# Whole-Genome Duplications and the Diversification of the Globin-X Genes of Vertebrates

Federico G. Hoffmann [1,2,*], Jay F. Storz[3], Shigehiro Kuraku [4,5,6], Michael W. Vandewege [7], and Juan C. Opazo[8,9,10]

[1]Department of Biochemistry, Molecular Biology, Entomology, and Plant Pathology, Mississippi State University, Starkville, MS, USA
[2]Institute for Genomics, Biocomputing and Biotechnology, Mississippi State University, Starkville, MS, USA
[3]School of Biological Sciences, University of Nebraska, Lincoln, NE, USA
[4]Molecular Life History Laboratory, Department of Genomics and Evolutionary Biology, National Institute of Genetics, Mishima, Japan
[5]Department of Genetics, Sokendai (Graduate University for Advanced Studies), Mishima, Japan
[6]Laboratory for Phyloinformatics, RIKEN Center for Biosystems Dynamics Research, Kobe, Japan
[7]Department of Biology, Eastern New Mexico University, Portales, NM, USA
[8]Integrative Biology Group, Universidad Austral de Chile, Valdivia, Chile
[9]Instituto de Ciencias Ambientales y Evolutivas, Facultad de Ciencias, Universidad Austral de Chile, Valdivia, Chile
[10]Millennium Nucleus of Ion Channel-Associated Diseases (MiNICAD), Valdivia, Chile

*Corresponding author: E-mails: federico.g.hoffmann@gmail.com, fgh19@msstate.edu

**Accepted:** 23 August 2021

## Abstract

Globin-X (GbX) is an enigmatic member of the vertebrate globin gene family 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, evidence suggests that the monomeric, membrane-bound GbX may play a role in cellular signaling or protection against the oxidation of membrane lipids. Recently released genomes from key vertebrates provide an excellent opportunity to address questions about the early stages of the evolution of GbX in vertebrates. We integrate bioinformatics, synteny, and phylogenetic analyses to characterize the diversity of *GbX* genes in nonteleost ray-finned fishes, resolve relationships between the *GbX* genes of cartilaginous fish and bony vertebrates, and demonstrate that the *GbX* genes of cyclostomes and gnathostomes derive from independent duplications. 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 relative to other globin genes and provide an evolutionary framework for interpreting results of experiments that examine functional properties of GbX and patterns of tissue-specific expression. By identifying *GbX* paralogs that are products of different WGDs, 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.

**Key words:** gene family evolution, comparative genomics, gene expansion, synteny, cyclostomes.

## Introduction

Globins are small, oxygen-binding hemoproteins found in all domains of life (Vinogradov et al. 2005; Storz 2019). The globin superfamily of vertebrates provides an excellent example of how local gene duplications, whole-genome duplications, and both structural and regulatory changes in individual genes can promote 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)

## Significance

Globins are small, oxygen-binding proteins found in all domains of life. The globin gene superfamily provides a textbook example of how the interplay between local gene duplications, whole-genome duplications (WGDs), and both structural and regulatory changes can promote the evolution of novel protein functions. Globin-X (GbX) is an enigmatic member of the vertebrate globin gene family with a broad phyletic distribution. Analyzing the genomes of early-diverging vertebrates, we find that the GbX family is much more widely represented in vertebrates than expected, and many gene copies in extant taxa are products of WGDs. Our results provide a robust evolutionary framework to interpret functional data and inform the design of further experiments to test whether genes that originate via WGDs have evolved novel functional properties relative to singleton or tandemly duplicated genes in the same species.

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, Opazo, Hoogewijs, et al. 2012). 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 before the split between protostomes and deuterostomes (Burmester et al. 2000, 2002; Roesner et al. 2005; Dröge and Makałowski 2011; Hoffmann, Opazo, Hoogewijs, et al. 2012; Blank and Burmester 2012), and despite 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 all vertebrates examined possess *hemoglobin* and *myoglobin* genes 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 distributions of *globin-E*, *globin-Y*, and *GbX* are 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, Opazo, Hoogewijs, et al. 2012; 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 cellular signaling, protection against the oxidation of membrane lipids and even appears to function as a nitrite reductase in red blood cells (Corti et al. 2016; Koch and Burmester 2016).

Whole-genome duplications (WGDs) have played a prominent role in the expansion and functional diversification of vertebrate-specific globins (Storz et al. 2011; Hoffmann, Opazo and Storz 2012; Storz et al. 2013) 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 squamate reptiles, 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, Opazo, Hoogewijs, et al. 2012). 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 WGDs were responsible for the expanded repertoire of *GbX* 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 relationships of *GbX* genes of cartilaginous fish and those 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* during the course of vertebrate evolution. 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.

## Results

### Data Description and Nomenclature

We obtained 139 putative vertebrate *GbX* sequences from the Ensembl database, release 101 (http://aug2020.archive.ensembl.org/index.html, 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, Supplementary Material online). 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 ENSLOCG00000014709, which is flanked by *PLEHKG2* and *SUPT5*, were labeled as *GbX1* genes, and orthologs of the spotted gar *GbX2* gene ENSLOCG00000012798, 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 (ENSELUG00000016373), 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 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*, *TMEM160*, and *PYGL*) as *GbX-C1*.

### Phylogenetic Analyses

We first estimated a maximum likelihood phylogenetic tree for the initial alignment of 190 sequences to ensure they were true *GbX* genes (available as Vert_GbX.190.fasta, Supplementary Material online). 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, Opazo, Hoogewijs, et al. 2012; Opazo et al. 2015), (fig. 1, supplementary fig. 1, Supplementary Material online). 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, Supplementary Material online), 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 copies that correspond to the same gene 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 (except for 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 Material online). 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

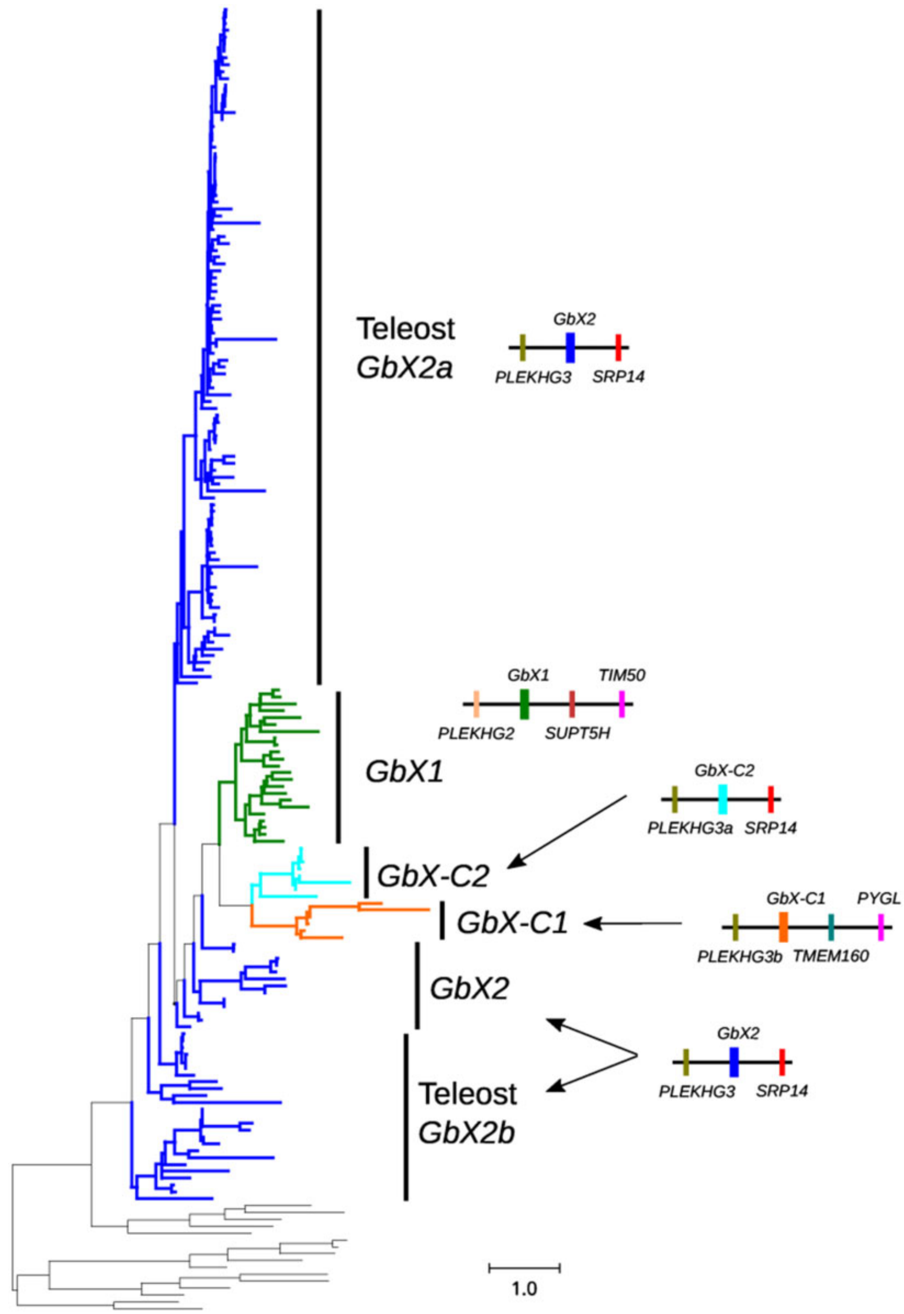


**Fig. 1.**—Maximum likelihood phylogram describing evolutionary relationships among the vertebrate globin-X candidates identified in our study. The tree was rooted with the full set of Acorn worm globins. The tree with the terminal labels is available as supplementary figure 1, Supplementary Material online, and the corresponding alignment is available as Vert_GbX.190.fasta, Supplementary Material online.

to the clade that included the *GbX1*, *GbX-C1*, and *GbX-C2* genes, (supplementary fig. 2, Supplementary Material online). 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 descendent lineages, and would also imply that the duplication giving rise to the teleost *GbX2a* and *2b* ohnologs occurred in the vertebrate ancestor. However, a tree that minimizes these independent deletions, where the *GbX1*

**Table 1**
Results of topology tests

| Tree | logL | ΔL | bp-RELL | p-KH | p-SH | c-ELW | p-AU |
|---|---|---|---|---|---|---|---|
| Unconstrained | −20,600.3 | 0 | 0.428 | 0.617 | 1 | 0.428 | 0.629 |
| *GbX2* monophyletic | −20,604.5 | 4.2 | 0.199 | 0.383 | 0.789 | 0.199 | 0.454 |
| *GbX2*, *GbX2a*, and *GbX2b* monophyletic | −20,604.9 | 4.6 | 0.359 | 0.412 | 0.6 | 0.357 | 0.441 |
| *GbX2* sister to *GbX-C2* | −20,640.4 | 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 2002).

and *GbX2* genes of gnathostomes were constrained to be monophyletic, and the *GbX2a* and *GbX2b* ohnologs of teleosts were constrained to be monophyletic within *GbX2* 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, and because it agrees well with assessments of conserved synteny among gnathostomes, we selected it as the most plausible phylogenetic hypothesis and we used it as the basis of our evolutionary inferences. In this constrained tree (fig. 2, supplementary fig. 3, Supplementary Material online), 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 possesses 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 lamprey *GbX-C2* sequences (also supported by the available synteny data) indicates that the duplication that gave rise to *GbX-C1* and *GbX-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 is an assembly artifact.

In the case of gnathostome *GbX* genes, the *GbX1* sequences from 1) cartilaginous fish, 2) ray-finned fishes, 3) caecilians, 4) squamates, and 5) 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 fig. 3, Supplementary Material online). 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 despite 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. In turn, the clade of sterlet and reedfish *GbX1* genes was placed sister to spotted gar *GbX1*. 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 *GbX1*, but again, the corresponding branches were very short, so we did not attach importance to this apparent discrepancy. Relationships among the *GbX1* sequences from amniotes matched the expected organismal relationships.

In the constrained tree (fig. 2, supplementary fig. 3, Supplementary Material online), 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 figs. 1–3, Supplementary Material online). The duplicate GbX2a paralogs of salmonids match expected organismal relationships (supplementary figs. 1–3, Supplementary Material online), consistent with the hypothesis that they trace their origins to the salmonid-specific WGD (Gallagher and Macqueen 2017). The cyprinin-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 cyprinin paralogs by genus, except for the tandem duplicate of the *GbX2a* paralog of the carp, ENSCCRG00000022746, which is grouped with the *GbX2a* paralogs of golden-line fishes (supplementary figs. 1–3, Supplementary Material online). 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*) have each experienced an independent

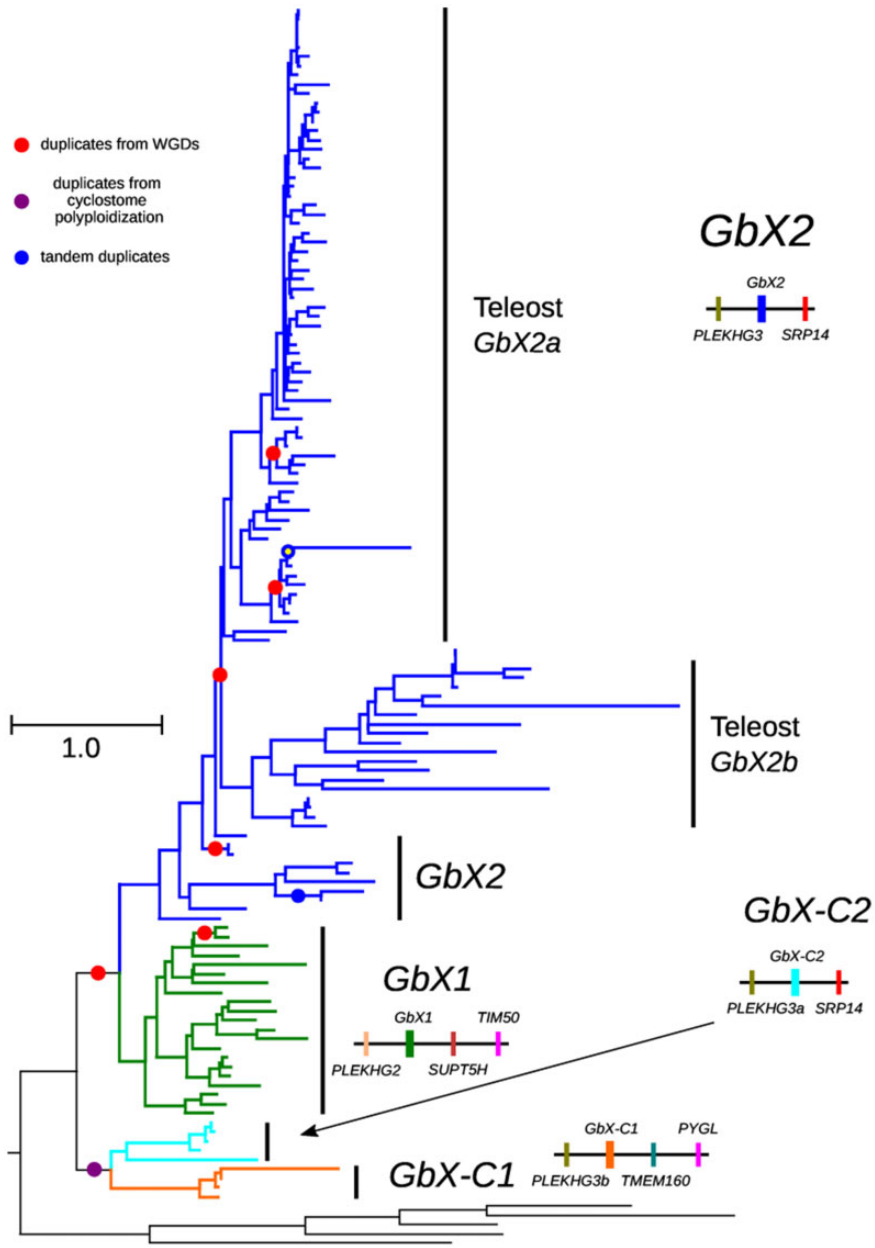


FIG. 2.—Maximum likelihood phylogram describing evolutionary relationships among the curated set of vertebrate globin-X candidates in our study, where the *GbX1* and *GbX2* genes of gnathostomes were constrained to be monophyletic, and the *GbX2a* and *GbX2b* ohnologs of teleosts were constrained to be monophyletic within *GbX2* and sister to each other. This tree was not statistically different from an unconstrained tree, which is available as supplementary figure 2, Supplementary Material online, 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 the corresponding alignment is available as Vert_GbX.134.fasta, Supplementary Material online.

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 (Hoffmann et al. 2010; 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 coduplicated 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 tested this expectation 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 *PLEKHG* phylogeny recapitulates the topology of the GbX tree (supplementary fig. 4, Supplementary Material online), placing the two *PLEKHG3* paralogs of cyclostomes in a monophyletic group (mirroring the relationship of 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 (mirroring the relationship of their neighboring *GbX1* and *GbX2* genes in the GbX tree). Thus, synteny and phylogenetic analyses both indicate that the *GbX-C1* and *GbX-C2* genes of cyclostomes and the *GbX1* and *GbX2* genes of gnathostomes are products of independent duplication events.

### Expression of the Different GbX Paralogs

The available evidence indicates that the *GbX* paralogs of elephant fish, spotted gar, and salmon have different tissue-specific expression profiles (Opazo et al. 2015; Gallagher and Macqueen 2017). In the elephant fish, RNA-seq data indicate 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 the brain, spleen, and testis, whereas GbX2 (XP_007891388), which is labeled a GbX1 in Opazo et al. (2015), is expressed in the 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 the 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 (ENSSSAG00000047904) 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 tissue-specific expression are variable across lineages.

## 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 longstanding 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 paradigm change 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 Makałowski 2011; Blank and Burmester 2012; Hoffmann, Opazo, Hoogewijs, et al. 2012; Opazo et al. 2015), which is estimated to have occurred ~ 730 million years ago (Kumar et al. 2017).

### Evolution of GbX in Early Vertebrates

Our reconstructions shed light on the early stages of the 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. The results suggest that the two *GbX* paralogs of gnathostomes, *GbX1* and *GbX2*, trace back to the last common ancestor of cartilaginous fish and bony vertebrates, and synteny analyses suggest that these two paralogs are 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

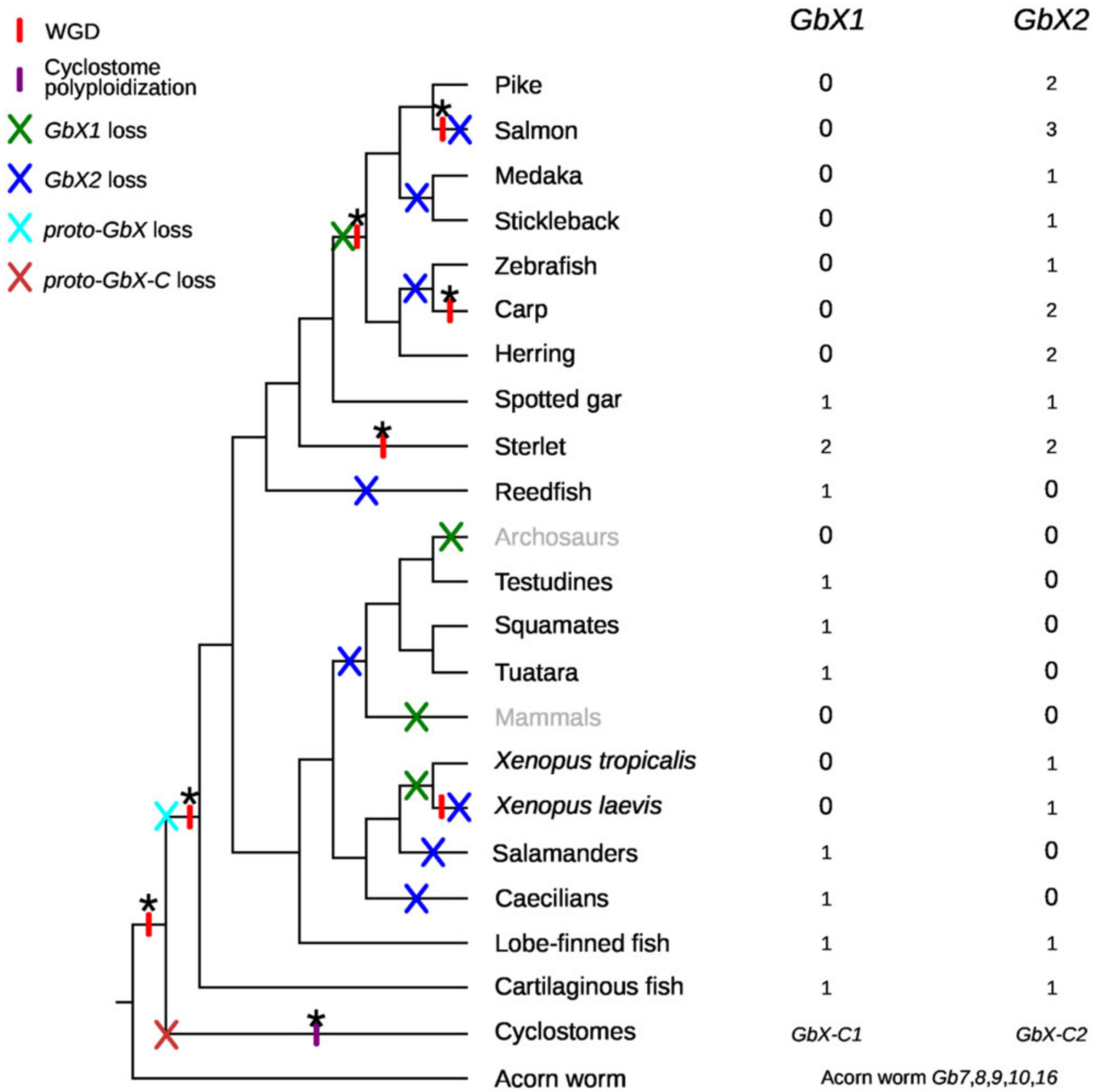


**Fig. 3.**—Graphical summary of the role of WGDs in the expansion of the vertebrate GbX repertoire. Organismal relationships on the right, and the number of *GbX* paralogs per lineage on the left. WGDs, polyploidizations, and gene losses are mapped to their corresponding branch. We placed the 1R and 2R WGDs following Simakov et al. (2020) and Nakatani et al. (2021). Symbols on a branch are arranged according to their relative order. Asterisks on top of the vertical bars denote WGDs that gave rise to *GbX* paralogs present in extant species. The polyploidization event on the cyclostome branch could be an additional WGD (Mehta et al. 2013) or an hexaploidization due to hybridization between diploid and tetraploid lineages (Nakatani et al. 2021). Note that a full tree of all species examined would include 2 additional tandem duplications and multiple additional gene losses.

segmental duplication or even a WGD. In this regard, our results are consistent with previous studies that indicate that lamprey genomes underwent an additional polyploidization by either a WGD (Mehta et al. 2013), hexaploidization via hybridization between tetraploid and diploid lineages (Nakatani et al. 2021), or extensive segmental duplications (Smith and Keinath 2015).

Our results provide strong evidence that the gnathostome *GbX* paralogs derive from one of the two vertebrate-specific WGDs (fig. 3), either 1R or 2R, confirming inferences from Opazo et al. (2015). In addition, the *GbX-C1* and *GbX-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 two 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). The most recent studies place 2R in the last common ancestor of gnathostomes (Simakov et al. 2020; Nakatani et al. 2021), and suggest that cyclostomes underwent an additional and independent polyploidization early in their evolution (Mehta et al. 2013; Nakatani et al. 2021), whereas other authors place 2R in the common ancestor of cyclostomes and gnathostomes (Sacerdot et al. 2018). Under the first scenario, 1R would have given rise to the proto GbX gene of gnathostomes and the proto GbX-C gene of cyclostomes, followed by reciprocal losses of proto-GbX-C in the common ancestor of gnathostomes and of proto-GbX in the ancestor of cyclostomes (fig. 3). The *GbX1* and *GbX2* of gnathostomes would derive from 2R, and the shared presence of *PLEKHG3* paralogs next to *GbX-C1* and *GbX-C2* would suggest these derive from a polyploidization event in the ancestor of cyclostomes Our results would require additional independent gene losses to fit the second scenario.

## 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. 2021), 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 *GbX*1 and *GbX2*, and teleosts have only retained copies of *GbX2* (fig. 1, supplementary figs. 1–3, Supplementary Material online). 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 is an artifact. 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 1, Supplementary Material online). The *GbX2a* and *GbX2b* paralogs of salmonid fishes also exhibit 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 figs. 1–3, Supplementary Material online). Similarly, the duplicate *GbX2* paralogs of the subfamily Cyprininae had contrasting fates. The *GbX2a* paralog duplicated in the cyprinin-specific WGD and has been retained in most cases, whereas the *GbX2b* paralog was apparently lost earlier, in the common ancestor of cyprinins and zebrafish.

In addition to different rates of retention between the *GbX2a* and *GbX2b* paralogs of teleost fish, the two genes also have different histories of transposition and substitution. 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 of *PLEKHG3* as well. The *GbX2a* and *GbX2b* genes also exhibit different 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 *GbX2b* 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, Opazo and Storz 2012; Opazo et al. 2013, Opazo et al. 2015; Storz et al. 2013). As with vertebrate-specific globins, the repertoire of *GbX* in extant taxa 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 seven pairs of duplicates that *GbX1* was lost at least four times independently, and that *GbX2* was lost at least eight times independently (fig. 3). The *GbX2b* gene has also been lost multiple times independently in teleost fishes. Starting with the oldest, the first pair of WGD-derived duplicates corresponds to the proto-GbX and proto-GbX-C genes of gnathostomes and cyclostomes, which derive from the 1R WGD, and were reciprocally lost in cyclostomes and gnathostomes (fig. 3). Then, in the ancestor of gnathostomes 2R gave rise to the second pair of ohnologs, *GbX1* and *GbX2*. The third pair corresponds to the *GbX2a* and *GbX2b* ohnologs of teleost fish, which derive from the teleost WGD, and the fourth 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 fifth pair of ohnologs corresponds to the duplicate copies of *GbX2a* in cyprinins, which derive from the cyprinin-specific WGD, and finally, the sixth and seventh 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. It is also possible that the *GbX-C1* and *GbX-C2* duplicates of lampreys and hagfish derive from a cyclostome-specific WGD.

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 the 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.

## Materials and Methods

### 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. 2021). 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, Opazo, Hoogewijs, et al. 2012; 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 nonteleost ray-finned fish (Bi et al. 2021; Du et al. 2020), 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; Chen et al. 2019; Xu 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, Opazo, Hoogewijs, et al. 2012). We verified the identity of candidate *GbX* genes by reciprocal BLAST, comparing putative *GbX* sequences against the nonredundant protein sequence database (nr) of deuterostomes.

### Sequence Alignment and Phylogenetic Analyses

We aligned amino acid sequences using the L-INS-i strategy from MAFFT v 7.471 (Katoh et al. 2019; Katoh 2005) 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 2002) as implemented in IQ-Tree v.2.0.6 (Minh et al. 2020). All trees, alignments, and search logs are available in the Supplementary Material online.

### 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, Opazo, Hoogewijs, et al. 2012; Prothmann et al. 2020). We removed redundant records and retained a single representative species per genus, except for the carp, where we kept two separate assemblies that include two separate duplications. Finally, we also removed truncated genes from the data set. In the cases of the 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 1, Supplementary Material online 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 1, Supplementary Material online.

### 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).

## Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

## Acknowledgments

F.G.H. thanks Amanda Coward Black for editorial assistance. J.C.O. wants to acknowledge the members of the Integrative Biology Group, Universidad Austral de Chile for their constant support, scientific enthusiasm, and creative feedback. Support for this research was provided by the National Science Foundation (OIA-1736026 to F.G.H.; OIA-1736249 to J.F.S.); 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 to F.G.H.; the Fondo Nacional de Desarrollo Científico y Tecnológico from Chile (FONDECYT 1210471 to J.C.O.); the Millennium Nucleus of Ion Channels Associated Diseases (MiNICAD), Iniciativa Científica Milenio, Ministry of Economy, Development and Tourism to J.C.O.; the National Institute of Health (HL087216 to J.F.S.); and the Japan Society for the Promotion of Science Grants-in-Aid for Scientific Research (KAKENHI 20H03269 to S.K.).

## Data Availability

All data used in this article is available as part of its online supplementary material.

## Literature cited

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol. 215(3):403–410.

Anisimova M, Gil M, Dufayard J-F, Dessimoz C, Gascuel O. 2011. Survey of branch support methods demonstrates accuracy, power, and robustness of fast likelihood-based approximation schemes. Syst Biol. 60(5):685–699.

Ascenzi P, Gustincich S, Marino M. 2014. Mammalian nerve globins in search of functions. IUBMB Life. 66(4):268–276.

Berthelot C, et al. 2014. The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates. Nat Commun. 5:3657.

Bi X, et al. 2021. Tracing the genetic footprints of vertebrate landing in non-teleost ray-finned fishes. Cell 184(5):1377–1391.

Blank M, et al. 2011. A membrane-bound vertebrate globin. PLoS One. 6(9):e25292.

Blank M, 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(11):3553–3561.

Burmester T, Ebner B, Weich B, Hankeln T. 2002. Cytoglobin: a novel globin type ubiquitously expressed in vertebrate tissues. Mol Biol Evol. 19(4):416–421.

Burmester T, Hankeln T. 2014. Function and evolution of vertebrate globins. Acta Physiol (Oxf). 211(3):501–514.

Burmester T, Weich B, Reinhardt S, Hankeln T. 2000. A vertebrate globin expressed in the brain. Nature 407(6803):520–523.

Campanini EB, et al. 2015. Early evolution of vertebrate *Mybs*: an integrative perspective combining synteny, phylogenetic, and gene expression analyses. Genome Biol Evol. 7(11):3009–3021.

Chen Z, et al.; NISC Comparative Sequencing Program. 2019. De novo assembly of the goldfish (*Carassius auratus*) genome and the evolution of genes after whole-genome duplication. Sci Adv. 5(6):eaav0547.

Corti P, 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(30):8538–8543.

Dehal P, Boore JL. 2005. Two rounds of whole genome duplication in the ancestral vertebrate. PLoS Biol. 3(10):e314.

Dröge J, Makałowski W. 2011. Phylogenetic analysis reveals wide distribution of globin X. Biol Direct. 6:54.

Du K, et al. 2020. The sterlet sturgeon genome sequence and the mechanisms of segmental rediploidization. Nat Ecol Evol. 4(6):841–852.

Fago A, Hundahl C, Malte H, Weber RE. 2004. Functional properties of neuroglobin and cytoglobin. Insights into the ancestral physiological roles of globins. IUBMB Life. 56(11-12):689–696.

Gallagher MD, Macqueen DJ. 2017. Evolution and expression of tissue globins in ray-finned fishes. Genome Biol Evol. 9(1):32–47.

Hess JE, et al. 2020. Genomic islands of divergence infer a phenotypic landscape in Pacific lamprey. Mol Ecol. 29(20):3841–3856.

Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. 2018. UFBoot2: improving the ultrafast bootstrap approximation. Mol Biol Evol. 35(2):518–522.

Hoffmann FG, Opazo JC, Hoogewijs D, et al. 2012. Evolution of the globin gene family in deuterostomes: lineage-specific patterns of diversification and attrition. Mol Biol Evol. 29(7):1735–1745.

Hoffmann FG, Opazo JC, Storz JF. 2010. Gene cooption and convergent evolution of oxygen transport hemoglobins in jawed and jawless vertebrates. Proc Natl Acad Sci U S A. 107(32):14274–14279.

Hoffmann FG, Opazo JC, Storz JF. 2011. Differential loss and retention of cytoglobin, myoglobin, and globin-E during the radiation of vertebrates. Genome Biol Evol. 3:588–600.

Hoffmann FG, Opazo JC, Storz JF. 2012. Whole-genome duplications spurred the functional diversification of the globin gene superfamily in vertebrates. Mol Biol Evol. 29(1):303–312.

Hoogewijs D, et al. 2012. Androglobin: a chimeric globin in metazoans that is preferentially expressed in Mammalian testes. Mol Biol Evol. 29(4):1105–1114.

Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat Methods. 14(6):587–589.

Katoh K. 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res. 33(2):511–518.

Katoh K, Rozewicki J, Yamada KD. 2019. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. Brief Bioinform. 20(4):1160–1166.

Keppner A, et al. 2020. Lessons from the post-genomic era: globin diversity beyond oxygen binding and transport. Redox Biol. 37:101687.

Kishino H, 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 Mol Evol. 29(2):170–179.

Kishino H, Miyata T, Hasegawa M. 1990. Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. J Mol Evol. 31(2):151–160.

Koch J, Burmester T. 2016. Membrane-bound globin X protects the cell from reactive oxygen species. Biochem Biophys Res Commun. 469(2):275–280.

Kumar S, Stecher G, Suleski M, Hedges SB. 2017. TimeTree: a resource for timelines, timetrees, and divergence times. Mol Biol Evol. 34(7):1812–1819.

Kuraku S. 2013. Impact of asymmetric gene repertoire between cyclostomes and gnathostomes. Semin Cell Dev Biol. 24(2):119–127.

Kuraku S. 2010. Palaeophylogenomics of the vertebrate ancestor–impact of hidden paralogy on hagfish and lamprey gene phylogeny. Integr Comp Biol. 50(1):124–129.

Kuraku S, 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, Kuratani S. 2009. Timing of genome duplications relative to the origin of the vertebrates: did cyclostomes diverge before or after? Mol Biol Evol. 26(1):47–59.

Lien S, et al. 2016. The Atlantic salmon genome provides insights into rediploidization. Nature 533(7602):200–205.

McLysaght A, Hokamp K, Wolfe KH. 2002. Extensive genomic duplication during early chordate evolution. Nat Genet. 31(2):200–204.

Mehta TK, 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.

Meyer A, 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(6):699–704.

Minh BQ, et al. 2020. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Mol Biol Evol. 37(5):1530–1534.

Nakatani Y, et al. 2021. Reconstruction of proto-vertebrate, proto-cyclostome and proto-gnathostome genomes provides new insights into early vertebrate evolution. Nat Commun. 12(1):4489.

Nguyen NTT, Vincens P, Roest Crollius H, Louis A. 2018. Genomicus 2018: karyotype evolutionary trees and on-the-fly synteny computing. Nucleic Acids Res. 46(D1):D816–D822.

Opazo JC, 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(7):1684–1694.

Opazo JC, Butts GT, Nery MF, Storz JF, Hoffmann FG. 2013. Whole-genome duplication and the functional diversification of teleost fish hemoglobins. Mol Biol Evol. 30(1):140–153.

Pál C, Papp B, Lercher MJ. 2006. An integrated view of protein evolution. Nat Rev Genet. 7(5):337–348.

Prothmann A, et al. 2020. The globin gene family in arthropods: evolution and functional diversity. Front Genet. 11:858.

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.

Queiroz JPF, Lima NCB, Rocha BAM. 2021. The rise and fall of globins in the amphibia. Comp Biochem Physiol Part D Genomics Proteomics. 37:100759.

Rhie A, et al. 2021. Towards complete and error-free genome assemblies of all vertebrate species. Nature 592(7856):737–746.

Roesner A, Fuchs C, Hankeln T, 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(1):12–20.

Sacerdot C, Louis A, Bon C, Berthelot C, Roest Crollius H. 2018. Chromosome evolution at the origin of the ancestral vertebrate genome. Genome Biol. 19(1):166.

Session AM, et al. 2016. Genome evolution in the allotetraploid frog *Xenopus laevis*. Nature 538(7625):336–343.

Sharma S, et al. 2018. The NCBI BioCollections database. Database 2018(2018):bay006.

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

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

Simakov O, et al. 2020. Deeply conserved synteny resolves early events in vertebrate evolution. Nat Ecol Evol. 4(6):820–830.

Smith JJ, Keinath MC. 2015. The sea lamprey meiotic map improves resolution of ancient vertebrate genome duplications. Genome Res. 25(8):1081–1090.

Song S, et al. 2020. Globins in the marine annelid *Platynereis dumerilii* shed new light on hemoglobin evolution in Bilaterians. BMC Evol Biol. 20(1):165.

Storz JF. 2019. Hemoglobin: insights into protein structure, function, and evolution. New York (NY): Oxford University Press.
Storz JF, Opazo JC, Hoffmann FG. 2013. Gene duplication, genome duplication, and the functional diversification of vertebrate globins. Mol Phylogenet Evol. 66(2):469–478.
Storz JF, Opazo JC, Hoffmann FG. 2011. Phylogenetic diversification of the globin gene superfamily in chordates. IUBMB Life 63(5):313–322.
Strimmer K, Rambaut A. 2002. Inferring confidence sets of possibly misspecified gene trees. Proc Biol Sci. 269(1487):137–142.
Tatusova TA, Madden TL. 1999. BLAST 2 sequences, a new tool for comparing protein and nucleotide sequences. FEMS Microbiol Lett. 174(2):247–250.
Vinogradov SN, 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(32):11385–11389.
Xu P, et al. 2014. Genome sequence and genetic diversity of the common carp, *Cyprinus carpio*. Nat Genet. 46(11):1212–1219.
Xu P, et al. 2019. The allotetraploid origin and asymmetrical genome evolution of the common carp *Cyprinus carpio*. Nat Commun. 10(1):4625.
Yates AD, et al. 2020. Ensembl 2020. Nucleic Acids Res. 48(D1):D682–D688.

**Associate editor:** Adam Eyre-Walker