Methods Mol. Biol. 855, 385–395.

Kuraku, S., Meyer, A., and Kuratani, S. (2009). Timing of genome duplications relative to the origin of the vertebrates: did cyclostomes diverge before or after? Mol. Biol. Evol. 26, 47–59.

Lien, S., Koop, B. F., Sandve, S. R., Miller, J. R., Kent, M. P., Nome, T., et al. (2016). The Atlantic salmon genome provides insights into rediploidization. Nature 533, 200–205.

McLysaght, A., Hokamp, K., and Wolfe, K. H. (2002). Extensive genomic duplication during early chordate evolution. Nat. Genet. 31, 200–204.

Mehta, T. K., Ravi, V., Yamasaki, S., Lee, A. P., Lian, M. M., Tay, B.-H., et al. (2013). Evidence for at least six Hox clusters in the Japanese lamprey (Lethenteron japonicum). Proc. Natl. Acad. Sci. U. S. A. 110, 16044–16049.

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

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., von Haeseler, A., et al. (2020). IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Mol. Biol. Evol. 37, 1530–1534.

Nguyen, N. T. T., Vincens, P., Roest Crollius, H., and Louis, A. (2018). Genomicus 2018: karyotype evolutionary trees and on-the-fly synteny computing. Nucleic Acids Res. 46, D816–D822.

Opazo, J. C., Butts, G. T., Nery, M. F., Storz, J. F., and Hoffmann, F. G. (2013). Whole-genome duplication and the functional diversification of teleost fish hemoglobins. Mol. Biol. Evol. 30, 140–153.

Opazo, J. C., Lee, A. P., Hoffmann, F. G., Toloza-Villalobos, J., Burmester, T., Venkatesh, B., et al. (2015). Ancient Duplications and Expression Divergence in the Globin Gene Superfamily of Vertebrates: Insights from the Elephant Shark Genome and Transcriptome. Mol. Biol. Evol. 32, 1684–1694.
Evolution of globin X genes

Pál, C., Papp, B., and Lercher, M. J. (2006). An integrated view of protein evolution. *Nature Reviews Genetics* 7, 337–348. doi:10.1038/nrg1838.

Prothmann, A., Hoffmann, F. G., Opazo, J. C., Herbener, P., Storz, J. F., Burmester, T., et al. (2020). The Globin Gene Family in Arthropods: Evolution and Functional Diversity. *Front. Genet.* 11, 858.

Qiu, H., Hildebrand, F., Kuraku, S., and Meyer, A. (2011). Unresolved orthology and peculiar coding sequence properties of lamprey genes: the KCNA gene family as test case. *BMC Genomics* 12, 325.

Queiroz, J. P. F., Lima, N. C. B., and Rocha, B. A. M. (2020). The rise and fall of globins in the amphibia. *Comp. Biochem. Physiol. Part D Genomics Proteomics* 37, 100759.

Rhie, A., McCarthy, S. A., Fedrigo, O., Damas, J., Formenti, G., Koren, S., et al. (2020). Towards complete and error-free genome assemblies of all vertebrate species. *Cold Spring Harbor Laboratory*, 2020.05.22.110833. doi:10.1101/2020.05.22.110833.

Roesner, A., Fuchs, C., Hankeln, T., and Burmester, T. (2005). A globin gene of ancient evolutionary origin in lower vertebrates: evidence for two distinct globin families in animals. *Mol. Biol. Evol.* 22, 12–20.

Sacerdot, C., Louis, A., Bon, C., Berthelot, C., and Roest Crollius, H. (2018). Chromosome evolution at the origin of the ancestral vertebrate genome. *Genome Biol.* 19, 166.

Session, A. M., Uno, Y., Kwon, T., Chapman, J. A., Toyoda, A., Takahashi, S., et al. (2016). Genome evolution in the allotetraploid frog Xenopus laevis. *Nature* 538, 336–343.

Sharma, S., Ciufo, S., Starchenko, E., Darji, D., Chlumsky, L., Karsch-Mizrachi, I., et al. (2018). The NCBI BioCollections Database. *Database* 2018. doi:10.1093/database/bay006.

Shimodaira, H. (2002a). An Approximately Unbiased Test of Phylogenetic Tree Selection. *Systematic Biology* 51, 492–508. doi:10.1080/10635150290069913.

Shimodaira, H. (2002b). An approximately unbiased test of phylogenetic tree selection. *Syst. Biol.* 51, 492–508.

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

Simakov, O., Marlétaz, F., Yue, J.-X., O’Connell, B., Jenkins, J., Brandt, A., et al. (2020). Deeply conserved synteny resolves early events in vertebrate evolution. *Nat Ecol Evol* 4, 820–830.

Smith, J. J., and Keinath, M. C. (2015). The sea lamprey meiotic map improves resolution of ancient vertebrate genome duplications. *Genome Res.* 25, 1081–1090.

Song, S., Starunov, V., Bailly, X., Ruta, C., Kern, P., Cornelissen, A. J. M., et al. (2020). Globins in the marine Annelid Platynereis dumerilii shed new light on hemoglobin evolution in Bilaterians. *Cold Spring Harbor Laboratory*, 789214. doi:10.1101/789214.

Storz, J. F., Opazo, J. C., and Hoffmann, F. G. (2011). Phylogenetic diversification of the globin gene
Evolution of globin X genes

superfamily in chordates. *IUBMB Life* 63, 313–322.

Storz, J. F., Opazo, J. C., and Hoffmann, F. G. (2013). Gene duplication, genome duplication, and the functional diversification of vertebrate globins. *Mol. Phylogenet. Evol.* 66, 469–478.

Strimmer, K., and Rambaut, A. (2002). Inferring confidence sets of possibly misspecified gene trees. *Proc. Biol. Sci.* 269, 137–142.

Tatusova, T. A., and Madden, T. L. (1999). BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. *FEMS Microbiol. Lett.* 174, 247–250.

Vinogradov, S. N., Hoogewijs, D., Bailly, X., Arredondo-Peter, R., Guertin, M., Gough, J., et al. (2005). Three globin lineages belonging to two structural classes in genomes from the three kingdoms of life. *Proc. Natl. Acad. Sci. U. S. A.* 102, 11385–11389.

Xu, P., Xu, J., Liu, G., Chen, L., Zhou, Z., Peng, W., et al. (2019). The allotetraploid origin and asymmetrical genome evolution of the common carp Cyprinus carpio. *Nat. Commun.* 10, 4625.

Xu, P., Zhang, X., Wang, X., Li, J., Liu, G., Kuang, Y., et al. (2014). Genome sequence and genetic diversity of the common carp, Cyprinus carpio. *Nat. Genet.* 46, 1212–1219.

Yates, A. D., Achuthan, P., Akanni, W., Allen, J., Allen, J., Alvarez-Jarreta, J., et al. (2020). Ensembl 2020. *Nucleic Acids Res.* 48, D682–D688.
