# Evolutionary Analysis of the Zinc Finger and Homeoboxes Family of Proteins Identifies Multiple Conserved Domains and a Common Early Chordate Ancestor

Alexandra N. Nail[1], Jeramiah J. Smith[2,3], Martha L. Peterson[1,3], and Brett T. Spear [1,3,*]

[1]Department of Microbiology, Immunology and Molecular Genetics, University of Kentucky

[2]Department of Biology, University of Kentucky

[3]Markey Cancer Center, University of Kentucky

*Corresponding author: E-mail: bspear@uky.edu.

Accepted: February 27, 2020

## Abstract

The Zinc Fingers and Homeoboxes (Zhx) proteins, Zhx1, Zhx2, and Zhx3, comprise a small family of proteins containing two amino-terminal $C_2$–$H_2$ zinc fingers and four or five carboxy-terminal homeodomains. These multiple homeodomains make Zhx proteins unusual because the majority of homeodomain-containing proteins contain a single homeodomain. Studies in cultured cells and mice suggest that Zhx proteins can function as positive or negative transcriptional regulators. Zhx2 regulates numerous hepatic genes, and all three Zhx proteins have been implicated in different cancers. Because Zhx proteins contain multiple predicted homeodomains, are associated with interesting physiological traits, and seem to be only present in the vertebrate lineage, we investigated the evolutionary history of this small family by comparing Zhx homologs from a wide range of chordates. This analysis indicates that the zinc finger motifs and homeodomains are highly similar among all Zhx proteins and also identifies additional Zhx-specific conserved regions, including a 13 amino acid amino-terminal motif that is nearly identical among all gnathostome Zhx proteins. We found single Zhx proteins in the sea lamprey (*Petromyzon marinus*) and in the nonvertebrate chordates sea squirt (*Ciona intestinalis*) and lancelet (*Branchiostoma floridae*); these Zhx proteins are most similar to gnathostome Zhx3. Based on our analyses, we propose that a duplication of the primordial *Zhx* gene gave rise to *Zhx3* and the precursor to *Zhx1* and *Zhx2*. A subsequent tandem duplication of this precursor generated *Zhx1* and *Zhx2* found in gnathostomes.

**Key words:** gnathostome, homeodomain, zinc finger, chordate evolution, transcription factor.

## Introduction

The Zinc Fingers and Homeoboxes (Zhx) protein family, consisting of Zhx1, Zhx2, and Zhx3, along with the more divergent Homez, comprise one of five families within the Zinc Finger class of homeodomain-containing proteins (Spear et al. 2006). Along with several other members of the Zinc Finger class, Zhx proteins are quite unusual among the hundreds of homeodomain proteins in that they contain multiple homeodomains rather than a single homeodomain, characteristic of most proteins in this class (Burglin and Affolter 2016). Specifically, Zhx proteins contain two amino-terminal $C_2$–$H_2$ zinc finger motifs and four or five carboxy-terminal homeodomains. The presence of homeodomains, an evolutionarily conserved 60 amino acid DNA-binding motif composed of three $\alpha$-helices (Gehring et al. 1994), is consistent with Zhx proteins functioning as transcriptional regulators. Homeodomain proteins are widespread, being found in fungi, plants, invertebrates, and vertebrates, and have essential roles in development and cellular differentiation and function (Luo et al. 2019). Mutations in homeodomain proteins can lead to profound developmental anomalies and are associated with a variety of diseases in humans, including cancer (Luo et al. 2019).

Mouse and human Zhx1 were initially identified by screening a murine endothelial-adipose cell line cDNA library and in a yeast 2-hybrid screen for proteins interacting with the transcription factor NF-YA, respectively (Barthelemy et al. 1996; Yamada et al. 1999). Subsequent yeast 2-hybrid experiments with ZHX1 as a bait isolated ZHX2 and ZHX3 and demonstrated that Zhx proteins form homodimers and heterodimers

with each other and with NF-YA in vitro (Yamada et al. 2002, 2003; Kawata et al. 2003a, 2003b).

Initial insight into the physiological role of Zhx proteins came from BALB/cJ mice, which have a natural *Zhx2* mutation that dramatically reduces Zhx2 mRNA levels (Olsson et al. 1977; Belayew and Tilghman 1982; Perincheri et al. 2005). A number of genes that are normally expressed in the fetal liver and repressed at birth, including *alpha-fetoprotein* (*AFP*), *H19*, and *Glypican 3* continue to be expressed in BALB/cJ adult liver (Belayew and Tilghman 1982; Pachnis et al. 1984; Morford et al. 2007). This incomplete repression of Zhx2 target genes in the BALB/cJ liver support in vitro studies indicating that Zhx2 functions as a transcriptional repressor (Yamada et al. 2002; Kawata et al. 2003a, 2003b; Yue et al. 2012). However, mouse *Major Urinary Protein* (*Mup*) genes are positively regulated by Zhx2 in the liver, indicating that Zhx2 may act in a context-dependent manner to positively or negatively control target gene expression (Jiang et al. 2017). The ability of Zhx proteins to interact with other transcription factors, including NF-YA and DNMT3B, suggests that Zhx proteins may also function as coregulators to control target gene expression (Yamada et al. 1999; Kim et al. 2007). Several studies have implicated a role for Zhx2 in hepatocellular carcinoma (Lv et al. 2006; Hu et al. 2007; Yue et al. 2012). Less is known about the function of Zhx1 and Zhx3, although several studies suggest a role for these proteins in kidney disease as well as certain cancers (Liu et al. 2006; Peterson et al. 2011).

Structural analysis of Zhx proteins is limited. NMR studies indicate that the tandem zinc finger motif of human Zhx1 has a unique structure that includes a conserved region on the carboxy side of the second zinc finger; this study suggested that the zinc fingers interact with proteins rather than DNA or RNA (Wienk et al. 2009). An effort to crystallize single and combined Zhx homeodomains succeeded with only human Zhx1 Homeodomain 4 (HD4) and Zhx2 HD2. X-ray crystallographic analysis indicated that these two homeodomains have somewhat atypical structures in comparison to other homeodomains (Bird et al. 2010).

Because the Zhx proteins contain multiple predicted homeodomains and are associated with interesting physiological traits, we investigated the evolutionary history of this small protein family by comparing Zhx proteins from a wide range of chordates. These studies identified Zhx proteins only in chordates, suggesting a relatively recent evolutionary origin. All gnathostomes analyzed, from elephant shark to humans, contain Zhx1, Zhx2, and Zhx3. Sea lamprey (*Petromyzon marinus*), sea squirt (*Ciona intestinalis)*, and lancelet (*Branchiostoma floridae*) genomes each appear to encode a single Zhx protein that is most similar to gnathostome Zhx3. Based on our analyses, we propose that a duplication of the primordial *Zhx* gene gave rise to *Zhx3* and the ancestral *Zhx1/2* locus, and a second tandem duplication gave rise to the two linked *Zhx1* and *Zhx2* loci.

## Materials and Methods

### Identification and Characterization of Zinc Finger and Homeobox Proteins from Chordate Species

BLAST analysis identified 3 Zhx proteins from 22 different gnathostome species, ranging from spotted gar to humans, which were downloaded from the NCBI database for further analysis (supplementary table 1, Supplementary Material online). The rabbit Zhx2 sequence contained a gap that spanned homeodomain 4, which resulted in the incorrect prediction of the carboxy terminus of the protein; aligning the rabbit Zhx2 genomic sequence to pig predicted the correct end of the rabbit Zhx2 protein. BLAST analysis of the newly assembled *Petromyzon marinus* (sea lamprey) genome identified three predicted genes designated *Zhx1*, *Zhx2*, and *Zhx3*. Further analysis revealed that the gene identified as *Zhx3* encodes an intact Zhx protein that is most similar to gnathostome Zhx3 (see Results). This lamprey Zhx protein sequence was confirmed by sequencing reverse transcription-polymerase chain reaction and 3′/5′ RACE amplicons of lamprey liver cDNA (Nail AN, Smith JJ, Peterson ML, Spear BT, in preparation). The genes designated as *Zhx1* and *Zhx2* appear to be a single mutated gene, indicating that it is a likely pseudogene that would not encode a full-length Zhx protein (Nail AN, Smith JJ, Peterson ML, Spear BT, in preparation). BLAST analysis of lancelet and sea squirt using full-length mouse Zhx proteins and the Zhx HD1 identified a single *Zhx* gene from each species. Identification of Zhx orthologs and syntenic genes was performed using the University of California–Santa Cruz and Ensembl databases.

Zhx proteins were manually annotated for zinc finger and homeodomain regions using Geneious software (Version 11.1.5). The $C_2H_2$ zinc finger domains were anchored with cysteines at amino acid positions 1 and 4 and histidines at positions 17 and 22 for each zinc finger. Homeodomains were anchored by positions 48 and 49 (W48/F48/Y48 and F49/Y49/W49 within each homeodomain) that are the most highly conserved residues among previously studied homeodomain proteins (Burglin and Affolter 2016). Conserved residues L16/F16/M16/A16 and R53/L53 within each homeodomain were also used as minimal criteria to annotate a region as a predicted homeodomain. Each homeodomain was annotated as 60 amino acids in length.

### Multiple Sequence Alignment and Phylogenetic Tree Building

Multiple sequence alignments (MSAs) were completed in Geneious 11.1.5 software using MUSCLE (Edgar 2004; Kearse et al. 2012). MSAs of full-length Zhx proteins, Zhx homeodomain regions (60 aa), and zinc finger regions (ZFRs) (22 aa for each zinc finger and the 10 aa region between zinc fingers) were used to build phylogenetic trees across species.

Distance tree building was completed using Geneious Tree Builder with the Jukes–Cantor distance model and Neighbor-Joining Tree building method. Lancelet Zhx full-length protein or lancelet ZFR was used for full-length Zhx tree building and ZFR tree building, respectively. No outgroup was used for gnathostome Zhx tree building. Consensus trees were built using bootstrap resampling with 100 replicates implementing a greedy clustering algorithm and a support threshold of 50%.

For Bayesian tree building, MUSCLE MSA were exported from Geneious software as phylip alignment files (.phy) and tested for best-fit models of evolution using Prottest3.4.2 software. Prottest3.4.2 execution was completed using the University of Kentucky Dell Intel Xenon64 Linux Cluster II. MUSCLE MSA were then exported from Geneious software as nexus alignment files (.nex) and analyzed using MrBayes3.2.6 software. MrBayes execution also was completed using the University of Kentucky Dell Intel Xenon64 Linux Cluster II. For Zhx full-length proteins, lancelet was used as an outgroup, and among site variation was set to invgamma for likelihood parameters. For the parameters of the phylogenetic model, we used a Jones Taylor Thornton rate matrix, with $10^6$ generations, sample frequency of 200, and burnin fraction to 0.25. For homeodomain analysis, no outgroups were used and among site variation was set to gamma for likelihood parameters. For the parameters of the phylogenetic model; we used a Jones Taylor Thornton rate matrix, with $3 \times 10^7$ generations, sample frequency of 200, and burnin fraction to 0.25. Sumt and Sump commands were completed after runs for both Zhx full-length and homeodomain runs, and consensus trees were uploaded into FigTree1.4.4 for visualization.

Compilation of MrBayes3.2.6 and Prottest3.4.2 to run on the cluster environment, including the incorporation of new Java script for Prottest3.4.2 compatibility, was completed by Vikram Gazula at the University of Kentucky Center for Computational Sciences.

## Results

Our preliminary analysis of several mammalian genomes identified genes predicted to encode all three Zhx proteins known to exist in mice and humans, all of which contained two zinc fingers and four or five homeodomains. To explore further the relationship of Zhx family members, we expanded our BLAST analysis to 22 vertebrate species ranging from elephant shark to humans (supplementary table 1, Supplementary Material online). Genomes of all species encoded single Zhx1, Zhx2, and Zhx3 proteins; in species where genomic data were available, the *Zhx1* and *Zhx2* genes were always found to be tightly linked, but unlinked to *Zhx3*, suggesting that they arose by tandem duplication of an ancestral gene; this linkage of *Zhx1* and *Zhx2* has been maintained across a wide range of gnathostomes (fig. 1). As in mouse and human (Peterson et al. 2011), the Zhx-protein-coding regions were contiguous (with the possible exception of the last several amino acids of Zhx3, at least in some species), suggesting a similar gene structure in all gnathostomes (Nail AN, Smith JJ, Peterson ML, Spear BT, in preparation). To verify that these predicted gnathostome Zhx proteins were homologous to mouse and human Zhx1, Zhx2, and Zhx3, phylogenetic analysis was performed. This confirmed that the predicted Zhx1, Zhx2, and Zhx3 proteins across all species clustered as expected, validating their use for further analysis (fig. 2).

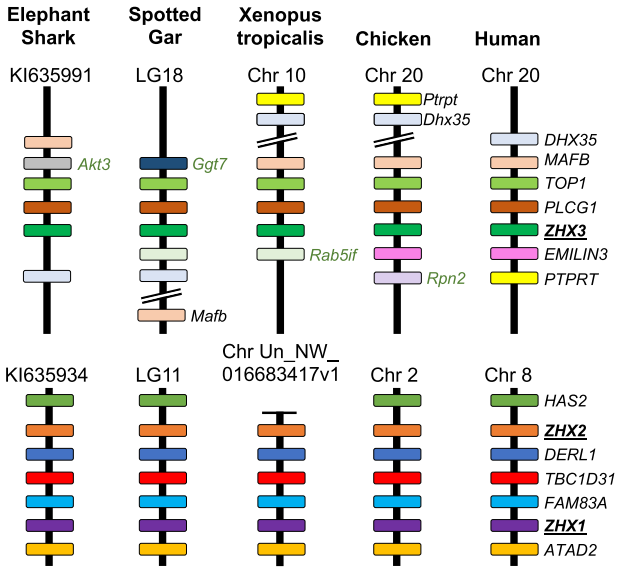


**Fig. 1.**—Conserved syntenic regions of *Zhx1*, *Zhx2*, and *Zhx3* in different vertebrate genomes. The Zhx3 region of synteny from elephant shark to humans is relatively small, with evidence of chromosomal rearrangements occurring near the *Zhx3* gene. The *Zhx1* and *Zhx2* genes are tightly linked in all species analyzed and the region containing these two genes has remained stable. *Akt3*, *thymoma viral proto-oncogene 3*; *ATAD2*, *ATPase family, AAA domain containing 2*; *DERL1*, *Der1-like domain family, member 1*; *DHX35*, *DEAH (Asp-Glu-Ala-His) box polypeptide 35*; *Emilin3*, *elastin microfibril interfacer 3*; *FAM83A*, *Family with sequence similarity 83, member A*; *Ggt7*, *gamma-glutamyltransferase 7*; *HAS2*, *hyaluronan synthase 2*; *MAFB*, *v-maf musculoaponeurotic fibrosarcoma oncogene family, protein B*; *PLCG1*, *phospholipase C, gamma 1*; *PTRPT*, *protein tyrosine phosphatase, receptor type, T*; *Rab5if*, *RAB5 interacting factor*; *Rpn2*, *ribophorin II*; *TBC1D31*, *TBC1 domain family, member 31*; *TOP1*, *topoisomerase (DNA) I*.

To identify conserved regions within gnathostome Zhx proteins, MUSCLE MSA was used to align these proteins from all 22 species. This analysis identified nine regions of homology among all Zhx proteins (fig. 3). As expected, the ZFR was highly conserved, as well as a short extension of roughly 22 amino acids on the carboxy side of the second zinc finger (ZFR-C in fig. 3). All homeodomains also showed high conservation although HD5, which appears to be present in a subset of Zhx proteins (see below), was the least conserved. In addition to these previously identified motifs, three additional regions of homology (A, B, and C in fig. 3) were

identified. Of these, the most intriguing was the highly conserved amino-terminal 13 amino acids (fig. 4). Eleven of the 13 amino acids are identical among all 66 gnathostome Zhx proteins; of the two that are not identical in all Zhx proteins, lysine is commonly found at position 4 and the related amino acids valine or isoleucine are always found at position 13. Downstream of this region, the similarity drops off dramatically. Regions "B" and "C," which are roughly 22 and 62 amino acids in length, are found between the ZFR and HD1 and between HD1 and HD2, respectively. BlastP analysis of these three regions against multiple protein databases failed to identify these motifs in other proteins, but their high conservation, particularly for the amino-terminal motif (region A), suggests that these regions are important for Zhx protein function.

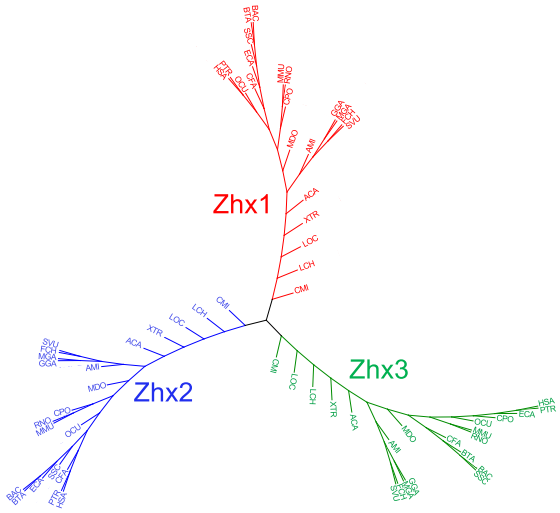


**Fig. 2.**—Zhx1, Zhx2, and Zhx3 proteins are found in 22 gnathostome species. Neighbor-Joining distance phylogenetic tree of full-length predicted Zhx proteins from species described in supplementary table 1, Supplementary Material online, using MUSCLE MSA. Zhx proteins cluster into Zhx1 (red), Zhx2 (blue), and Zhx3 (green) as predicted. Alligator (AMI), Anole (ACA), Chicken (GGA), Chimpanzee (PTR), Clawed Frog (XTR), Coelacanth (LCH), Cow (BTA), Dog (CFA), Elephant Shark (CMI), Guinea Pig (CPO), Horse (ECA), Human (HSA), Minke Whale (BAC), Mouse (MMU), Opossum (MDO), Pig (SSC), Rabbit (OCU), Rat (RNO), Saker Falcon (FCH), Spotted Gar (LOC), Starling (SVU), and Turkey (MGA).

To explore further the relationship of the zinc fingers among the different Zhx proteins, alignments were generated among gnathostome consensus sequences for the three Zhx paralogs. These alignments focused on a 54 amino acid region that was present within each Zhx paralog, containing two 22 amino acid $C_2$–$H_2$ zinc fingers for zinc finger 1 and zinc finger 2 and an intervening 10 amino acid spacer (fig. 5). The cysteines and histidines (fig. 5, black arrows) are invariant in both zinc fingers of all Zhx proteins. In addition to these, F14 in zinc finger 1 and F38, K41, L46, and N50 in zinc finger 2 (fig. 5, red arrows) are nearly invariant. Taken together, a total of 13 amino acids are highly conserved between all three Zhx proteins (fig. 5), with 8 of these conserved amino acids in zinc finger 2. These data also revealed that the Zhx1 and Zhx2 ZFRs are much more similar to each other than either is to the ZFR of Zhx3; in addition to the 13 amino acids conserved among all Zhx proteins, 22 are conserved only between Zhx1 and Zhx2, whereas 3 amino acids and 1 amino acid are shared between Zhx1 and Zhx3 and between Zhx2 and Zhx3, respectively (color coded in fig. 5). This similarity between Zhx1 and Zhx2 is not only in the ZFRs (59% and 68% for zinc finger 1 and zinc finger 2, respectively) but also in the 10 amino acid spacer region (70%). These data also indicate that the ZFRs of Zhx1 and Zhx2 are much more highly conserved among all gnathostomes (>90% identical) than Zhx3 (~66% identical) (table 1). This suggests that the function of the ZFR of Zhx1 and Zhx2 has been maintained across gnathostomes. These analyses also argue that Zhx3 zinc finger binding sites may be different than those of Zhx1 and Zhx2 and, because its homology among gnathostomes is lower (~66%), may be more divergent across gnathostomes.

To compare the homeodomain regions of different gnathostomes, alignments of 60 amino acid homeodomains were anchored using highly conserved L16, W48, F49, and R53 residues (fig. 6, black arrows). Because conservation of

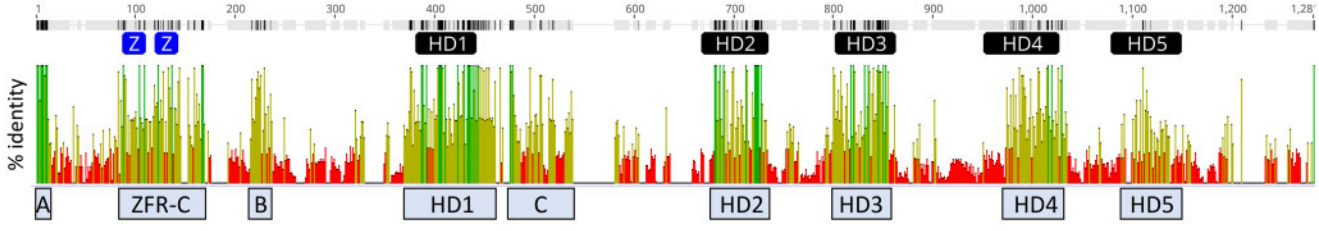


**Fig. 3.**—Zhx proteins from 22 gnathostome species contain 9 regions of high sequence similarity. Conservation from MSA of all gnathostome Zhx proteins; green indicates 100% identity, yellow represents 99–30% identity, and red indicates <30% identity. In the protein alignment along top, regions of higher similarity have darker shading; sequence gaps are also indicated. As expected, the ZFR including a region on the carboxy side of zinc finger 2 (entire region designated ZFR-C) and homeodomains (HD1-HD5) exhibits high similarity, although to a lesser extent with HD5 than the other homeodomains. Three other regions of high similarity (A–C) are noted. The highest homology is found within the first 13 amino acids at the amino terminus (region A).

HD5 was weak and predicted to be present in only a subset of Zhx1 and Zhx3 proteins, we focused on HDs 1–4. This analysis indicated that HD1 was the most highly conserved homeodomain among all Zhx proteins (80.8% overall identity), followed by HD3, HD2, and HD4 (68.7%, 68.0%, and 64.2% overall identify, respectively). Amino acids L16, W48, F49, and R53 were nearly invariant in all homeodomains. One striking exception to this was seen in Zhx2 HD1, which often had methionine rather than leucine at position 16. Several other highly conserved residues were P26, E30, and L34, the two latter of which were in helix 2 (fig. 6, red arrows). In contrast to the ZFR, in which Zhx1 and Zhx2 were much more similar to each other than to Zhx3, the homeodomain regions did not exhibit such a striking similarity between Zhx1 and Zhx2 (table 2). For example, Zhx1 HD1 and Zhx2 HD1 contained more residues that were common to Zhx3 HD1 (8/60 [highlighted in blue] and 7/60 [highlighted in green], respectively, in fig. 6), than to each other (3/60 [highlighted in orange]). For all four homeodomains, helix 3, which fits into the major groove of the DNA, is more conserved than helix 1 and helix 2, which are arranged in an antiparallel manner and span the major groove of the DNA (table 3) (Kornberg 1993; Gehring et al. 1994).

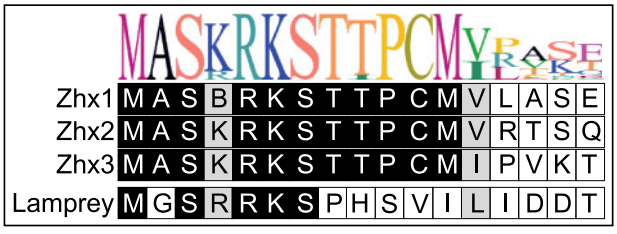


**Fig. 4.**—The N-terminal region of Zhx proteins is highly conserved among gnathostomes. MSA shows that 13 amino-terminal amino acids are almost invariant among gnathostome Zhx1, Zhx2, and Zhx3 proteins, although this homology quickly declines after this region. The top shows the amino acid sequence logo generated from alignment of all 66 gnathostome Zhx proteins. Zhx1, Zhx2, and Zhx3 amino acid sequences are shown below. "B" for Zhx1 represents K or R. Lamprey Zhx protein sequence shows a lower level of homology with gnathostome Zhx proteins.

The presence of three Zhx proteins in all analyzed gnathostomes led us to search for Zhx proteins in more divergent vertebrate and chordate lineages. BLAST analyses identified a single intact Zhx-like protein in sea lamprey (supplementary fig. 2*A*, Supplementary Material online), sea squirt, and lancelet that are respectively 1,079, 662, and 1394 amino acids in length (fig. 7; elephant shark Zhx proteins are included for comparison). The lamprey Zhx protein sequence was confirmed by sequencing of RT-PCR and 5′/3′ RACE amplicons from lamprey liver cDNA (GenBank MN823071). The lancelet and sea squirt Zhx proteins were predicted based on available sequence data, EST databases, and alignment with gnathostome Zhx proteins. The lamprey and lancelet Zhx proteins contained two zinc fingers and five or six homeodomains, respectively. In contrast, the sea squirt contained a single zinc finger and two homeodomains. Extensive BLAST analysis of nonchordate genomes, using full proteins, ZFRs and homeodomains, failed to identify any Zhx-like genes, indicating that the initial *Zhx*-like gene likely appeared in a common chordate ancestor.

To determine the evolutionary relationship of different Zhx proteins, phylogenetic analysis was used to compare lamprey and gnathostome Zhx proteins, using lancelet as an outgroup. The sea squirt Zhx protein, with only a single zinc finger and two homeodomains, was not included in this analysis because this protein was much smaller than other Zhx proteins. In addition, other studies have shown that sea squirt proteins often have high rates of amino acid substitution (Putnam et al. 2008; Alexandra-Louis et al. 2012). Our initial phylogenetic analysis using the Neighbor-Joining distance-based tree method (as shown in fig. 2) did not definitively determine the position of sea lamprey Zhx compared with gnathostome Zhx proteins (supplementary fig. 1*A*, Supplementary Material online). As distance-based trees often underestimate true branch distance because some sites may have undergone multiple substitution events or different rates of substitution events, we utilized Bayesian phylogenetic analysis as an alternative approach. Bayesian analysis, like distance tree building, was unable to resolve the position of sea lamprey Zhx relative to gnathostome proteins (supplementary fig. 1*B*, Supplementary Material online). However, in agreement

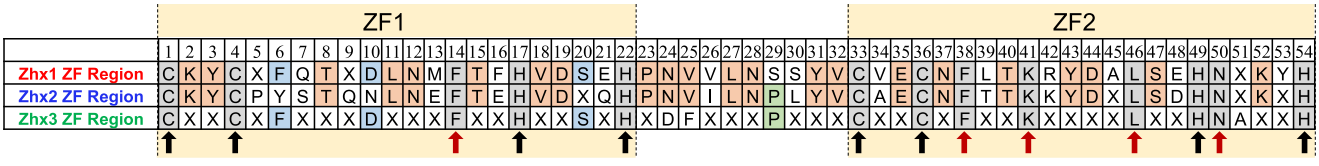


**Fig. 5.**—The $C_2$–$H_2$ zinc fingers regions of Zhx1 and Zhx2 are more similar to each other than to the same region of Zhx3. Alignment of the Zhx ZFR of all gnathostome Zhx proteins, composed of zinc finger 1 (ZF1; 22 aa), zinc finger 2 (ZF2; 22 aa), and the 10 amino acid intervening sequence. Residues common to all Zhx proteins are shaded in gray, those common to Zhx1 and Zhx2, Zhx1 and Zhx3, and Zhx2 and Zhx3 are shaded in orange, blue, and green, respectively. The cysteines and histidines that anchor zinc atoms are designated with black arrows; additional amino acids common to all proteins are designated with red arrows. "X" represents nonconserved amino acids. In addition to the Zhx1 and Zhx2 ZFRs being more similar to each other than either one is to Zhx3, zinc finger 2 is more conserved among the three Zhx proteins than zinc finger 1.

with distance tree building, Bayesian analysis indicated that gnathostome Zhx1 and Zhx2 are more similar to each other than to Zhx3.

Because our earlier analysis showed strong conservation of ZFRs across all gnathostome taxa (fig. 5), we considered whether this region could determine the relationship of lamprey and gnathostome Zhx proteins. Again, lancelet was used as an outgroup and sea squirt was not included. Using distance-based methods, this analysis yielded results indicating that the sea lamprey ZFR was most similar to the gnathostome Zhx3 ZFR (fig. 8). Consistent with our earlier analyses, gnathostome Zhx1 and Zhx2 clustered separately from Zhx3, albeit with low bootstrap support. We further examined the similarities among chordate Zhx proteins by homeodomain phylogenetic analyses. MUSCLE MSA was used to align 326 Zhx homeodomains from lancelet, sea squirt, lamprey, and all gnathostomes (fig. 9). For the most part, this analysis could not determine whether homeodomains of lamprey and lower chordates were most similar to gnathostome Zhx1, Zhx2, or Zhx3. The one exception was HD1, which is the most conserved homeodomain; lamprey HD1 was most similar the gnathostome HD1. Taken together, these sequence comparisons suggest that gnathostome Zhx3 proteins are generally more similar to Zhx proteins of outgroup chordates.

**Table 1**

Percent Identity of Zinc Finger 1 (ZF1), Zinc Finger 2 (ZF2), and Entire 54 Amino Acid ZFR among All 22 Gnathostome Zhx Proteins

| Zhx Protein | ZF 1 | ZF 2 | ZFR |
|---|---|---|---|
| Zhx1 | 94.9 | 93.3 | 94.3 |
| Zhx2 | 92.4 | 88.1 | 91.9 |
| Zhx3 | 65.7 | 66.8 | 65.8 |

## Discussion

In contrast to the majority of homeodomain proteins, which contain only a single homeodomain, Zhx proteins contain multiple homeodomains. Here, we describe a comprehensive analysis of Zhx proteins from 22 gnathostomes, lamprey, and 2 invertebrate chordates. These analyses have identified sequence regions that are common to all Zhx proteins and therefore likely to be important for Zhx protein function. Comparisons of Zhx proteins and their various domains provide insight into the evolutionary history of Zhx proteins subsequent to the origination of the first *Zhx* gene in a common chordate ancestor.

The amino-terminal ZFR is one of the most conserved domains among all Zhx proteins. $C_2H_2$ zinc fingers, the most common DNA-binding motif in metazoans, do not only bind DNA but are also known to bind RNA and mediate protein–protein interactions (Laity et al. 2001; Stubbs et al. 2011). In addition to the two cysteines and two histidines in each Zhx zinc finger, other conserved amino acids include phenylalanine located two amino acids after the second cysteine in zinc finger 1 and phenylalanine or leucine found three residues before the first histidine in both zinc fingers. Other conserved amino acids found in Zhx proteins, but not in other $C_2H_2$ zinc finger families, include lysine and asparagine in zinc finger 2 (positions 41 and 50 in fig. 5). We have also found that the stretch of ~30 amino acids after zinc finger 2, which was previously noted in comparisons involving a smaller

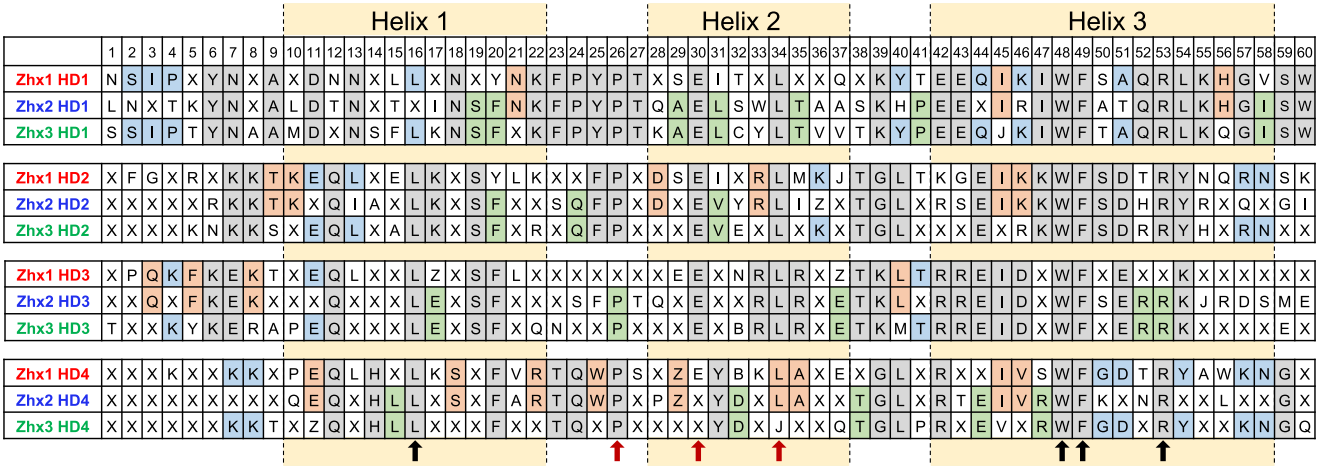


FIG. 6.—HD1 is the most conserved homeodomain among all gnathostome Zhx proteins. Alignment of the 60 amino acid homeodomains HD1–HD4 of all Zhx proteins, with the amino acid number and the helix 1, helix 2, and helix 3 regions, which are important for the homeodomain tertiary structure and DNA binding, shown at the top. Residues within each homeodomain that are common to Zhx1, Zhx2, and Zhx3 are shaded in gray, those common to Zhx1 and Zhx2, Zhx1 and Zhx3, or Zhx2 and Zhx3 are shaded in orange, blue, and green, respectively. "X" represents nonconserved amino acids among the 22 compared gnathostomes. The highly conserved L16, W48, F49, and R53 residues that were used to anchor the 60 amino acid homeodomain region are designated with black arrows. Additional residues that are highly conserved, P26, E30, and L34 are designated with red arrows. The homeodomain regions of Zhx1 and Zhx2 are no more similar to each other than to Zhx3, which is in contrast to the ZFR.

number of Zhx proteins (Wienk et al. 2009), is highly conserved among all Zhx proteins (fig. 3), albeit to a lesser degree outside of the gnathostomes. Previous NMR studies indicated that this zinc finger extension region, which is unique to Zhx proteins, forms two $\beta$-sheets that interact with the two $\beta$-sheets of zinc finger 2 (Wienk et al. 2009). Although the ZFRs of Zhx1 and Zhx2 are much more highly conserved between gnathostomes than the ZFR of Zhx3, structural analysis predicted that all three have similar biological function but may interact with different substrates (Wienk et al. 2009). This previous NMR study suggested that the Zhx zinc fingers are involved in protein binding rather than DNA or RNA binding (Wienk et al. 2009).

Zhx homeodomains contain many residues that conform to the homeodomain consensus sequence. Nearly, all Zhx homeodomains contained highly conserved L16, W48, F49, and R53 residues, in agreement with most other homeodomains; the primary exception being position 16 in Zhx2 HD1, which was often methionine (also in HD4 of elephant shark, fig. 7). These conserved residues are important for formation of a hydrophobic core that helps maintain the overall tertiary structure of the homeodomain. HD5 was not present in all Zhx proteins and, in many cases, lacked these key conserved homeodomain residues, suggesting that HD5 may not be required for the core function of Zhx proteins. HD1 was the most conserved homeodomain among all Zhx proteins, which might suggest that this homeodomain has maintained the same function in all Zhx proteins. These findings seem to be generally consistent with previous efforts to express and crystallize multiple domains of all three human ZHX proteins only obtained crystals for ZHX1 HD4 and ZHX2 HD2 (Bird et al. 2010). The structural analyses identified an extra C-terminal helix in ZHX1 HD4, compared with most other homeodomain structures, that is predicted to increase stability. The ZHX2 HD2 structure had a low level of similarity to other characterized homeodomains and an unusual conformation that was predicted to disrupt DNA binding (Bird et al. 2010). Unique aspects of these two characterized Zhx homeodomains raise the possibility that the presence of multiple homeodomains might allow some of them to gain new structures and functions, whereas this flexibility might not be possible in proteins that contain a single homeodomain. Structural and functional analysis of additional Zhx homeodomains will be needed to address this possibility. Another interesting question is how Zhx proteins may have obtained multiple homeodomains. It seems likely that duplication of a primordial homeodomain gave rise to the homeodomains found in chordate Zhx proteins. However, our homology analyses did not reveal a clear relationship between the multiple homeodomains or with homeodomains on nonchordate species, including echinoderms.

In addition to the zinc fingers and homeodomains, three additional regions of high similarity were found among Zhx proteins, all of which appear to be unique to the Zhx protein

**Table 2**

Number of Common Amino Acids within HD Regions among Zhx1, Zhx2, and Zhx3 Proteins from 22 Gnathostomes

| Amino Acids within HD | HD1 | HD2 | HD3 | HD4 |
|---|---|---|---|---|
| Common to 1–3 | 27 | 21 | 20 | 15 |
| Common only to 1 and 2 | 3 | 6 | 4 | 9 |
| Common only to 1 and 3 | 8 | 5 | 4 | 7 |
| Common only to 2 and 3 | 7 | 3 | 6 | 5 |

**Table 3**

Percent Identity of Helix 1, Helix 2, and Helix 3 for HD1, HD2, HD4, and HD4 across 22 Gnathostome Zhx Proteins

| | Helix 1 | Helix 2 | Helix 3 |
|---|---|---|---|
| HD 1 | 59.5 | 52.2 | 80.8 |
| HD 2 | 61.8 | 54.3 | 68.0 |
| HD 3 | 58.4 | 61.8 | 68.7 |
| HD 4 | 58.7 | 51.1 | 64.2 |

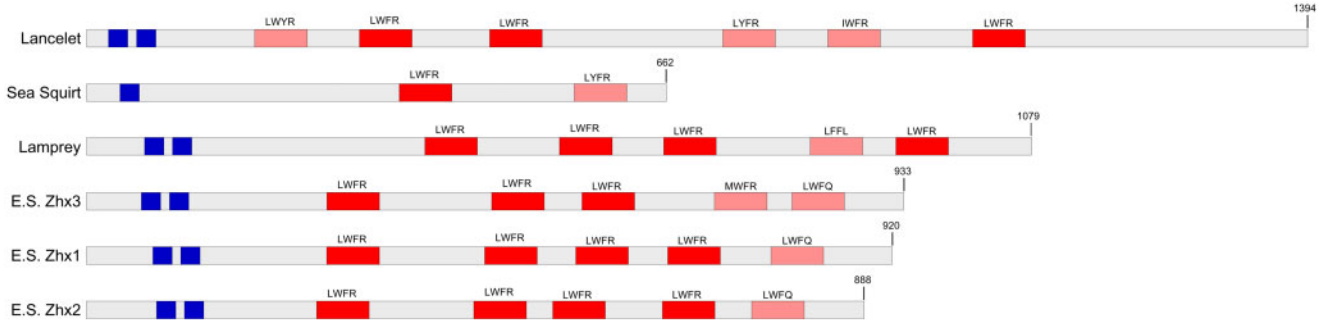


**Fig. 7.**—Diagram of lancelet, sea squirt, sea lamprey, and elephant shark Zhx proteins. Proteins are based on analysis of genome databases and alignment with known Zhx proteins except for sea lamprey, in which protein sequence is based on RT-PCR and 3′/5′ race of sea lamprey liver cDNA. The length of Zhx proteins from these four species and location of zinc fingers (blue) and homeodomains (red/pink) are shown. Amino acids for positions 16, 48, 49, and 53 of the homeodomains are shown above boxes, those with canonical L16, W48, F49, and R53 are in red, those with noncanonical amino acids at any of these four positions are in pink.


**Fig. 8.**—Neighbor-distance joining phylogeny using ZFRs places sea lamprey within the gnathostome Zhx3 lineage. ZFRs across 22 gnathostome species, lancelet, and sea lamprey were aligned using MUSCLE MSA followed by phylogenetic distance tree analysis (Geneious Tree Builder) with Jukes–Cantor genetic distance model and Neighbor-Joining tree building. Consensus tree was built using 100 replicates for bootstrapping and majority greedy clustering, and lancelet Zhx (BFL_Zhx) was used as an outgroup. Consensus support based on bootstrapping is shown at each branch base. Amino acid substitution length is shown at the bottom of each tree with a scale bar (0.3). Sea lamprey Zhx ZFR is most closely related to that of gnathostome Zhx3. Zhx1 (red) and Zhx2 (blue) ZFRs are most closely related to each other than to Zhx3 (green), which also shows a higher sequence divergence across taxa. Alligator (AMI), Anole (ACA), Chicken (GGA), Chimpanzee (PTR), Clawed Frog (XTR), Coelacanth (LCH), Cow (BTA), Dog (CFA), Elephant Shark (CMI), Guinea Pig (CPO), Horse (ECA), Human (HSA), Lancelet (BFL), Minke Whale (BAC), Mouse (MMU), Opossum (MDO), Pig (SSC), Rabbit (OCU), Rat (RNO), Saker Falcon (FCH), Sea Lamprey (PMA), Spotted Gar (LOC), Starling (SVU), and Turkey (MGA).

family. Of these, we are particularly intrigued by the amino terminus of gnathostome Zhx proteins in which 11 of the first 13 amino acids are conserved. Protein termini have numerous unique features that can influence function, stability, and localization (Lange and Overall 2013). Ends of proteins are often less rigid than other protein regions and extend from the

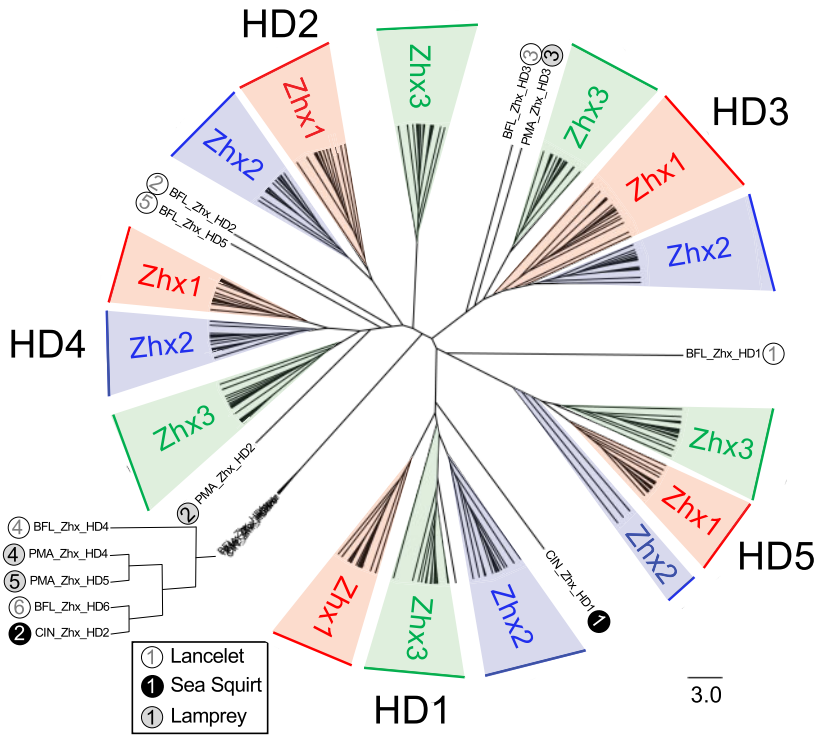


**Fig. 9.**—Bayesian phylogeny of all Zhx homeodomains. Bayesian phylogenetic analysis of MUSCLE MSA sequence alignment from 326 homeodomains across all Zhx proteins. Trees were run for 30,000,000 generations (standard deviation of split frequency = 0.008, see Materials and Methods for more details). No outgroups were used for this analysis. Shown is the consensus tree from combined simultaneous runs that were visualized in FigTree1.4.4 software. All Zhx homeodomains are shown. Zhx1, Zhx2, and Zhx3 HDs are highlighted in red, blue, and green, respectively. Gray circles represent lamprey homeodomains (1–5), black circles indicate sea squirt HD1 and HD2, and open circles represent lancelet homeodomains (1–6). Panel on the left is zoomed in to show the relationship of lancelet HD4 to sea lamprey HD4 and HD5, lancelet HD6, and sea squirt HD2. The homeodomains of Zhx1 and Zhx2 cluster with each other. Lancelet and lamprey HD3 clusters with gnathostome HD3. Sea squirt HD1 clusters with gnathostome HD1, whereas lamprey HD1 clusters specifically with Zhx3 HD1.

globular protein structure (Carugo 2017). Furthermore, amino termini, the initial translated region, are frequently subjected to posttranslational modification (Lange and Overall 2013). Interestingly, many of the conserved residues in the Zhx amino terminus, including two serines, two threonines, two lysines, cysteine, arginine, and proline, are frequent targets of posttranslational modification (Seo and Lee 2004). Although only 5 of the 13 amino-terminal amino acids of the lamprey Zhx protein are conserved with gnathostomes (fig. 4), 8 of these 13 are amino acids that are frequent targets of modification.

Our analysis confirmed the presence of one Zhx protein in lamprey, sea squirt, and lancelet, but we were unable to find evidence that any Zhx-like proteins existed prior to the emergence of a common chordate ancestor, indicating that the primordial *Zhx* gene likely arose roughly 700 Ma (Kumar et al. 2017). Although we describe here a single lamprey protein, the publicly available sea lamprey germline genome assembly (https://genomes.stowers.org/organism/Petromyzon/marinus; last accessed March 9, 2020) contains two annotated *Zhx*-like genes (PMZ_0033201-RA and PMZ_0033200-RA) that are adjacent to each other (Smith et al. 2018). Both predicted genes are in the same orientation and separated by 30 nucleotides. The 5′ gene is predicted to encode two zinc fingers and two homeodomains that are most similar to HD1 and HD2 of other Zhx proteins, whereas the second gene is predicted to contain a single homeodomain that is most similar to HD4; the 30-bp region encodes a stop codon (supplementary fig. 2, Supplementary Material online). We propose that these two predicted genes are in fact a single pseudogene (Nail AN, Smith JJ, Peterson ML, Spear BT, in preparation). Notably, this same stop codon is present in a newer high-quality reference genome that was released by the Vertebrate Genomes Project during revision of this article (https://vgp.github.io/genomeark/Petromyzon_marinus/; last accessed March 9, 2020).

In contrast to sea squirt, lancelet, and lamprey, all gnathostomes analyzed contained three Zhx proteins. We propose that a whole genome duplication event preceding the emergence of lamprey duplicated the original *Zhx* gene, resulting in Zhx3, which maintained a structure and presumably a function similar to the primordial protein, and the Zhx1/Zhx2 precursor. This precursor may be related to the *Zhx*-like pseudogene in lamprey. We believe that a tandem duplication in the ancestral gnathostome lineage gave rise to Zhx1 and Zhx2 from the Zhx1/Zhx2 precursor, based on the high

similarity of Zhx1 and Zhx2 proteins and that the Zhx1 and Zhx2 genes are tightly linked in all gnathostomes for which genome assemblies are currently available, including both chondrichthyans and bony vertebrates (Kent et al. 2002). Searches of available fish genomes on the University of California–Santa Cruz genome browser and ENSEMBL indicate that duplicated copies of all three *Zhx* genes exist in most teleost fish species (Nail AN, Smith JJ, Peterson ML, Spear BT, in preparation), consistent with the whole genome duplication that occurred in this clade.

Little is known about the function of Zhx proteins, including the extent to which their functions are redundant or distinct. Data indicating that Zhx proteins homodimerize and heterodimerize with each other as well as with the transcriptional activator NF-YA could be considered evidence supporting the idea that their functions overlap, at least to some extent. Zhx1 binds the DNA methyltransferase DNMT3B (Kim et al. 2007) and transcriptional corepressor BS69 (Ogata-Kawata et al. 2007), although neither of these studies examined Zhx2 or Zhx3 binding to these proteins. Zhx2, but neither Zhx1 nor Zhx3, contains a proline-rich region between HD1 and HD2. A recent study found that three proline residues in human Zhx2, two in this region and one within HD2, can be hydroxylated and that this modification leads to Zhx2 degradation by von Hippel-Lindau (VHL) E3 ubiquitin ligase (Zhang et al. 2018). The two prolines in the proline-rich region are conserved among Zhx2 proteins but are not present in Zhx1 or Zhx3, and neither of these proteins appear to be degraded by VHL protein (Zhang et al. 2018). Curiously, the proline within HD2, which appears to be most important for binding VHL, is highly conserved in HD2 of all Zhx proteins, including those in lamprey and lancelet, but is not commonly found in other homeodomains.

Studies in BALB/cJ mice and Zhx2 knock-out mice have identified numerous genes that are dysregulated in the absence of Zhx2 in the liver, although it is not clear which of these genes are direct or indirect Zhx2 targets or whether Zhx1 and/or Zhx3 also regulate these genes. Several human GWAS studies have implicated Zhx2 in cardiovascular disease, which is consistent with mouse data linking Zhx2 with serum lipid levels and atherosclerosis (Bis et al. 2011; Li et al. 2015), and one small-scale human GWAS study found a possible association between Zhx3 and serum lipid levels (Johansen et al. 2011). Changes in Zhx proteins have been associated with various cancers, including HCC, renal cell carcinoma, multiple myeloma, and Hodgkin lymphoma (Armellini et al. 2008; Nagel Schneider et al. 2011; Yue et al. 2012; Zhang et al. 2018). Future studies, using both animal and in vitro models combined with structural analysis, will be needed to understand the regulatory functions of the Zhx protein family. Although these proteins appear to be transcriptional regulators, a full understanding of their gene targets, essential domains, and interacting proteins will be needed to elucidate their function during development and disease. In addition, continued analysis of the proteins and genes of this small family should provide insight into the evolution of regulatory mechanisms that have contributed to chordate evolution.

## Supplementary Material

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

## Acknowledgments

The authors thank Kristofer Schroder and Jordan Laferty for technical assistance and early efforts on this project and Vikram Gazula at the University of Kentucky Center for Computational Sciences for providing support and computing time on the Lipscomb High Performance Computing Cluster. This work was supported by grants DK059866 and DK074816 from the National Institute of Diabetes and Digestive and Kidney Diseases.

## Literature Cited

Alexandra-Louis A, Roest Crollius H, Robinson-Rechavi M. 2012. How much does the amphioxus genome represent the ancestor of chordates? Brief Funct Genomics. 11(2):89–95.

Armellini A, et al. 2008. Low expression of ZHX2, but not RCBTB2 or RAN, is associated with poor outcome in multiple myeloma. Br J Haematol. 141(2):212–215.

Barthelemy I, et al. 1996. zhx-1: a novel homeodomain protein containing two zinc-fingers and five homeodomains. Biochem Biophys Res Commun. 224(3):870–876.

Belayew A, Tilghman SM. 1982. Genetic analysis of α-fetoprotein synthesis in mice. Mol Cell Biol. 2(11):1427–1435.

Bird LE, et al. 2010. Novel structural features in two ZHX homeodomains derived from a systematic study of single and multiple domains. BMC Struct Biol. 10(1):13.

Bis JC, et al. 2011. Meta-analysis of genome-wide association studies from the CHARGE consortium identifies common variants associated with carotid intima media thickness and plaque. Nat Genet. 43(10):940–947.

Burglin TR, Affolter M. 2016. Homeodomain proteins: an update. Chromosoma 125:497–521.

Carugo O. 2017. Protein termini. Curr Protein Pept Sci. 18(3):211–216.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32(5):1792–1797.

Gehring WJ, Affolter M, Burglin T. 1994. Homeodomain proteins. Annu Rev Biochem. 63(1):487–526.

Hu S, et al. 2007. Expression of zinc-fingers and homeoboxes 2 in hepatocellular carcinogenesis: a tissue microarray and clinicopathological analysis. Neoplasma 54(3):207–211.

Jiang J, Creasy KT, Purnell J, Peterson ML, Spear BT. 2017. Zhx2 (zinc fingers and homeoboxes 2) regulates major urinary protein gene expression in the mouse liver. J Biol Chem. 292(16):6765–6774.

Johansen CT, Kathiresan S, Hegele RA. 2011. Genetic determinants of plasma triglycerides. J Lipid Res. 52(2):189–206.

Kawata H, et al. 2003a. The mouse zinc-fingers and homeoboxes (ZHX) family; ZHX2 forms a heterodimer with ZHX3. Gene 323:133–140.

Kawata H, et al. 2003b. Zinc-fingers and homeoboxes (ZHX) 2, a novel member of the ZHX family, functions as a transcriptional repressor. Biochem J. 373:747–757.

Kearse M, et al. 2012. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 28(12):1647–1649.
Kent WJ, et al. 2002. The human genome browser at UCSC. Genome Res. 12(6):996–1006.
Kim SH, et al. 2007. Zinc-fingers and homeoboxes 1 (ZHX1) binds DNA methyltransferase (DNMT) 3B to enhance DNMT3B-mediated transcriptional repression. Biochem Biophys Res Commun. 355(2):318–323.
Kornberg TB. 1993. Understanding the homeodomain. J Biol Chem. 268(36):26813–26816.
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.
Laity JH, Lee BM, Wright PE. 2001. Zinc finger proteins: new insights into structural and functional diversity. Curr Opin Struct Biol. 11(1):39–46.
Lange PF, Overall CM. 2013. Protein TAILS: when termini tell tales of proteolysis and function. Curr Opin Chem Biol. 17(1):73–82.
Li C, et al. 2015. Genetic association and gene-smoking interaction study of carotid intima-media thickness at five GWAS-indicated genes: the Bogalusa Heart Study. Gene 562(2):226–231.
Liu G, Clement LC, Kanwar YS, Avila-Casado C, Chugh SS. 2006. ZHX proteins regulate podocyte gene expression during the development of nephrotic syndrome. J Biol Chem. 281(51):39681–39692.
Luo Z, Rhie SK, Farnham PJ. 2019. The enigmatic *HOX* genes: can we crack their code? Cancers (Basel). 11(3):323.
Lv Z, et al. 2006. Promoter hypermethylation of a novel gene, *ZHX2*, in hepatocellular carcinoma. Am J Clin Pathol. 125(5):740–746.
Morford LA, et al. 2007. The oncofetal gene *glypican 3* is regulated in the postnatal liver by zinc fingers and homeoboxes 2 and in the regenerating liver by alpha-fetoprotein regulator 2. Hepatology 46(5):1541–1547.
Nagel S, et al. 2011. t(4;8)(q27;q24) in Hodgkin lymphoma cells targets phosphodiesterase PDE5A and homeobox gene *ZHX2*. Genes Chromosomes Cancer 50:996–1009.
Ogata-Kawata H, Yamada K, Uesaka-Yoshino M, Kagawa N, Miyamoto K. 2007. BS69, a corepressor interacting with ZHX1, is a bifunctional transcription factor. Front Biosci. 12(1):1911–1926.
Olsson M, Lindahl G, Roushlahti E. 1977. Genetic control of alpha-fetoprotein synthesis in the mouse. J Exp Med. 145(4):819–830.
Pachnis V, Belayew A, Tilghman SM. 1984. Locus unlinked to α-fetoprotein under the control of the murine *raf* and *Rif* genes. Proc Natl Acad Sci U S A. 81(17):5523–5527.
Perincheri S, Dingle RW, Peterson ML, Spear BT. 2005. Hereditary persistence of alpha-fetoprotein and H19 expression in liver of BALB/cJ mice is due to a retrovirus insertion in the *Zhx2* gene. Proc Natl Acad Sci U S A. 102(2):396–401.
Peterson ML, Ma C, Spear BT. 2011. Zhx2 and Zbtb20: novel regulators of postnatal alpha-fetoprotein repression and their potential role in gene reactivation during liver cancer. Semin Cancer Biol. 21(1):21–27.
Putnam NH, et al. 2008. The amphioxus genome and the evolution of the chordate karyotype. Nature 453(7198):1064–1071.
Seo J, Lee KJ. 2004. Post-translational modifications and their biological functions: proteomic analysis and systematic approaches. J Biochem Mol Biol. 37:35–44.
Smith JJ, et al. 2018. The sea lamprey germline genome provides insights into programmed genome rearrangement and vertebrate evolution. Nat Genet. 50(2):270–277.
Spear BT, Jin L, Ramasamy S, Dobierzewska A. 2006. Transcriptional control in the mammalian liver: liver development, perinatal repression, and zonal gene regulation. Cell Mol Life Sci. 63(24):2922–2938.
Stubbs L, Sun Y, Caetano-Anolles D. 2011. Function and evolution of C2H2 zinc finger arrays. Subcell Biochem. 52:75–94.
Wienk HI, et al. 2009. The tandem zinc-finger region of human ZHX adopts a novel C2H2 zinc finger structure with a C-terminal extension. Biochemistry 48(21):4431–4439.
Yamada K, et al. 2002. Functional analysis and the molecular dissection of zinc-fingers and homeoboxes 1 (ZHX1). Biochem Biophys Res Commun. 297(2):368–374.
Yamada K, et al. 2003. Analysis of zinc-fingers and homeoboxes (ZHX)-1-interacting proteins: molecular cloning and characterization of a member of the ZHX family, ZHX3. Biochem J. 373(1):167–178.
Yamada K, Printz RL, Osawa H, Granner DK. 1999. Human ZHX1: cloning, chromosomal location, and interaction with transcription factor NF-Y. Biochem Biophys Res Commun. 261(3):614–621.
Yue X, et al. 2012. Zinc fingers and homeoboxes 2 inhibits hepatocellular carcinoma cell proliferation and represses expression of Cyclins A and E. Gastroenterology 142(7):1559–1570.
Zhang J, et al. 2018. VHL substrate transcription factor ZHX2 as an oncogenic driver in clear cell renal cell carcinoma. Science 361(6399):290–295.

**Associate editor:** Sabyasachi Das