Connectivity of vertebrate genomes: Paired-related homeobox (Prrx) genes in spotted gar, basal teleosts, and tetrapods.

Ingo Braasch, Yann Guiguen, Ryan Loker, John H Letaw, Allyse Ferrara, Julien Bobe, John H Postlethwait

Research Article

Connectivity of vertebrate genomes: *Paired-related homeobox* (*Prrx*) genes in spotted gar, basal teleosts, and tetrapods

Ingo Braasch[a], Yann Guiguen[b], Ryan Loker[a], John H. Letaw[a,1], Allyse Ferrara[c], Julien Bobe[b], and John H. Postlethwait[a]

[a]Institute of Neuroscience, University of Oregon, Eugene 97403-1254, OR, USA; [b]INRA, UR1037 LPGP, Campus de Beaulieu, F-35000 Rennes, France; [c]Department of Biological Sciences, Nicholls State University, Thibodaux, LA 70310, USA

Email addresses:
Ingo Braasch: ibraasch@uoneuro.uoregon.edu
Yann Guigen: yann.guiguen@rennes.inra.fr
Ryan Loker: loker@uoregon.edu
John H. Letaw: letaw@ohsu.edu
Allyse Ferrara: allyse.ferrara@nicholls.edu
Julien Bobe: Julien.Bobe@rennes.inra.fr
John H. Postlethwait: jpostle@uoneuro.uoregon.edu

Correspondings author:
John H. Postlethwait, Institute of Neuroscience, University of Oregon, 1425 E. 13th Avenue, Eugene OR 97403, USA
Phone: +1 (541) 346-4538; Email: jpostle@uoneuro.uoregon.edu

[1]Present address: Oregon National Primate Research Center, Division of Neuroscience, Oregon Health and Science University, 505 NW 185th Ave, Beaverton, OR 97006, USA

Abstract

Teleost fish are important models for human biology, health, and disease. Because genome duplication in a teleost ancestor (TGD) impacts the evolution of teleost genome structure and gene repertoires, we must discriminate gene functions that are shared and ancestral from those that are lineage-specific in teleosts or tetrapods to accurately apply inferences from teleost disease models to human health. Generalizations must account both for the TGD and for divergent evolution between teleosts and tetrapods after the likely two rounds of genome duplication shared by all vertebrates. Progress in sequencing techniques provides new opportunities to generate genomic and transcriptomic information from a broad range of phylogenetically informative taxa that facilitate detailed understanding of gene family and gene function evolution.

We illustrate here the use of new sequence resources from spotted gar (*Lepisosteus oculatus*), a rayfin fish that diverged from teleosts before the TGD, as well as RNA-Seq data from gar and multiple teleosts lineages to reconstruct the evolution of the *Paired-related homeobox* (*Prrx*) transcription factor gene family involved in the development of mesoderm and neural crest-derived mesenchyme. We show that for *Prrx* genes, the spotted gar genome and gene expression patterns mimic mammals better than teleosts do. Analyses force the seemingly paradoxical conclusion that the limb expression domains of *Prrx* genes existed before the evolution of paired appendages. Detailed evolutionary analyses like those reported here are required to identify fish species most similar to the human genome to optimally connect fish models to human gene functions in health and disease.

Keywords: Prrx1, Prrx2, aquatic medical model, vertebrate, genome duplication, ohnolog, RNA-Seq, paired appendages, fin/limb bud, craniofacial

1. Introduction

Several teleost fish species, most prominently zebrafish, medaka, platyfish, stickleback, and killifish, are used as model species for human development, physiology, health, and disease (reviewed in Schartl, 2013). Despite their advantages as laboratory research models – such as easy husbandry, high fertility, external fertilization, transparency of embryos, tractable genetics, and amenability for high-throughput drug screens – interpretations of results obtained from teleosts species are challenged by ~900 million years of independent evolution that separates them from the *condition humaine*: Since the last fish-like bony vertebrate (euteleostome) ancestor that lived ~450 million years ago (Hedges et al., 2006), the lobefin vertebrate (sarcopterygian) lineage that led to tetrapods, mammals and later humans has evolved independently of the rayfin vertebrate (actinopterygian) lineage, including teleost fishes, which have significantly remodeled their morphology and, importantly, their genomes since the euteleostome ancestor.

Two rounds of vertebrate genome duplication (VGD1 and VGD2; Fig. 1) likely occurred at the stem of the vertebrate branch (Dehal and Boore, 2005; Putnam et al., 2008) and, despite the overall high level of conservation of genes between teleosts and tetrapods (Howe et al., 2013), important differences characterize the arrangement of specific gene families in both lineages. The lineages of teleosts and tetrapods have divergently retained gene duplicates from whole genome duplications, a class of paralogs generally termed 'ohnologs' (Wolfe, 2001). For example, a central regulator of pluripotency in mammals, *Pou5f1* (also known as *Oct3* or *Oct3/4*), was lost in the rayfin fish lineage (Frankenberg and Renfree, 2013). The tetrapod genome, on the other hand, has undergone substantial loss of ancestral genes during the water-to-land transition as well, for example *actinodin* and *fgf24* (Amemiya et al., 2013).

In addition to these instances of 'ohnologs gone missing' (Postlethwait, 2007) leading to the divergence of gene contents in teleost and tetrapod genomes, an additional round of whole genome duplication occurred in an ancestor of the teleost lineage, the teleost genome duplication or TGD; 12-24% of genes have been retained

as two paralogous genes in teleosts compared to one gene in tetrapods (reviewed in Braasch and Postlethwait, 2012). TGD paralogs, also termed co-orthologs, like VGD1 and VGD2 ohnologs have often changed functions by mechanisms such as subfunctionalization, *i.e.* the partitioning of ancestral gene functions among duplicates, and/or neofunctionalization, *i.e.* the acquisition of new gene functions in one or both co-orthologs (Force et al., 1999; Postlethwait et al., 2004). These differences in gene repertoires and gene functions between teleosts and tetrapods can make it difficult to transfer knowledge obtained in a teleost model species to the human condition.

Until recently, genomic sequence information was restricted to a few teleost model species, *i.e.* zebrafish (Howe et al., 2013), medaka (Kasahara et al., 2007), stickleback (Jones et al., 2012), and pufferfishes (Aparicio et al., 2002; Jaillon et al., 2004), but progress in sequencing techniques has enabled the relatively cheap and fast generation of genomic and transcriptomic sequence data from numerous fish species.

The present study illustrates how the availability of sequence information from a wide range of phylogenetically informative fish species can improve our understanding of gene function evolution among vertebrates, thereby better informing the suitability of teleost models for the analysis of specific gene functions and associated diseases. To this end, we take advantage of genomic sequence information from the spotted gar (*Lepisosteus oculatus*), a member of the closest living sister lineage to the teleosts, the holosteans (gars and bowfin), which diverged from teleosts before the TGD (Amores et al., 2011) (Fig. 1). The spotted gar offers a unique opportunity to study gene functions in a non-teleost, unduplicated rayfin species that is suitable for gene function analysis in a laboratory environment (Amores et al., 2011; Braasch and Postlethwait, 2012). Basally diverging teleost lineages, such as elopomorphs and osteoglossomorphs (Fig. 1), on the other hand, offer the opportunity to gain better understanding of the evolutionary mechanisms leading to the divergence of gene repertoires and gene functions shortly after the rise of the teleost lineage and the TGD.

We illustrate the use of these new sequence resources by the example of *Paired-related homeobox* (*Prrx*) genes. *Prrx* genes encode transcription factors characterized by a paired type DNA-binding homeodomain and a C-terminal OAR domain functioning as an interface for cofactor interactions (Norris and Kern, 2001). *Prrx* genes play essential roles in several developmental processes, particularly of mesoderm and neural crest-derived mesenchyme (Cserjesi et al., 1992; Kern et al., 1992; Opstelten et al., 1991). In tetrapods, *Prrx* genes are expressed in mesenchymal tissues like the limb buds, cranial and axial mesoderm, and branchial arches as well as during brain and heart development (Beverdam and Meijlink, 2001; Kuratani et al., 1994; Leussink et al., 1995; Lu et al., 1999; Nohno et al., 1993; Opstelten et al., 1991; Takahashi et al., 1998). Loss of *Prrx* functions in mice result in malformations in the skull, jaws, limbs, ears, and vasculature (Bergwerff et al., 2000; Martin et al., 1995; ten Berge et al., 1998) and mutations in human PRRX1 are associated with the agnathia-otocephaly complex (OMIM: #202650). Prrx1 is involved in pancreas development and regeneration (Reichert et al., 2013), has been identified as an inducer of epithelial-to-mesenchymal transitions, and its loss is required for cancer metastasis (Ocana et al., 2012). Furthermore, Prrx1 is important for the maintenance of adult neural stem cells in mammals (Shimozaki et al., 2013).

The expression domains of *Prrx2*, a second *Prrx* gene, overlap with many expression domains of *Prrx1* in tetrapods (Leussink et al., 1995; Lu et al., 1999), yet *Prrx2* seems to be absent from teleosts (Hernandez-Vega and Minguillon, 2011). The importance of *PRRX* genes for human development and health and specific differences in the repertoire of *Prrx* genes between teleost and tetrapods reported in the literature make this gene family a good example to illustrate the use of new fish sequence resources to improve the connectivity of vertebrate gene functions.

2. Material and Methods

2.1 Animals and tissue collection

In the Phylofish project, ten tissues/organs (ovary, testis, brain, gills, intestine, liver, bones, kidney, muscle, heart) were collected from adult spotted gar, zebrafish (*Danio rerio*), medaka (*Oryzias latipes*), Northern pike (*Esox lucius*), allis shad (*Alosa alosa*), striped catfish (*Pangasianodon hypophthalmus*), butterflyfish (*Pantodon buchholzi*), and European eel (*Anguilla anguilla*). See Fig. 1 for species relationships. In some species (spotted gar, zebrafish, medaka, Northern pike, striped catfish), embryos were also collected around the eyed stage. In European eel, a leptocephalus stage was used in place of the embryonic sample.

2.2 RNA-seq

Tissues were homogenized in Tri-reagent (Sigma, St-Louis, USA) at a ratio of 100 mg of tissue per ml of reagent and total RNA was extracted according to manufacturer's instructions. RNA quality was checked on a Bioanalyzer 2100 (Aligent, Santa Clara, CA). Sequencing libraries were prepared using a TruSeq RNA sample preparation kit according to the manufacturer's instructions (Illumina, San Diego, CA). Poly-A-containing mRNA was isolated from total RNA using poly-T oligo-attached magnetic beads, and chemically fragmented. First strand cDNA was generated using SuperScript II reverse transcriptase and random primers. Following the second strand cDNA synthesis and adaptor ligation, cDNA fragments were amplified by PCR. Amplification products were loaded onto an Illumina HiSeq2000 instrument and subjected to multiplexed paired-end (2 × 100 bp) sequencing. The processing of fluorescent images into sequences, base-calling and quality value calculations were performed using the Illumina data processing pipeline. Coding sequences of *prrx* genes were obtained from *de novo* assembly of cDNAs performed in a library-specific manner using Velvet and Oases (Schulz et al., 2012). Sequences of *prrx* transcripts were submitted to GenBank (accession numbers **KF841587 - KF841600**).

2.3 Identification and genomic annotation of *prrx* genes

RNA-Seq derived transcripts were used to annotate *prrx* genes in the genomes of spotted gar (genome assembly *LepOcu1*; GenBank Assembly ID: GCA_000242695.1), European eel (Henkel et al., 2012a), and Japanese eel (*Anguilla japonica*) (Henkel et al., 2012b). Agnathan *prrx* was identified by tblastn–guided manual annotation of genomic sequences in the genome of Artic lamprey (*Lethenteron camtschaticum*) (GenBank Assembly ID: GCA_000466285.1; Mehta et al., 2013). Chondrichthyan *prrx* transcripts were identified by tblastn in the transcriptomes of little skate (*Leucoraja erinacea*), small spotted catshark (*Scyliorhinus canicula*), and elephant shark (*Callorhinchus milii*) (http://skatebase.org/downloads). Other *Prrx* sequences were obtained from Ensembl release 73 (http://www.ensembl.org). Suppl. Tab. 1 summarized sequence information and accession numbers.

2.4 Phylogenetic analysis

Sequence alignments were performed with MUSCLE (Edgar, 2004) implemented in GENEIOUS 6.1.6 (Biomatters, Ltd.). The appropriate models for nucleotide evolution (GTR+I+G) and protein evolution (JTT+G+F) for phylogentic tree inference were determined using jModelTest 2 (Darriba et al., 2012) and ProtTest 2.4 (Abascal et al., 2005), respectively. Maximum likelihood trees were generated with PhyML 3.0 (Guindon et al., 2010) with 100 bootstrap replications and a 50% consensus rule.

2.5 Synteny analysis

The spotted gar genome assembly (*LepOcu1*) and a preliminary gene annotation were downloaded from PreEnsembl (http://pre.ensembl.org/Lepisosteus_oculatus), integrated into the Synteny Database (Catchen et al., 2009), and used to generate dot plots, orthologous pairwise, and paralogous pairwise clusters. Pairwise clusters were generated with sliding window sizes of 100 and 200 genes and paralogous pairwise clusters within the genomes of spotted gar and human were identified with amphioxus (*Branchiostoma floridae*) as outgroup (Fig. 1).

Genes directly surrounding *Prrx* genes in human, zebrafish, and stickleback were identified in Ensembl release 73; genes directly surrounding the spotted gar, Japanese eel, and Arctic lamprey *prrx* genes were predicted with AUGUSTUS (Stanke et al., 2004) and then annotated with blastp against the nr database of GenBank (http://blast.ncbi.nlm.nih.gov/Blast.cgi).

2.6 Spotted gar breeding, husbandry and ethics statement

Wild adult spotted gar were collected from the Atchafalaya River Basin, Louisiana, using electrofishing. Spotted gar broodstock were injected with Ovaprim© (0.5 ml/kg) to induce spawning and cultured in a 2-m diameter tank containing artificial spawning substrate. Mean total length/weight for female and male broodstock was 691mm/1235g and 571mm/616g, respectively. Fertilized eggs from a group spawn were shipped to the University of Oregon and arrived two days after fertilization. Embryos were manually dechorionated with forceps and raised in fish water (salinity 1ppt) at 24°C in a 14h light, 10h dark cycle. To inhibit melanin formation, 0.015 g/l N-Phenylthiourea (Sigma-Aldrich) was added to the fish water without any signs of developmental delay or malformations. Developmental stages were determined following (Long and Ballard, 2001). Animals were handled in accordance with good animal practice as approved by the University of Oregon Institutional Animal Care and Use Committee (Animal Welfare Assurance Number A-3009-01, IACUC protocol 12-02RA).

2.7 Spotted gar RNA *in situ* hybridization

RNA *in situ* hybridization probes were cloned from spotted gar genomic DNA with the following primer sets: Loc-prrx1-ex4F CTACCTACCCACCCACATGC, Loc-prrx1-3'UTRR CTTCAGATGTGCTTGGCAGAT (1,572bp); Loc-prrx2-ex4F GGCCAAGGAGTACAGTCTGC, Loc-prrx2-3'UTRR GTAAAAAGATCGGCCAGCTG (1,568bp). Spotted gar whole mount *in situ* hybridizations were performed following a standard zebrafish protocol (Jowett and Yan, 1996) with proteinase K (10µg/ml) digestion times of 35min for Long/Ballard stage (st.) 28-29 and 50min for st.33-34 55min.

2.8 Tissue expression profiling

For spotted gar, zebrafish, medaka, and European eel, expression profiles among the tissue set were generated through re-mapping of Illumina reads using BWA (Li and Durbin, 2009) and SAMtools (Li et al., 2009) implemented in Galaxy (Goecks et al., 2010). For each library, reads were re-aligned onto the coding sequence of the deduced cDNAs using BWA (maximum of two mismatches allowed, -aln 2), and counted using SAMtools (with a maximum alignment quality value –q 30, to discard ambiguous mapping reads). Read counts were subsequently normalized and expressed as reads per kb per million reads (RPKM).

3. Results

3.1 *Prrx1* and *Prrx2* are ancient vertebrate paralogs

A single *prrx* gene is present in non-vertebrate genomes such as fruitfly and the cephalochordate amphioxus (Fig. 1), but two *Prrx* genes have been identified in vertebrates (Zhong and Holland, 2011). *Prrx1* has been described in tetrapods and zebrafish, but *Prrx2* has only been found so far in tetrapods and is absent from the zebrafish genome (Hernandez-Vega and Minguillon, 2011). The genome of the African coelacanth (*Latimeria chalumnae*) (Amemiya et al., 2013) harbors both *Prrx* paralogs, suggesting that *Prrx2* could be a gene specific to the lobefin lineage.

We examined the spotted gar genome and its transcriptomes and found that both resources contained both *prrx* genes, *prrx1* and *prrx2*. The *prrx1* ortholog is located on spotted gar linkage group 10 (LG10) with extensive conserved macrosynteny to the human *PRRX1* genes on human chromosome Hsa1 (Figs. 2A, C). Likewise, the spotted gar *prrx2* ortholog on LG21 shows long-range conserved macrosynteny to the human *PRRX2* region on Hsa9 (Figs. 2B, D). These data show that the duplication of the *Prrx1/2* precursor predates the divergence of rayfin and lobefin vertebrates.

A *prrx1* gene was previously reported for small spotted catshark (Compagnucci et al., 2013). To further date the duplication of the *Prrx1/2* precursor, we looked for *Prrx* genes in the transcriptomes and genomes of three cartilaginous fish, whose lineage diverged from the bony vertebrates before the divergence of rayfin and lobefin fish, and in the genomes of two lamprey species as representatives of jawless vertebrates, the deepest diverging living vertebrates (Fig. 1). Both *Prrx* paralogs were found for small spotted catshark, little skate, and elephant shark showing that the *Prrx* duplication predates the origin of extant jawed (gnathostome) vertebrates. The situation in lampreys, however, is less conclusive: While no *prrx* gene could be located in the genome assembly of the sea lamprey (*Petromyzon marinus*), the genome of the Arctic lamprey possesses a clear *prrx* gene. Thus, based on the *prrx* gene repertoire in agnathans, it remains unclear whether the duplication of *Prrx*

predates the origin of all living vertebrate lineages or whether this was a gnathostome-specific duplication (but see section 3.4 below).

3.2 Evolution of *prrx1* genes in teleost model species

Next, we analyzed the situation for *prrx* genes in Clupeocephala, the teleost order containing the major fish model species zebrafish, medaka, platyfish, and stickleback (Fig. 1). Two paralogs of *prrx1*, *prrx1a* and *prrx1b*, are present in the genome of zebrafish (Hernandez-Vega and Minguillon, 2011), an ostariophysian teleost (Fig. 1). The genome of another ostariophysian, the Mexican tetra (*Astyanax mexicanus*), harbors *prrx1a* and *prrx1b* as well. In contrast, the genomes of the acanthomorph teleosts – medaka, platyfish (*Xiphophorus maculatus*), Nile tilapia (*Oreochromis niloticus*), three-spined stickleback (*Gasterosteus aculeatus*), two pufferfishes (both *Tetraodon nigroviridis* and *Takifugu rubripes*), and Atlantic cod (*Gadus morhua*) – contain only a single *prrx1* gene. Therefore, *prrx1a* and *prrx1b* may be either paralogs from an ostariophysian-specific gene duplication or older duplicates generated, for example, during the TGD with a subsequent loss of one of the two genes in the lineage leading to acanthomorphs.

Synteny data support the origin-in-TGD hypothesis: the *prrx1* region on spotted gar LG10 shows double conserved synteny to zebrafish chromosomes Dre2 and Dre20, which contain the *prrx1a* and *prrx1b* genes, respectively (Fig. 3A). The synteny block on Dre20 is relatively short as some LG10 co-orthlogs jump to Dre8 and Dre6. Furthermore, the *prrx1* region in spotted gar LG10 shows double conserved synteny to medaka chromosome Ola4, containing the single *prrx1* gene, and to Ola17, which does not harbor a *prrx* gene (Fig. 3B). Finally, the *prrx1b* region on Dre20 shows conserved synteny with Ola4 but also Ola24 (Suppl. Fig. 1A), while the *prrx1a* region on Dre2 shows extensive conserved synteny to Ola17 (Suppl. Fig. 1B). Dre2, as well as Ola17 and Ola4 are derived from the reconstructed ancestral pre-TGD protochromosome *m* (Kasahara et al., 2007; Nakatani et al., 2007). Thus, we conclude *(i)* that the zebrafish *prrx1* co-orthologs were generated in the TGD; *(ii)* that in the zebrafish lineage the chromosomal region surrounding the *prrx1b* gene was translocated to a different chromosome that became Dre20 after the TGD; and

(iii) that the *prrx1a* co-ortholog was lost in the lineage leading to acanthomorphs, leaving this teleost group with only the *prrx1b* co-ortholog.

When did the *prrx1a* gene get lost during the course of teleost evolution? To answer this question, we surveyed the transcriptomes of several phylogenetic informative taxa (Fig. 1). Transcripts of both genes, *prrx1a* and *prrx1b*, were found for the striped catfish, another ostariophysian, for the allis shad, a clupeomorph, and for the Northern pike, an esociform that represents a sister lineage to acanthomorph teleosts (Near et al., 2012). These data would be expected if *prrx1a* was lost in an ancestor of acanthomorph teleosts.

3.3 Three *prrx* genes in basal teleosts and the problem of orthology

Until recently, historical relationsships of the three major teleost lineages, Clupeocephala (including zebrafish, medaka, stickleback, and others), Osteoglossomorpha (bony tongues and mooneyes), and Elopomorpha (eels, tarpons, ladyfishes, bonefishes) were unclear. Fresh molecular phylogenetic analyses now place elopomorphs basally in the phylogeny (Fig. 1) as a sister group to the lineage of osteoglossomorphs and clupeocephalans, which together constitute the monophylum Osteoglossocephala (Betancur et al., 2013; Faircloth et al., 2013; Near et al., 2012). We were interested in the status of *prrx* genes in elopomorphs and osteoglossomorphs and therefore analyzed the genomes of two eels species, the European eel (Henkel et al., 2012a) and the Japanese eel (Henkel et al., 2012b); we also searched our transcriptomes from the European eel and the butterflyfish, an osteoglossomorph.

Surprisingly, and in contrast to clupeocephalans, we found a *prrx2* gene for all three basally diverging teleosts. Therefore, *prrx2* was still present in the pre-TGD teleost ancestor and after the TGD, one of the two *prrx2* TGD duplicates was lost in all teleost, but the other TGD duplicate of them was retained both in elopomorphs and in osteoglossomorphs, but was lost in an ancestor of the clupeocephalan lineage.

Similar to zebrafish, eels and butterflyfish both have two *prrx1* genes. Thus, the lineages of elopomorphs and osteoglossomorphs contain three *prrx* genes, more than any other vertebrate so far investigated.

But which of the *prrx1* co-orthlogs in eels and butterflyfish are orthologous to *prrx1a* or *prrx1b* in zebrafish? To answer this question, we generated multiple phylogenetic trees of vertebrate *Prrx* genes based both on nucleotide and protein sequences. A bootstrap consensus (50%) maximum likelihood phylogeny of the nucleotide phylogeny is shown in Fig. 3C, other trees appear in Suppl. Fig. 2.

Common to all inferred phylogenies is that *Prrx1* genes are clearly separated from *Prrx2*. In the *Prrx2* clade, coelacanth groups anomalously with rayfin fish rather than lobefins. In contrast to *Prrx2*, however, in the *Prrx1* clade the vast majority of nodes are not well supported because the relatively short *Prrx1* genes (738bp) and accordingly Prrx1 proteins (246aa) lack phylogenetic signal. It is thus impossible to determine orthology of *prrx1* duplicates within teleosts based on phylogenetic trees.

Using the draft genome assemblies of eel species, we were able to assign orthology of the eel genes based on conserved microsynteny for genes directly surrounding the *prrx* genes (Fig. 4). In Japanese eel, scaffold222 appears to be more conserved with the *prrx1b* neighborhood in clupeocephalans: 6/9 eel genes from scaffold222 contribute to conserved synteny with zebrafish chromosome Dre20 (including *slc25a24*, which is specific to the zebrafish *prrx1b* region) and stickleback chromosome GacVIII, compared to 2/9 genes shared with the *prrx1a* neighborhoods on Dre2 and GacIII. The genomic region on Japanese eel scaffold1220 on the other hand is less decisive: 3/5 eel genes contribute to conserved synteny with both *prrx1* paralogons in zebrafish, yet eel scaffold1220 contains *ivns1ap*, which is in the vicinity of the *prrx1a* paralogon on Dre2. A genomic assembly of butterflyfish is currently lacking precluding a similar analysis for this osteoglossomorph.

Alternative splicing patterns can provide phylogenetically informative characters independent of sequence similarities and the preservation of genome neighborhoods. In mammals, *Prrx1* gene generates two major splice isoforms (Norris and Kern, 2001), hereafter called splisoform A and splisoform B. While

Prrx1 splisoform A contains both the homeodomain and the C-terminal OAR protein domain similar to the single *Prrx2* splisoform, the shorter *Prrx1* splisoform B uses an alternative fourth exon lacking the OAR domain (Norris and Kern, 2001) (Fig. 5). Likewise, in the spotted gar transcriptome two splice variants were present for *prrx1*, splisoform A and splisoform B', with B' meaning that, although this gene uses an alternative fourth exon, this exon is not conserved with the mammalian alternative fourth exon. The resulting protein in spotted gar lacks the OAR domain and thus appears to be functionally equivalent to the mammalian splisoform B, hence the name splisoform B'. In zebrafish, only *prrx1b* (and likewise the acanthomorph *prrx1b* genes) generate splisoform B'. The alternative fourth exon is gone from zebrafish *prrx1a* (Fig. 5). In the eel genomes and supported by the eel transcriptome, for one of the two *prrx1* co-orthologs we also found evidence for the presence of the alternative fourth exon that characterizes splisoform B' in spotted gar and zebrafish. This eel co-ortholog (scaffold222 in Japanese eel) is thus annotated as *prrx1b*, which is consistent with our conserved synteny results (Figs. 4/5). In the butterflyfish transcriptome, however, we did not find evidence for an isoform B' for either *prrx1* genes; therefore, the orthology of butterflyfish *prrx1* genes remains unresolved and so these genes were called *prrx1x* and *prrx1y* following nomenclature conventions for unclear phylogenetic relationships following genome duplications (Force et al., 2002).

3.4 Origin of *Prrx1* and *Prrx2* in the early vertebrate genome duplications

We next analyzed the origin of gnathostome *Prrx1* and *Prrx2* genes. As we have seen, Arctic lamprey contains only a single *prrx* gene (see section 3.1) and phylogenetic trees tend to place lamprey *prrx* as outgroup to the gnathostome genes (Fig. 3C, Suppl. Fig. 2). The direct genomic environment of lamprey *prrx*, however, appears to be more similar to gnathostome *Prrx1* than to *Prrx2*. Among the ten genes on lamprey scaffold416 surrounding *prrx*, three genes (*kifap3*, *mettl11b*, and *pde4dip*) are also found in the neighborhood of gnathostome *Prrx1* genes, but none of the genes on scaffold416 provides conserved synteny with gnathostome *Prrx2* (Fig. 4). In addition, sequence comparisons uncovered no evidence for an alternative

fourth exon could. From these data, it is thus difficult to conclude whether the duplication of *Prrx* genes occurred before or after the divergence on agnathans and gnathostomes.

As with the TGD, intragenomic patterns of conserved synteny can be informative to infer gene family history with regard to the vertebrate whole genome duplications (*e.g.* Braasch et al., 2009; Canestro et al., 2009; Lagman et al., 2013; Nakatani et al., 2007). We therefore used the Synteny Database (Catchen et al., 2009) to generate pairwise paralogous clusters and paralogy dotplots for spotted gar and human with amphioxus as outgroup (Fig. 6). Within spotted gar and human, the two *Prrx* gene regions show extensive pairwise paralogy with each other (Fig. 6A, B) suggesting that the *Prrx*-containing paralogons were generated in a large-scale duplication event. Dotplots further provide evidence that the two *Prrx* paralogons found in spotted gar were indeed generated during the earlier rounds of vertebrate whole genome duplication: In spotted gar, the two *prrx* paralogons on LG10 and LG21 also show conserved synteny with LG2, LG6, and LG19 (Fig. 6C); in human, the two *PRRX* paralogons on Hsa1 and Hsa9 share additional conserved synteny with regions on Hsa6 and Hsa19 (Fig. 6D). Similar observations were made for the chicken genome (data not shown). The paralogous regions evident in the human and chicken dotplots were previously shown to be derived from the inferred vertebrate ancestral protochromosome *A* (Nakatani et al., 2007). Thus we conclude that the gnathostome *Prrx* genes were generated as result of the early vertebrate genome duplications VGD1 and VGD2.

3.5 Expression of *prrx* ohnologs in spotted gar and teleosts

To gain insight into the expression of *prrx* genes in the spotted gar, we performed RNA whole mount *in situ* hybridization experiments on spotted gar embryos (Fig. 7A). Based on the prominent expression of *Prrx1* and *Prrx2* in the fore- and hindlimb buds of mouse and chicken (Beverdam and Meijlink, 2001; Kuratani et al., 1994; Lu et al., 1999; Nohno et al., 1993) and of *prrx1* co-orthologs in pectoral fin buds of zebrafish (Hernandez-Vega and Minguillon, 2011), we selected two critical developmental stages for our analysis. At Long/Ballard stage 28-29 (5-6

days post fertilization), the pectoral fins are visible as low discs; at stage 32-33 (11-12 days post fertilization), pelvic fin buds become apparent for the first time (Long and Ballard, 2001). Indeed, we find that *prrx1* and *prrx2* are both expressed in the pectoral fins buds at stage 28-29. Furthermore, both genes are expressed in branchial arches and head structures at stage 28-29. At stage 32-33, *prrx1* and *prrx2* are both expressed in the pelvic fin buds while the expression in pectoral find is no longer detectable (Fig. 7A).

Using our RNA-Seq data we furthermore determined expression levels of *pprx* genes in adult tissues of spotted gar, European eel, zebrafish, and medaka (Fig. 7B). A general observation is that in all species the level of *prrx* gene expression is higher in whole embryos than in any adult tissue. Furthermore, in the two species that have a *prrx2* gene, spotted gar and European eel, the expression of *prrx2* is lower than the expression of *prrx1* (spotted gar) or *prrx1* co-orthologs combined (European eel).

In spotted gar, *prrx1* is most prominently expressed in testis and bone, while *prrx2* is barely expressed in these tissues. Gar *prrx1* is also expressed in gills, heart, muscle and kidney, with *prrx2* expression being equally high in the heart (Fig. 7B).

In the eel, *prrx1a* is generally more highly expressed than *prrx1b*. While *prrx1a* is expressed in bone, testis, gills, brain, and muscle, *prrx1b* is only expressed in gills (Suppl. Fig. 3). Eel *prrx2* expression is found mostly in the gills, in contrast to spotted gar *prrx2* (Fig. 7B).

In zebrafish as in eel, *prrx1a* is more highly expressed than *prrx1b*. Zebrafish *prrx1a* is expressed in gills, muscle, brain, bone, testis, and kidney, and *prrx1b* is expressed in testis, gills, and muscle (Fig. 7B, Suppl. Fig.3). When comparing *prrx1* co-ortholog expression between zebrafish and eel (Suppl. Fig. 3), we see that despite differences in absolute expression levels (which are generally higher in zebrafish), the ratio of expression of *prrx1a vs. prrx1b* is similar in both species.

In medaka, *prrx1b* is expressed in bone, muscle, and gills and thereby more closely resembles the expression of *prrx1a* in zebrafish and eel than the expression of *prrx1b* in these species (Fig. 7B).

4. Discussion

4.1 Evolution of vertebrate *Prrx* genes by three rounds of genome duplication

The analyses presented here provide evidence to support a model for the evolution of the *Prrx* gene family in vertebrates through three rounds of whole genome duplication (VGD1, VGD2, TGD) (Fig. 8). The two vertebrate genome duplications led to the amplification of many gene families in vertebrates, generating for example the four vertebrate *Hox* clusters (reviewed in (Canestro, 2012)). In the case of the *prrx* genes, gnathostomes have retained only two of the theoretically four VGD ohnologs. This result could be explained by the hypothesis that one of the two VGD1-paralogs was lost before VGD2 so that only a single *Prrx* gene was available for duplication in the VGD2 event. The genomic location of the two *PRRX* genes in human and chicken relative to the reconstruction of vertebrate karyotype evolution by Nakatani et al. (2007), however, is inconsistent with this hypothesis. The *PRRX* paralogons in human and chicken genomes are located in chromosomal blocks *A2* (*PRRX1*) and *A0* (*PRRX2*), which are connected by VGD1, not VGD2 (Nakatani et al., 2007). Thus, it appears that after duplication of the ancestral chromosome *A* during VGD1 and VGD2 four *Prrx* ohnologs were initially present but the VGD2 ohnologs of the two extant *Prrx* genes were lost independently after VGD2 (Fig. 8). The presence of a single *prrx* gene in lamprey, whose lineage most likely diverged from gnathostomes after VGD2 (Smith et al., 2013), neither supports nor refutes this model. The phylogenetic position of the lamprey *prrx* gene remains unresolved, a common problem when inferring gene family histories among the two major, early diverging vertebrate lineages, agnathans and gnathostomes (Kuraku, 2013).

The common ancestor of bony vertebrates (euteleostomes) still contained both *Prrx* genes because we find both genes in many lobefins as well as in spotted gar among rayfins. Our analysis of teleost transcriptomes further shows that *prrx2* is not an ohnolog gone missing in all teleosts, but that it was lost in the lineage leading to clupeocephalan teleosts, and was retained in the two other major teleost lineages, elopomorphs and osteoglossomorphs. Following the TGD, *prrx1* was still present in

two copies in the ancestors of the elopomorph, osteoglossomorph, and clupeocephalan lineages. Within clupeocephalans, the *prrx1a* gene was later lost in the lineage leading to acanthomorphs.

Based on the rapid succession of the TGD and the diversification of the three major teleost lineages (elopomorphs, osteoglossomorphs, and clupeocephalans), phylogenetic reconstructions can fail in clearly inferring orthology of TGD duplicates among the three teleost lineages (Kuraku, 2013), particularly in situations with low phylogenetic signal like the one described here for *prrx* genes. Therefore, other measures, such as conserved synteny and splicing patterns, become essential to determine orthology as successfully applied here for clupeocephalan and eel *prrx1* genes.

Our analysis of *Prrx* gene family history shows that none of the teleost model species resembles the situation of *PRRX* genes in human. The makeup of the *prrx* gene family in neither zebrafish nor medaka is representative of teleosts in general, and ironically, despite the TGD, the *prrx* gene repertoire is most reduced among bony vertebrates in acanthomorphs, including model species like medaka and platyfish. The basal teleost lineages, in contrast, have the most expanded *prrx* gene repertoire of all vertebrates analyzed so far. Among investigated rayfins, the spotted gar is the only one that is the same as human with regard to its *prrx* gene repertoire.

This census of rayfin *prrx* genes illustrates that the choice of model species for the study of gene functions can be rather random at the level of the specific gene family and that the specifics of the gene regulatory network components in the teleost model under investigation should be taken into account when generalizing to teleosts and translating the knowledge to the human situation.

4.2 Implications of gene family history for gene function evolution and modeling human conditions

Prrx genes are most prominently known for their function during the development of limbs and craniofacial structures (ten Berge et al., 1998). Our *in situ* hybridization analysis during spotted gar development shows for the first time that the overlapping expression of *Prrx1* and *Prrx2* genes in paired appendages and

branchial arches is not restricted to tetrapods, but is most likely derived from a gene regulatory program that predates the divergence of bony vertebrates. Besides reports on the expression of *prrx1* in the branchial arches in shark (Compagnucci et al., 2013), expression data are currently lacking for species basal to bony vertebrates, but will be required to gain more insight into the evolutionary origin of vertebrate *Prrx* gene expression patterns.

Following the parsimony principle, the similarity in expression of bony vertebrate *Prrx1* and *Prrx2* suggests that their common single precursor gene potentially already had *cis*-regulatory elements enabling expression in branchial arches and paired appendages that both daughter genes then retained after the duplication. Genomic evidence shows that the VGD1 event initially generated *Prrx1* and *Prrx2* (Fig. 8), an event that predates the divergence of gnathostome and agnathan vertebrates (Smith et al., 2013). These arguments lead to the paradoxical situation that the paired appendage expression domain evolved before the evolution of paired appendages in gnathostomes (Tulenko et al., 2013)! This conclusion is reminiscent of the situation for *Tbx5* and *Tbx4*, which are expressed in gnathostomes in anterior and posterior paired appendages, respectively, suggesting that their expression program is derived from the common *Tbx4/5* precursor gene, and yet the duplication that generated these paralogs presumably predates the origin of paired appendages itself (Horton et al., 2008; Minguillon et al., 2009; Tanaka et al., 2002). Considering both gene families (*Prrx* and *Tbx4/5*), we conclude that the evolution of paired appendages may involve the exaptation of preexisting regulatory elements driving the expression of limb-specific genes in paired mesodermal domains.

The present study furthermore shows that the spotted gar will be very useful to functionally model the role of *prrx1* and *prrx2* in the ancestor of tetrapods before the fin-to-limb transition. Lobefin fish, *i.e.* coelacanths and lungfishes, are not suitable for functional and/or genetic studies, while teleost model species are variants as proxies due to the loss of *prrx2* and the independent duplication of *prrx1* during the TGD.

The expression of the two *prrx1* co-orthologs during zebrafish development largely overlapps in pectoral fin buds, head mesenchyme, and branchial arches (Hernandez-Vega and Minguillon, 2011) and matches similar expression of the single *prrx1* gene in spotted gar embryos (Fig. 7A). The expression of *prrx1* in embryonic and adult spotted gar, in comparison to zebrafish and eel *prrx1* co-orthologs, does not clearly indicate of sub- and/or neofunctionalization after the TGD. Generally, in both zebrafish and eel, *prrx1a* is the more highly expressed TGD co-ortholog, and thus presumably already was more highly expressed than *prrx1b* in the last common ancestor of elopomorphs and clupeocephalans. Surprisingly, however, within clupeocephalans, *prrx1a,* the highly expressed gene, rather than *prrx1b*, the lower expressed gene, was lost in acanthomorphs. We hypothesize that *prrx1a* was more prone to gene loss than *prrx1b* despite its presumable higher and broader expression because it had lost its ability to generate splisoform B' soon after the TGD.

Our RNA-Seq analysis of adult rayfin tissues revealed that *prrx* genes are predominantly expressed in bone, gills, heart, muscle, brain, and testis. *Prrx* genes are expressed in bone, heart, and muscle in human and mouse as well (Leussink et al., 1995; Norris et al., 2000) suggesting similar functions in rayfins and mammals. *Prrx1* has been recently shown to be important for the maintenance of adult neural stem cells in mammals (Shimozaki et al., 2013). Zebrafish is becoming an important model organism for adult neurogenesis in vertebrates (see *e.g.* (Grandel et al., 2006)) and our expression data suggest that *prrx1a* is a primary candidate to test for similar functions in teleosts. In medaka, in contrast, we did not find expression of its only remaining gene, *prrx1b*, in the adult brain (Fig. 7B) suggesting that zebrafish potentially better than medaka models adult neurogenic functions of *PRRX1* in human.

Interestingly, testis had the highest expression of *prrx1* among adult tissues in spotted gar, and testis expression of *prrx1a* as well as of *prrx1a* and *prrx1b* was also found in eel and zebrafish, respectively. To the best of our knowledge, *Prrx1* has not been studied in the context of testis development and testis function and it will be

interesting to analyze if this is a rayfin-specific function or shared with other vertebrates.

5. Conclusions

The present study illustrates, based on the example of *Prrx* genes, the diversity of the developmental genetic toolbox components among vertebrates and within teleost fishes in particular. These differences are the results of lineage-specific gene family expansions by genome duplications and multiple subsequent gene loss events. We also uncovered significant differences in the *prrx* gene repertoires among teleost model and non-model species and learned that no single species represents teleosts in general, emphasizing the need to study several different teleost models.

Our study highlights the suitability of spotted gar to infer gene family histories in vertebrates. The spotted gar lends itself as the best possible model for the situation of the *Prrx* gene family before the fin-to-limb transition and to study the function of *prrx2* in a fish. Future biomedical studies concerned with *Prrx* genes in teleost model species need to take into account differences with respect to the human gene family. The currently rapid progress in fish genomics illuminates the 'genomic black boxes' of teleosts and enables us to make better predictions about which particular fish model to use for modeling specific human gene functions, thereby improving the interpretation and translation of insights from fish models to human disease states.

Acknowledgements

We would like to thank the BROAD Institute for generating the spotted gar genome assembly, MGX-Montpellier GenomiX for performing RNA-seq, INRA-Sigenae for generating *de novo* cDNA assemblies, Yi-Lin Yan for advice on spotted gar *in situ* hybridizations, Julian Catchen for help with integrating the spotted gar into the Synteny Database, and the Postlethwait lab, Julia Ganz, and Dylan Farnsworth for stimulating discussions. This work was supported by NIH grant R01 RR020833 (R01 OD011116) to JHP and by French National Research Agency grant PHYLOFISH (ANR-10-GENM-017) to JB.

References

Abascal, F., Zardoya, R., Posada, D., 2005. ProtTest: selection of best-fit models of protein evolution. Bioinformatics 21, 2104-2105.

Amemiya, C.T., Alfoldi, J., Lee, A.P., Fan, S., Philippe, H., Maccallum, I., Braasch, I., Manousaki, T., Schneider, I., Rohner, N., Organ, C., Chalopin, D., Smith, J.J., Robinson, M., Dorrington, R.A., Gerdol, M., Aken, B., Biscotti, M.A., Barucca, M., Baurain, D., Berlin, A.M., Blatch, G.L., Buonocore, F., Burmester, T., Campbell, M.S., Canapa, A., Cannon, J.P., Christoffels, A., De Moro, G., Edkins, A.L., Fan, L., Fausto, A.M., Feiner, N., Forconi, M., Gamieldien, J., Gnerre, S., Gnirke, A., Goldstone, J.V., Haerty, W., Hahn, M.E., Hesse, U., Hoffmann, S., Johnson, J., Karchner, S.I., Kuraku, S., Lara, M., Levin, J.Z., Litman, G.W., Mauceli, E., Miyake, T., Mueller, M.G., Nelson, D.R., Nitsche, A., Olmo, E., Ota, T., Pallavicini, A., Panji, S., Picone, B., Ponting, C.P., Prohaska, S.J., Przybylski, D., Saha, N.R., Ravi, V., Ribeiro, F.J., Sauka-Spengler, T., Scapigliati, G., Searle, S.M., Sharpe, T., Simakov, O., Stadler, P.F., Stegeman, J.J., Sumiyama, K., Tabbaa, D., Tafer, H., Turner-Maier, J., van Heusden, P., White, S., Williams, L., Yandell, M., Brinkmann, H., Volff, J.N., Tabin, C.J., Shubin, N., Schartl, M., Jaffe, D.B., Postlethwait, J.H., Venkatesh, B., Di Palma, F., Lander, E.S., Meyer, A., Lindblad-Toh, K., 2013. The African coelacanth genome provides insights into tetrapod evolution. Nature 496, 311-316.

Amores, A., Catchen, J., Ferrara, A., Fontenot, Q., Postlethwait, J.H., 2011. Genome evolution and meiotic maps by massively parallel DNA sequencing: spotted gar, an outgroup for the teleost genome duplication. Genetics 188, 799-808.

Aparicio, S., Chapman, J., Stupka, E., Putnam, N., Chia, J.M., Dehal, P., Christoffels, A., Rash, S., Hoon, S., Smit, A., Gelpke, M.D., Roach, J., Oh, T., Ho, I.Y., Wong, M., Detter, C., Verhoef, F., Predki, P., Tay, A., Lucas, S., Richardson, P., Smith, S.F., Clark, M.S., Edwards, Y.J., Doggett, N., Zharkikh, A., Tavtigian, S.V., Pruss, D., Barnstead, M.,

Evans, C., Baden, H., Powell, J., Glusman, G., Rowen, L., Hood, L., Tan, Y.H., Elgar, G., Hawkins, T., Venkatesh, B., Rokhsar, D., Brenner, S., 2002. Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. Science 297, 1301-1310.
Bergwerff, M., Gittenberger-de Groot, A.C., Wisse, L.J., DeRuiter, M.C., Wessels, A., Martin, J.F., Olson, E.N., Kern, M.J., 2000. Loss of function of the Prx1 and Prx2 homeobox genes alters architecture of the great elastic arteries and ductus arteriosus. Virchows Arch 436, 12-19.
Betancur, R.R., Broughton, R.E., Wiley, E.O., Carpenter, K., Lopez, J.A., Li, C., Holcroft, N.I., Arcila, D., Sanciangco, M., Cureton Ii, J.C., Zhang, F., Buser, T., Campbell, M.A., Ballesteros, J.A., Roa-Varon, A., Willis, S., Borden, W.C., Rowley, T., Reneau, P.C., Hough, D.J., Lu, G., Grande, T., Arratia, G., Orti, G., 2013. The tree of life and a new classification of bony fishes. PLoS Curr 5.
Beverdam, A., Meijlink, F., 2001. Expression patterns of group-I aristaless-related genes during craniofacial and limb development. Mech Dev 107, 163-167.
Braasch, I., Postlethwait, J.H., 2012. Polyploidy in Fish and the Teleost Genome Duplication, in: P.S. Soltis, D.E. Soltis (Eds.), Polyploidy and Genome Evolution. Springer, Berlin Heidelberg, 341-383.
Braasch, I., Volff, J.N., Schartl, M., 2009. The endothelin system: evolution of vertebrate-specific ligand-receptor interactions by three rounds of genome duplication. Mol Biol Evol 26, 783-799.
Canestro, C., 2012. Two round of Whole-Genome Duplication: Evidence and Impact on the Evoluion of Vertebrate Innovations, in: P.S. Soltis, D.E. Soltis (Eds.), Polyploidy and Genome Evolution. Springer, Berlin Heidelberg, 309-339.
Canestro, C., Catchen, J.M., Rodriguez-Mari, A., Yokoi, H., Postlethwait, J.H., 2009. Consequences of lineage-specific gene loss on functional evolution of surviving paralogs: ALDH1A and retinoic acid signaling in vertebrate genomes. PLoS Genet 5, e1000496.
Catchen, J.M., Conery, J.S., Postlethwait, J.H., 2009. Automated identification of conserved synteny after whole-genome duplication. Genome Res 19, 1497-1505.
Compagnucci, C., Debiais-Thibaud, M., Coolen, M., Fish, J., Griffin, J.N., Bertocchini, F., Minoux, M., Rijli, F.M., Borday-Birraux, V., Casane, D., Mazan, S., Depew, M.J., 2013. Pattern and polarity in the development and evolution of the gnathostome jaw: both conservation and heterotopy in the branchial arches of the shark, Scyliorhinus canicula. Dev Biol 377, 428-448.
Cserjesi, P., Lilly, B., Bryson, L., Wang, Y., Sassoon, D.A., Olson, E.N., 1992. MHox: a mesodermally restricted homeodomain protein that binds an essential site in the muscle creatine kinase enhancer. Development 115, 1087-1101.
Darriba, D., Taboada, G.L., Doallo, R., Posada, D., 2012. jModelTest 2: more models, new heuristics and parallel computing. Nat Methods 9, 772.
Dehal, P., Boore, J.L., 2005. Two rounds of whole genome duplication in the ancestral vertebrate. PLoS Biol 3, e314.
Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32, 1792-1797.
Faircloth, B.C., Sorenson, L., Santini, F., Alfaro, M.E., 2013. A Phylogenomic Perspective on the Radiation of Ray-Finned Fishes Based upon Targeted Sequencing of Ultraconserved Elements (UCEs). PLoS One 8, e65923.

Force, A., Amores, A., Postlethwait, J.H., 2002. Hox cluster organization in the jawless vertebrate Petromyzon marinus. J Exp Zool 294, 30-46.
Force, A., Lynch, M., Pickett, F.B., Amores, A., Yan, Y.L., Postlethwait, J., 1999. Preservation of duplicate genes by complementary, degenerative mutations. Genetics 151, 1531-1545.
Frankenberg, S., Renfree, M.B., 2013. On the origin of POU5F1. BMC Biol 11, 56.
Goecks, J., Nekrutenko, A., Taylor, J., 2010. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol 11, R86.
Grandel, H., Kaslin, J., Ganz, J., Wenzel, I., Brand, M., 2006. Neural stem cells and neurogenesis in the adult zebrafish brain: origin, proliferation dynamics, migration and cell fate. Dev Biol 295, 263-277.
Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W., Gascuel, O., 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol 59, 307-321.
Hedges, S.B., Dudley, J., Kumar, S., 2006. TimeTree: a public knowledge-base of divergence times among organisms. Bioinformatics 22, 2971-2972.
Henkel, C.V., Burgerhout, E., de Wijze, D.L., Dirks, R.P., Minegishi, Y., Jansen, H.J., Spaink, H.P., Dufour, S., Weltzien, F.A., Tsukamoto, K., van den Thillart, G.E., 2012a. Primitive duplicate Hox clusters in the European eel's genome. PLoS One 7, e32231.
Henkel, C.V., Dirks, R.P., de Wijze, D.L., Minegishi, Y., Aoyama, J., Jansen, H.J., Turner, B., Knudsen, B., Bundgaard, M., Hvam, K.L., Boetzer, M., Pirovano, W., Weltzien, F.A., Dufour, S., Tsukamoto, K., Spaink, H.P., van den Thillart, G.E., 2012b. First draft genome sequence of the Japanese eel, Anguilla japonica. Gene 511, 195-201.
Hernandez-Vega, A., Minguillon, C., 2011. The Prx1 limb enhancers: targeted gene expression in developing zebrafish pectoral fins. Dev Dyn 240, 1977-1988.
Horton, A.C., Mahadevan, N.R., Minguillon, C., Osoegawa, K., Rokhsar, D.S., Ruvinsky, I., de Jong, P.J., Logan, M.P., Gibson-Brown, J.J., 2008. Conservation of linkage and evolution of developmental function within the Tbx2/3/4/5 subfamily of T-box genes: implications for the origin of vertebrate limbs. Dev Genes Evol 218, 613-628.
Howe, K., Clark, M.D., Torroja, C.F., Torrance, J., Berthelot, C., Muffato, M., Collins, J.E., Humphray, S., McLaren, K., Matthews, L., McLaren, S., Sealy, I., Caccamo, M., Churcher, C., Scott, C., Barrett, J.C., Koch, R., Rauch, G.J., White, S., Chow, W., Kilian, B., Quintais, L.T., Guerra-Assuncao, J.A., Zhou, Y., Gu, Y., Yen, J., Vogel, J.H., Eyre, T., Redmond, S., Banerjee, R., Chi, J., Fu, B., Langley, E., Maguire, S.F., Laird, G.K., Lloyd, D., Kenyon, E., Donaldson, S., Sehra, H., Almeida-King, J., Loveland, J., Trevanion, S., Jones, M., Quail, M., Willey, D., Hunt, A., Burton, J., Sims, S., McLay, K., Plumb, B., Davis, J., Clee, C., Oliver, K., Clark, R., Riddle, C., Eliott, D., Threadgold, G., Harden, G., Ware, D., Mortimer, B., Kerry, G., Heath, P., Phillimore, B., Tracey, A., Corby, N., Dunn, M., Johnson, C., Wood, J., Clark, S., Pelan, S., Griffiths, G., Smith, M., Glithero, R., Howden, P., Barker, N., Stevens, C., Harley, J., Holt, K., Panagiotidis, G., Lovell, J., Beasley, H., Henderson, C., Gordon, D., Auger, K., Wright, D., Collins, J., Raisen, C., Dyer, L., Leung, K., Robertson, L., Ambridge, K., Leongamornlert, D., McGuire, S., Gilderthorp, R., Griffiths, C., Manthravadi, D., Nichol, S., Barker, G., Whitehead, S., Kay, M., Brown, J., Murnane, C., Gray, E., Humphries, M., Sycamore, N., Barker, D., Saunders, D., Wallis, J., Babbage, A., Hammond, S., Mashreghi-Mohammadi, M., Barr, L., Martin, S., Wray, P.,

Ellington, A., Matthews, N., Ellwood, M., Woodmansey, R., Clark, G., Cooper, J., Tromans, A., Grafham, D., Skuce, C., Pandian, R., Andrews, R., Harrison, E., Kimberley, A., Garnett, J., Fosker, N., Hall, R., Garner, P., Kelly, D., Bird, C., Palmer, S., Gehring, I., Berger, A., Dooley, C.M., Ersan-Urun, Z., Eser, C., Geiger, H., Geisler, M., Karotki, L., Kirn, A., Konantz, J., Konantz, M., Oberlander, M., Rudolph-Geiger, S., Teucke, M., Osoegawa, K., Zhu, B., Rapp, A., Widaa, S., Langford, C., Yang, F., Carter, N.P., Harrow, J., Ning, Z., Herrero, J., Searle, S.M., Enright, A., Geisler, R., Plasterk, R.H., Lee, C., Westerfield, M., de Jong, P.J., Zon, L.I., Postlethwait, J.H., Nusslein-Volhard, C., Hubbard, T.J., Roest Crollius, H., Rogers, J., Stemple, D.L., 2013. The zebrafish reference genome sequence and its relationship to the human genome. Nature 496, 498-503.

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

Jones, F.C., Grabherr, M.G., Chan, Y.F., Russell, P., Mauceli, E., Johnson, J., Swofford, R., Pirun, M., Zody, M.C., White, S., Birney, E., Searle, S., Schmutz, J., Grimwood, J., Dickson, M.C., Myers, R.M., Miller, C.T., Summers, B.R., Knecht, A.K., Brady, S.D., Zhang, H., Pollen, A.A., Howes, T., Amemiya, C., Baldwin, J., Bloom, T., Jaffe, D.B., Nicol, R., Wilkinson, J., Lander, E.S., Di Palma, F., Lindblad-Toh, K., Kingsley, D.M., 2012. The genomic basis of adaptive evolution in threespine sticklebacks. Nature 484, 55-61.

Jowett, T., Yan, Y.L., 1996. Double fluorescent in situ hybridization to zebrafish embryos. Trends Genet 12, 387-389.

Kasahara, M., Naruse, K., Sasaki, S., Nakatani, Y., Qu, W., Ahsan, B., Yamada, T., Nagayasu, Y., Doi, K., Kasai, Y., Jindo, T., Kobayashi, D., Shimada, A., Toyoda, A., Kuroki, Y., Fujiyama, A., Sasaki, T., Shimizu, A., Asakawa, S., Shimizu, N., Hashimoto, S., Yang, J., Lee, Y., Matsushima, K., Sugano, S., Sakaizumi, M., Narita, T., Ohishi, K., Haga, S., Ohta, F., Nomoto, H., Nogata, K., Morishita, T., Endo, T., Shin, I.T., Takeda, H., Morishita, S., Kohara, Y., 2007. The medaka draft genome and insights into vertebrate genome evolution. Nature 447, 714-719.

Kern, M.J., Witte, D.P., Valerius, M.T., Aronow, B.J., Potter, S.S., 1992. A novel murine homeobox gene isolated by a tissue specific PCR cloning strategy. Nucleic Acids Res 20, 5189-5195.

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

Kuratani, S., Martin, J.F., Wawersik, S., Lilly, B., Eichele, G., Olson, E.N., 1994. The expression pattern of the chick homeobox gene gMHox suggests a role in patterning of the limbs and face and in compartmentalization of somites. Dev Biol 161, 357-369.

Lagman, D., Ocampo Daza, D., Widmark, J., Abalo, X.M., Sundstrom, G., Larhammar, D., 2013. The vertebrate ancestral repertoire of visual opsins, transducin alpha subunits and oxytocin/vasopressin receptors was established by duplication of their shared genomic region in the two rounds of early vertebrate genome duplications. BMC Evol Biol 13, 238.
Leussink, B., Brouwer, A., el Khattabi, M., Poelmann, R.E., Gittenberger-de Groot, A.C., Meijlink, F., 1995. Expression patterns of the paired-related homeobox genes MHox/Prx1 and S8/Prx2 suggest roles in development of the heart and the forebrain. Mech Dev 52, 51-64.
Li, H., Durbin, R., 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754-1760.
Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078-2079.
Long, W.L., Ballard, W.W., 2001. Normal embryonic stages of the longnose gar, Lepisosteus osseus. BMC Dev Biol 1, 6.
Lu, M.F., Cheng, H.T., Lacy, A.R., Kern, M.J., Argao, E.A., Potter, S.S., Olson, E.N., Martin, J.F., 1999. Paired-related homeobox genes cooperate in handplate and hindlimb zeugopod morphogenesis. Dev Biol 205, 145-157.
Martin, J.F., Bradley, A., Olson, E.N., 1995. The paired-like homeo box gene MHox is required for early events of skeletogenesis in multiple lineages. Genes Dev 9, 1237-1249.
Mehta, T.K., Ravi, V., Yamasaki, S., Lee, A.P., Lian, M.M., Tay, B.H., Tohari, S., Yanai, S., Tay, A., Brenner, S., Venkatesh, B., 2013. Evidence for at least six Hox clusters in the Japanese lamprey (Lethenteron japonicum). Proc Natl Acad Sci U S A 110, 16044-16049.
Minguillon, C., Gibson-Brown, J.J., Logan, M.P., 2009. Tbx4/5 gene duplication and the origin of vertebrate paired appendages. Proc Natl Acad Sci U S A 106, 21726-21730.
Nakatani, Y., Takeda, H., Kohara, Y., Morishita, S., 2007. Reconstruction of the vertebrate ancestral genome reveals dynamic genome reorganization in early vertebrates. Genome Res 17, 1254-1265.
Near, T.J., Eytan, R.I., Dornburg, A., Kuhn, K.L., Moore, J.A., Davis, M.P., Wainwright, P.C., Friedman, M., Smith, W.L., 2012. Resolution of ray-finned fish phylogeny and timing of diversification. Proc Natl Acad Sci U S A 109, 13698-13703.
Nohno, T., Koyama, E., Myokai, F., Taniguchi, S., Ohuchi, H., Saito, T., Noji, S., 1993. A chicken homeobox gene related to Drosophila paired is predominantly expressed in the developing limb. Dev Biol 158, 254-264.
Norris, R.A., Kern, M.J., 2001. The identification of Prx1 transcription regulatory domains provides a mechanism for unequal compensation by the Prx1 and Prx2 loci. J Biol Chem 276, 26829-26837.
Norris, R.A., Scott, K.K., Moore, C.S., Stetten, G., Brown, C.R., Jabs, E.W., Wulfsberg, E.A., Yu, J., Kern, M.J., 2000. Human PRRX1 and PRRX2 genes: cloning, expression, genomic localization, and exclusion as disease genes for Nager syndrome. Mamm Genome 11, 1000-1005.

Ocana, O.H., Corcoles, R., Fabra, A., Moreno-Bueno, G., Acloque, H., Vega, S., Barrallo-Gimeno, A., Cano, A., Nieto, M.A., 2012. Metastatic colonization requires the repression of the epithelial-mesenchymal transition inducer Prrx1. Cancer Cell 22, 709-724.
Opstelten, D.J., Vogels, R., Robert, B., Kalkhoven, E., Zwartkruis, F., de Laaf, L., Destree, O.H., Deschamps, J., Lawson, K.A., Meijlink, F., 1991. The mouse homeobox gene, S8, is expressed during embryogenesis predominantly in mesenchyme. Mech Dev 34, 29-41.
Postlethwait, J., Amores, A., Cresko, W., Singer, A., Yan, Y.L., 2004. Subfunction partitioning, the teleost radiation and the annotation of the human genome. Trends Genet 20, 481-490.
Postlethwait, J.H., 2007. The zebrafish genome in context: ohnologs gone missing. J Exp Zool B Mol Dev Evol 308, 563-577.
Putnam, N.H., Butts, T., Ferrier, D.E., Furlong, R.F., Hellsten, U., Kawashima, T., Robinson-Rechavi, M., Shoguchi, E., Terry, A., Yu, J.K., Benito-Gutierrez, E.L., Dubchak, I., Garcia-Fernandez, J., Gibson-Brown, J.J., Grigoriev, I.V., Horton, A.C., de Jong, P.J., Jurka, J., Kapitonov, V.V., Kohara, Y., Kuroki, Y., Lindquist, E., Lucas, S., Osoegawa, K., Pennacchio, L.A., Salamov, A.A., Satou, Y., Sauka-Spengler, T., Schmutz, J., Shin, I.T., Toyoda, A., Bronner-Fraser, M., Fujiyama, A., Holland, L.Z., Holland, P.W., Satoh, N., Rokhsar, D.S., 2008. The amphioxus genome and the evolution of the chordate karyotype. Nature 453, 1064-1071.
Reichert, M., Takano, S., von Burstin, J., Kim, S.B., Lee, J.S., Ihida-Stansbury, K., Hahn, C., Heeg, S., Schneider, G., Rhim, A.D., Stanger, B.Z., Rustgi, A.K., 2013. The Prrx1 homeodomain transcription factor plays a central role in pancreatic regeneration and carcinogenesis. Genes Dev 27, 288-300.
Schartl, M., 2013. Beyond the zebrafish: diverse fish species for modeling human disease. Dis Model Mech.
Schulz, M.H., Zerbino, D.R., Vingron, M., Birney, E., 2012. Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics 28, 1086-1092.
Shimozaki, K., Clemenson, G.D., Jr., Gage, F.H., 2013. Paired related homeobox protein 1 is a regulator of stemness in adult neural stem/progenitor cells. J Neurosci 33, 4066-4075.
Smith, J.J., Kuraku, S., Holt, C., Sauka-Spengler, T., Jiang, N., Campbell, M.S., Yandell, M.D., Manousaki, T., Meyer, A., Bloom, O.E., Morgan, J.R., Buxbaum, J.D., Sachidanandam, R., Sims, C., Garruss, A.S., Cook, M., Krumlauf, R., Wiedemann, L.M., Sower, S.A., Decatur, W.A., Hall, J.A., Amemiya, C.T., Saha, N.R., Buckley, K.M., Rast, J.P., Das, S., Hirano, M., McCurley, N., Guo, P., Rohner, N., Tabin, C.J., Piccinelli, P., Elgar, G., Ruffier, M., Aken, B.L., Searle, S.M., Muffato, M., Pignatelli, M., Herrero, J., Jones, M., Brown, C.T., Chung-Davidson, Y.W., Nanlohy, K.G., Libants, S.V., Yeh, C.Y., McCauley, D.W., Langeland, J.A., Pancer, Z., Fritzsch, B., de Jong, P.J., Zhu, B., Fulton, L.L., Theising, B., Flicek, P., Bronner, M.E., Warren, W.C., Clifton, S.W., Wilson, R.K., Li, W., 2013. Sequencing of the sea lamprey (Petromyzon marinus) genome provides insights into vertebrate evolution. Nat Genet 45, 415-421, 421e411-412.
Stanke, M., Steinkamp, R., Waack, S., Morgenstern, B., 2004. AUGUSTUS: a web server for gene finding in eukaryotes. Nucleic Acids Res 32, W309-312.

Takahashi, S., Uochi, T., Kawakami, Y., Nohno, T., Yokota, C., Kinoshita, K., Asashima, M., 1998. Cloning and expression pattern of Xenopus prx-1 (Xprx-1) during embryonic development. Dev Growth Differ 40, 97-104.
Tanaka, M., Munsterberg, A., Anderson, W.G., Prescott, A.R., Hazon, N., Tickle, C., 2002. Fin development in a cartilaginous fish and the origin of vertebrate limbs. Nature 416, 527-531.
ten Berge, D., Brouwer, A., Korving, J., Martin, J.F., Meijlink, F., 1998. Prx1 and Prx2 in skeletogenesis: roles in the craniofacial region, inner ear and limbs. Development 125, 3831-3842.
Tulenko, F.J., McCauley, D.W., Mackenzie, E.L., Mazan, S., Kuratani, S., Sugahara, F., Kusakabe, R., Burke, A.C., 2013. Body wall development in lamprey and a new perspective on the origin of vertebrate paired fins. Proc Natl Acad Sci U S A 110, 11899-11904.
Wolfe, K.H., 2001. Yesterday's polyploids and the mystery of diploidization. Nat Rev Genet 2, 333-341.
Zhong, Y.F., Holland, P.W., 2011. HomeoDB2: functional expansion of a comparative homeobox gene database for evolutionary developmental biology. Evol Dev 13, 567-568.

Figure Captions

Fig. 1: Cladogram showing phylogenetic relationships among vertebrates analyzed in the present study. Tree topology was adopted from Near et al. (2012). VGD1/2: vertebrate genome duplication 1/2; TGD: teleost genome duplication. The results of our *Prrx* gene surveys are shown to the right. Presence/absence of genes is indicated by colored/white boxes. The relationship of the single lamprey *prrx* gene remains unresolved.

Fig. 2: Conserved synteny between human and spotted gar *Prrx* genes. A) A dotplot comparing the human *PRRX1* gene region on chromosome 1 (Hsa1) to the spotted gar (Loc) genome shows extensive conserved synteny to spotted gar LG10. B) Conserved synteny of human and spotted gar *Prrx2* genes. C, D) Orthologous pairwise clusters between human and spotted gar.

Fig. 3: Rayfin conserved synteny and phylogeny of *Prrx* genes. A, B) A dotplot analysis comparing the chromosomal neighborhood surrounding spotted gar *prrx1* on LG10 shows double conserved synteny with chromosome segments containing *prrx1* co-orthologs in zebrafish on Dre2 and Dre20 (A) and in medaka to *prrx1b* on Ola4 and to Ola17, which lacks a *prrx1* gene. C) Nucleotide maximum likelihood phylogeny of vertebrate *Prrx* genes (GTR+I+G model, 50% bootstrap consensus) rooted on amphioxus *prrx*. Node values indicate % bootstrap support. Note that gene names were assigned taking conserved synteny and/or splicing information into account. Aal: allis shad; Aja: Japanese eel; Ame: Mexican tetra; Bfl: amphioxus; Cmi: elephant shark; Elu: Northern pike; Gac: stickleback; Gga: chicken; Gmo: cod; Hsa: human; Lca: Arctic lamprey; Lch: coelacanth; Loc: spotted gar; Mmu: mouse; Ola: medaka; Oni: tilapia; Pbu: butterflyfish; Phy: striped catfish; Xtr: frog.

Fig. 4: Genomic environments of vertebrate *Prrx* genes. Orthologous genes contributing to conserved synteny are similarly color-coded.

Fig. 5: Splice variants of *Prrx* genes. Three domains characterize Prrx proteins: the prx domain, the homeodomain and the OAR domain. The OAR domain is missing from the amniote Prrx1 splisoform B and from rayfin splisoforms B' of Prrx1 (spotted gar) and Prrx1b (teleosts) because of the use of alternative fourth exons.

Fig. 6: Conserved synteny of *Prrx* paralogons after VGD1 and VGD2. *Prrx* genes in spotted gar (A) and human (B) are part of larger paralogons. Intragenomic paralogy dotplots in spotted gar (C) and human (D) show that additional paralogons are present in vertebrate genomes despite the absence of additional *Prrx* genes, as would be expected after two rounds of whole genome duplication followed by gene loss.

Fig. 7: Expression of *prrx* genes in rayfins. A) Expression of *prrx* genes during spotted gar development. RNA whole mount *in situ* hybridizations are shown for *prrx1* to the left, *prrx2* to the right. The upper row shows embryos at Long/Ballard stage (st.) 28-29 at left with a dorsal view of the gar embryo (heads to the left) and at the right a lateral view of the head and anterior trunk region. Both genes are expressed in the pectoral fin buds (fb), branchial arches (white asterisks) and head structures. The lower row shows embryos at stage 32-33. At the left, lateral views of the embryos are shown with a box insert highlighting the region around the pelvic fin bud (fb) that is shown in the magnification on the right. Both *prrx* genes are expressed in the pelvic find buds at this stage, while expression is no longer detectable in the pectorals. B) RNA-Seq based expression analysis of *prrx* genes in tissues of spotted gar, European eel, zebrafish, and medaka. Expression level is measured as reads per kb per million reads (rpkm).

Fig. 8: Model for the evolution of vertebrate *Prrx* genes. The *Prrx* gene family evolved through three round of whole genome duplication, vertebrate genome duplications 1 and 2 (VGD1/VGD2), and the teleost genome duplication (TGD) followed by several instances of gene loss in multiple lineages (OGM: ohnolog gone missing).

Suppl. Fig. 1: Orthology of clupeocephalan teleost *prrx1* genes. Dot plots of zebrafish chromosomes Dre2 (A) and Dre20 (B) containing *prrx1a* and *prrx1b*, respectively, against medaka (Ola) chromosomes. A) Zebrafish *prrx1a* shares highly conserved syntenies with Ola17, while the medaka *prrx1* gene is located on Ola4. B) The zebrafish *prrx1b*-containing chromosome shares conserved syntenies with Ola4, but even more pronounced with Ola 24. Thus, the gene lost in medaka is *prrx1a* and the retained gene is *prrx1b*. Because Ola4 and Ola17 are clearly TGD-derived paralogous chromosomes (see Fig. 3 and Kasahara et al., 2007), we conclude that the *prrx1b* gene region has been translocated to a different chromosomes after the TGD in the zebrafish lineage.

Suppl. Figure 2: Phylogenetic trees of the vertebrate Prrx transcription factors. A) Maximum likelihood (ML) phylogeny of *Prrx* nucleotide sequences (GTR+I+G model). B) ML bootstrap 50% consensus phylogeny, C) ML phylogeny of Prrx proteins (JTT+G+F model). Node values indicate % bootstrap support. Trees were rooted with the amphioxus sequence. Gene/protein names were assigned taking synteny and/or splicing information into account. Aal: Allis shad; Aja: Japanese eel; Ame: Mexican tetra; Bfl: amphioxus; Cmi: elephant shark; Elu: Northern pike; Gac: stickleback; Gga: chicken; Gmo: cod; Hsa: human; Lca: Arctic lamprey; Lch: coelacanth; Loc: spotted gar; Mmu: mouse; Ola: medaka; Oni: tilapia; Pbu: butterflyfish; Phy: striped catfish; Xtr: frog.

Suppl. Figure 3: Expression of *prrx1* co-orthologs in zebrafish and European eel. RNA-Seq based expression level is measured as reads per kb per million reads (rpkm). Despite the differences in absolute expression levels between zebrafish and eel, the relative tissue specific expression patterns of *prrx1* co-orthologs is similar across species with *prrx1a* being generally more highly expressed than *prrx1b*.

Figures

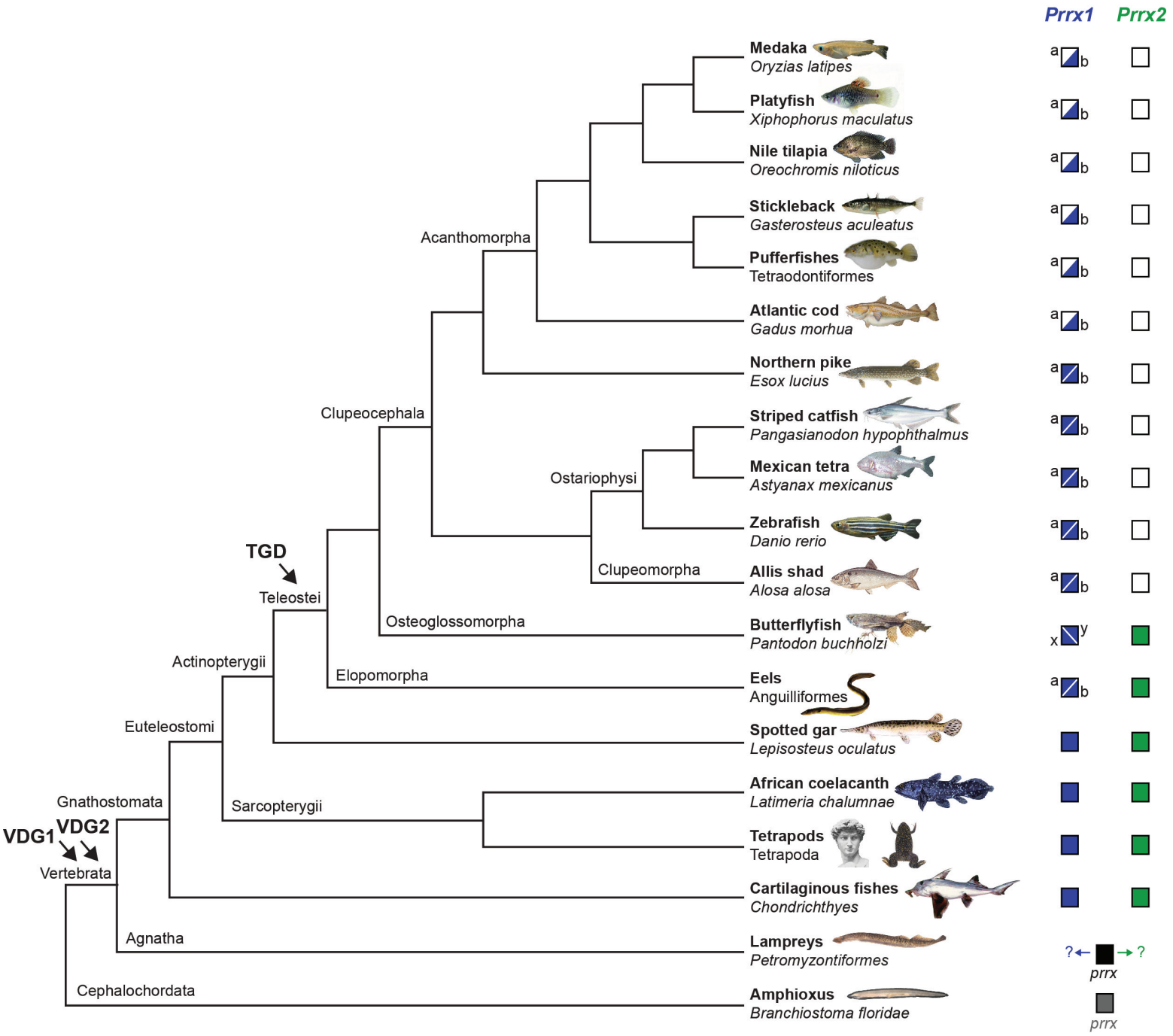


Figure 1

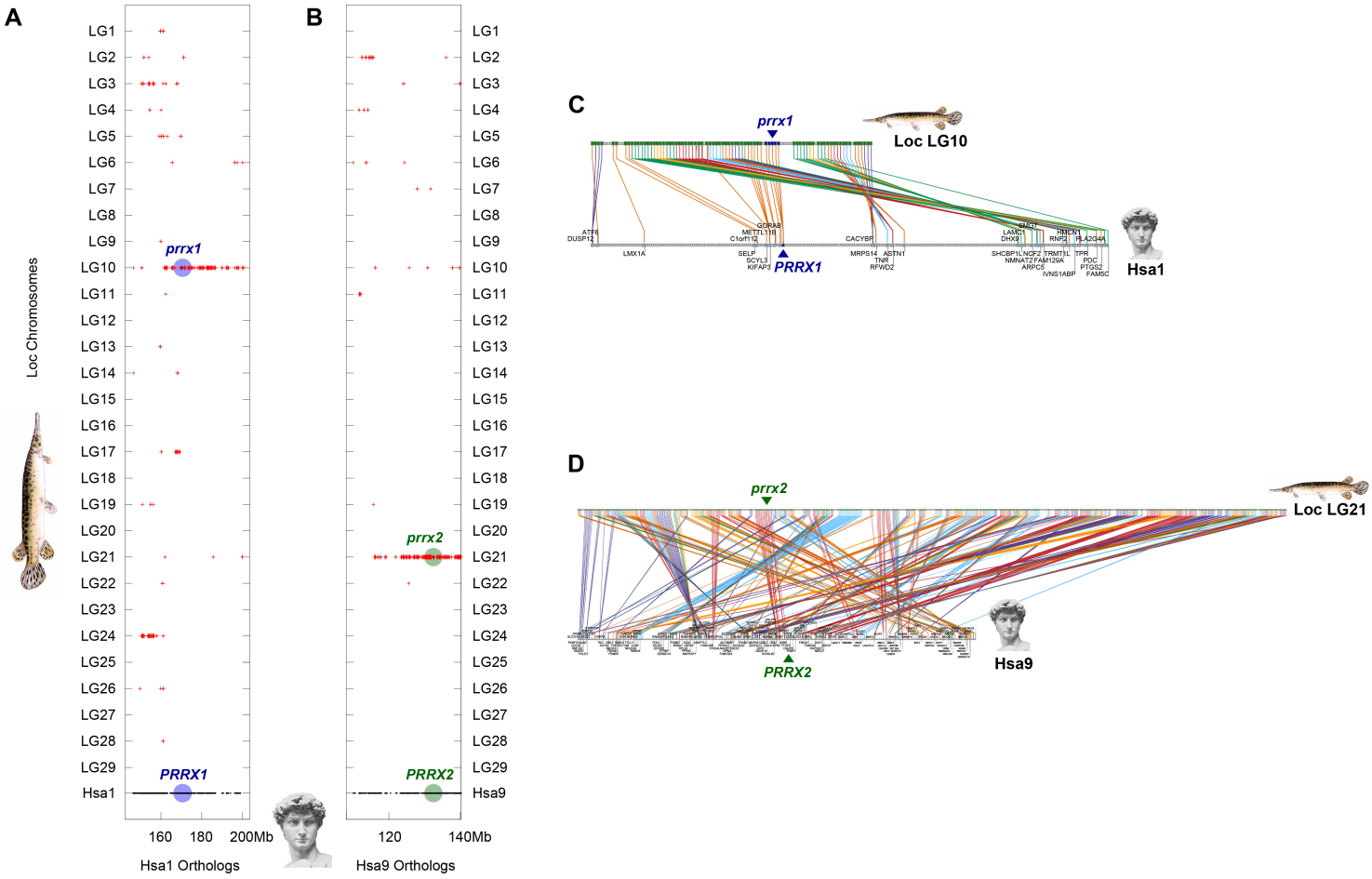

A
B
C
D
Loc Chromosomes
LG1
LG2
LG3
LG4
LG5
LG6
LG7
LG8
LG9
LG10
LG11
LG12
LG13
LG14
LG15
LG16
LG17
LG18
LG19
LG20
LG21
LG22
LG23
LG24
LG25
LG26
LG27
LG28
LG29
Hsa1
Hsa9
prrx1
PRRX1
prrx2
PRRX2
160
180
200Mb
120
140Mb
Hsa1 Orthologs
Hsa9 Orthologs
Loc LG10
Loc LG21
Hsa1
Hsa9

Figure 2

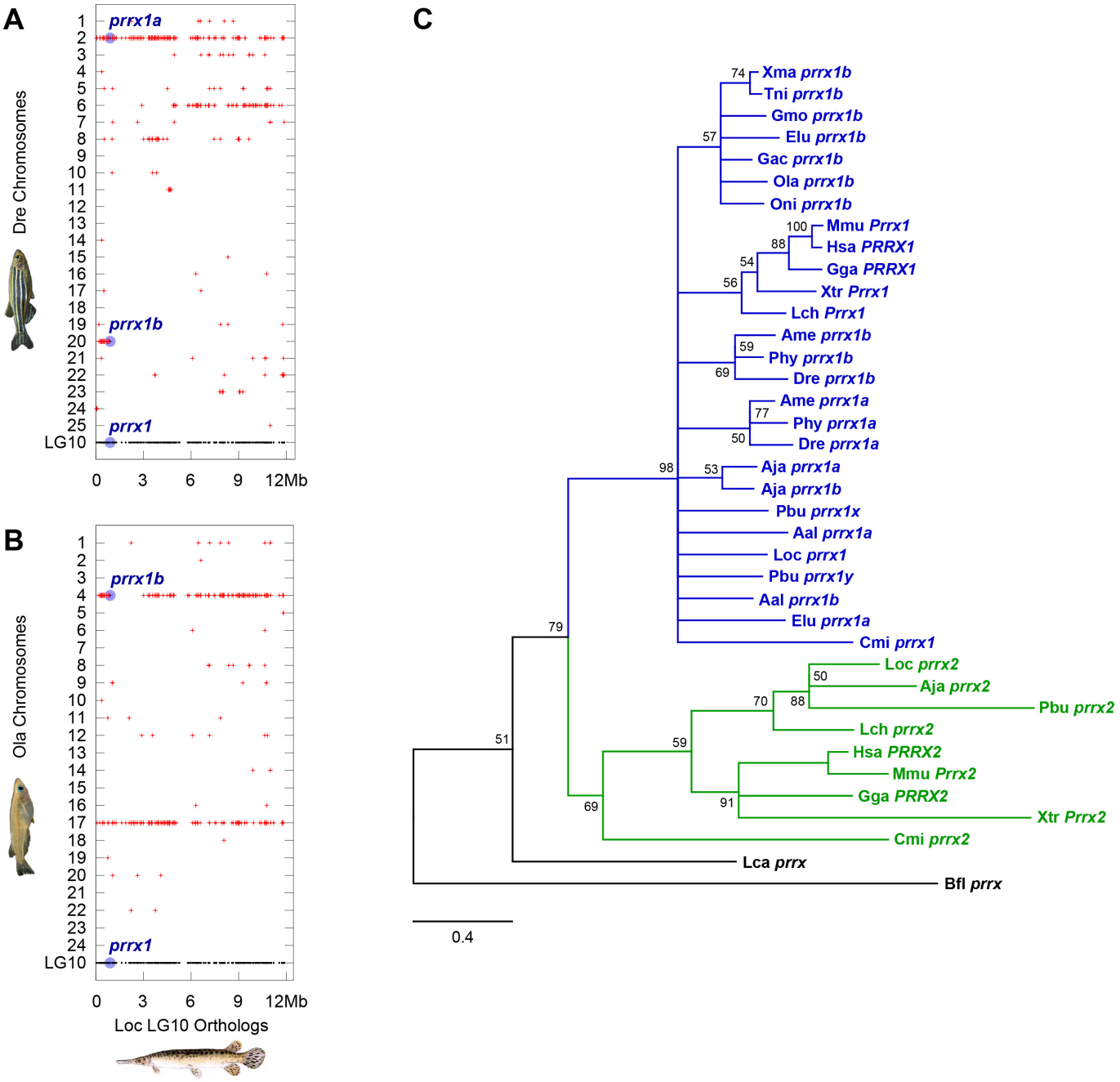

A
Dre Chromosomes
prrx1a
prrx1b
prrx1
LG10
0
3
6
9
12Mb
B
Ola Chromosomes
prrx1b
prrx1
LG10
0
3
6
9
12Mb
Loc LG10 Orthologs
C
Xma prrx1b
Tni prrx1b
Gmo prrx1b
Elu prrx1b
Gac prrx1b
Ola prrx1b
Oni prrx1b
Mmu Prrx1
Hsa PRRX1
Gga PRRX1
Xtr Prrx1
Lch Prrx1
Ame prrx1b
Phy prrx1b
Dre prrx1b
Ame prrx1a
Phy prrx1a
Dre prrx1a
Aja prrx1a
Aja prrx1b
Pbu prrx1x
Aal prrx1a
Loc prrx1
Pbu prrx1y
Aal prrx1b
Elu prrx1a
Cmi prrx1
Loc prrx2
Aja prrx2
Pbu prrx2
Lch prrx2
Hsa PRRX2
Mmu Prrx2
Gga PRRX2
Xtr Prrx2
Cmi prrx2
Lca prrx
Bfl prrx
0.4

Figure 3

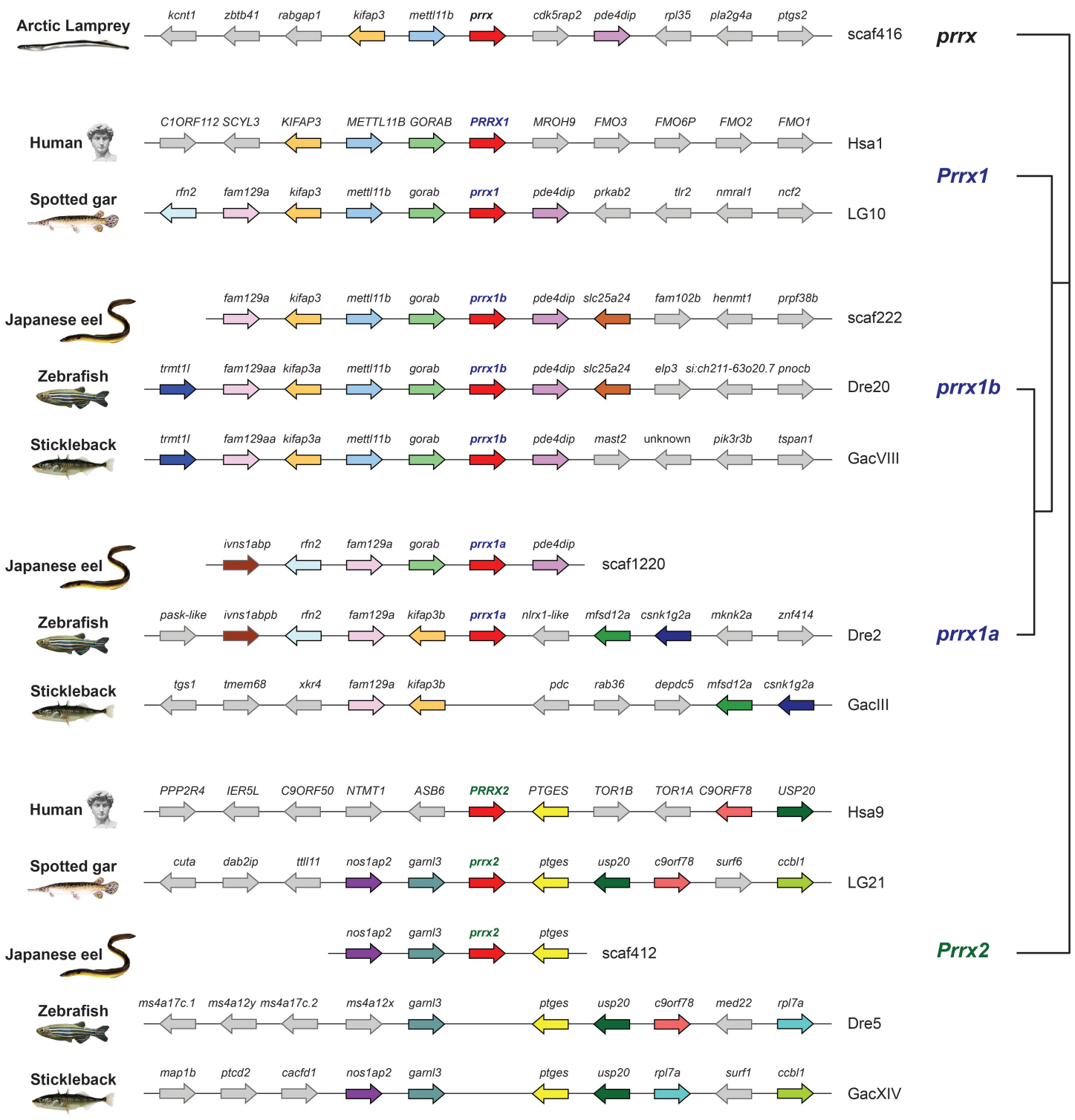

Arctic Lamprey
kcnt1 zbtb41 rabgap1 kifap3 mettl11b prrx cdk5rap2 pde4dip rpl35 pla2g4a ptgs2
scaf416
prrx
Human
C1ORF112 SCYL3 KIFAP3 METTL11B GORAB PRRX1 MROH9 FMO3 FMO6P FMO2 FMO1
Hsa1
Spotted gar
rfn2 fam129a kifap3 mettl11b gorab prrx1 pde4dip prkab2 tlr2 nmral1 ncf2
LG10
Prrx1
Japanese eel
fam129a kifap3 mettl11b gorab prrx1b pde4dip slc25a24 fam102b henmt1 prpf38b
scaf222
Zebrafish
trmt1l fam129aa kifap3a mettl11b gorab prrx1b pde4dip slc25a24 elp3 si:ch211-63o20.7 pnocb
Dre20
prrx1b
Stickleback
trmt1l fam129aa kifap3a mettl11b gorab prrx1b pde4dip mast2 unknown pik3r3b tspan1
GacVIII
Japanese eel
ivns1abp rfn2 fam129a gorab prrx1a pde4dip
scaf1220
Zebrafish
pask-like ivns1abpb rfn2 fam129a kifap3b prrx1a nlrx1-like mfsd12a csnk1g2a mknk2a znf414
Dre2
prrx1a
Stickleback
tgs1 tmem68 xkr4 fam129a kifap3b pdc rab36 depdc5 mfsd12a csnk1g2a
GacIII
Human
PPP2R4 IER5L C9ORF50 NTMT1 ASB6 PRRX2 PTGES TOR1B TOR1A C9ORF78 USP20
Hsa9
Spotted gar
cuta dab2ip ttll11 nos1ap2 garnl3 prrx2 ptges usp20 c9orf78 surf6 ccbl1
LG21
Japanese eel
nos1ap2 garnl3 prrx2 ptges
scaf412
Prrx2
Zebrafish
ms4a17c.1 ms4a12y ms4a17c.2 ms4a12x garnl3 ptges usp20 c9orf78 med22 rpl7a
Dre5
Stickleback
map1b ptcd2 cacfd1 nos1ap2 garnl3 ptges usp20 rpl7a surf1 ccbl1
GacXIV

Figure 4

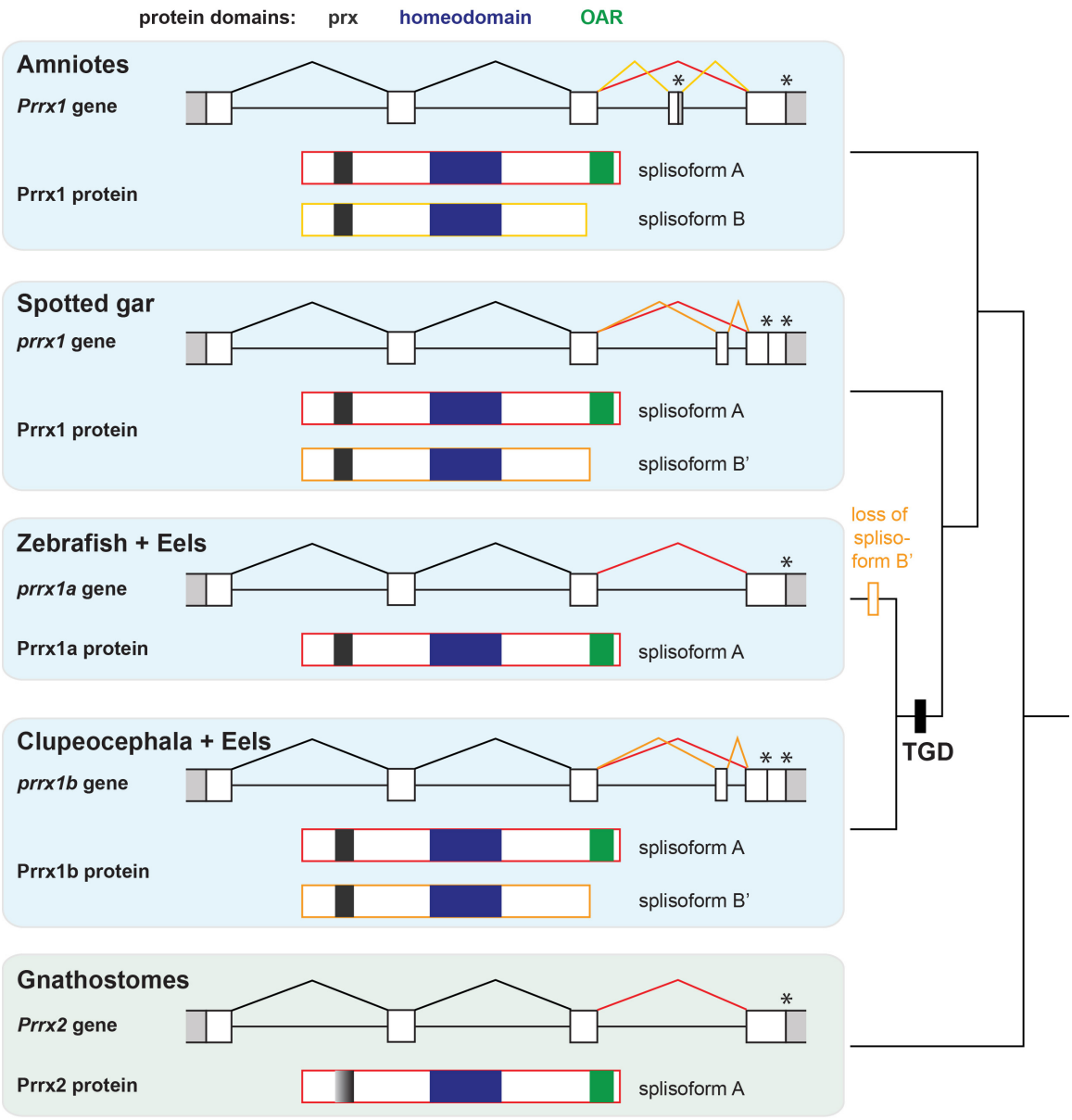
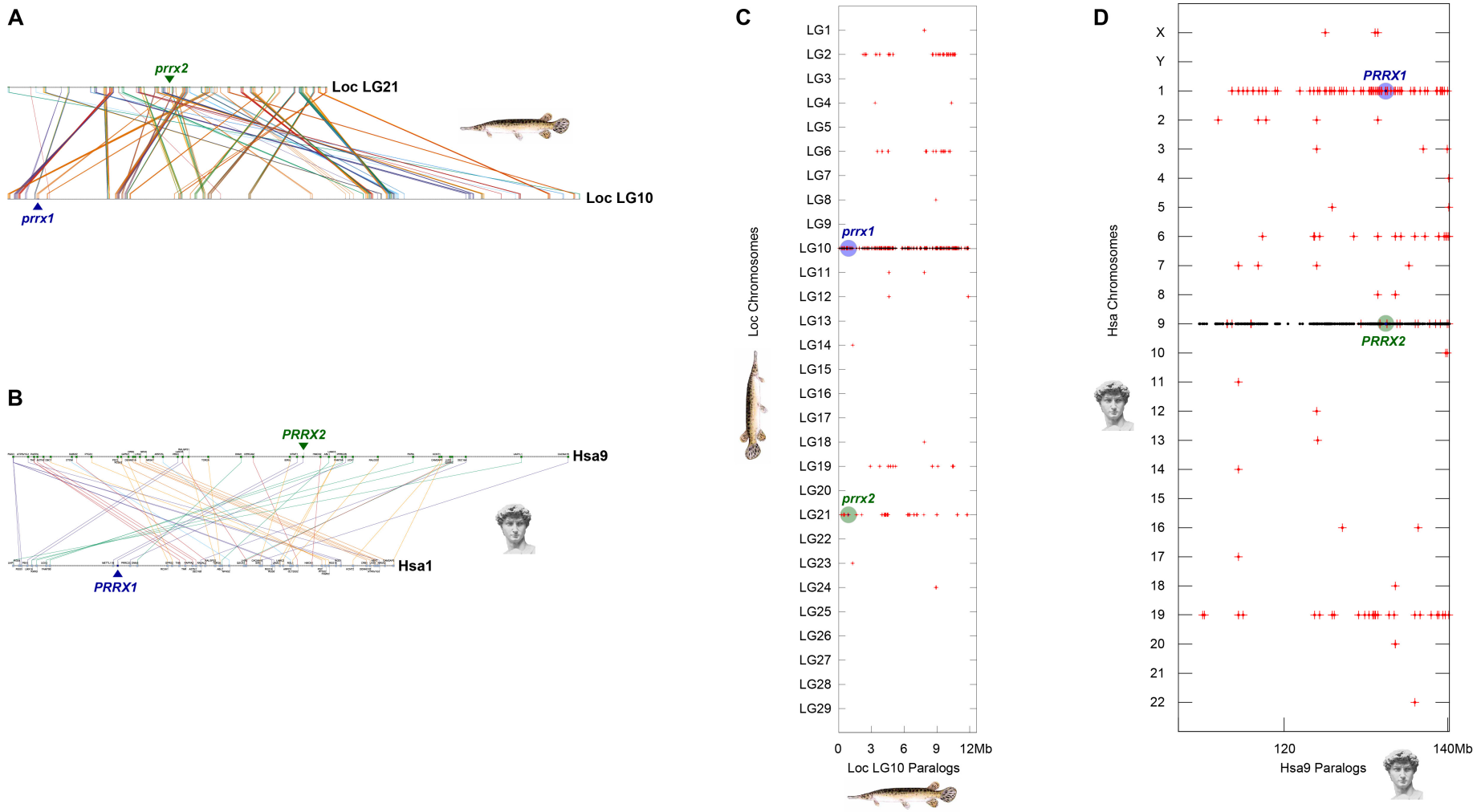

protein domains:
prx
homeodomain
OAR
Amniotes
Prrx1 gene
Prrx1 protein
splisoform A
splisoform B
Spotted gar
prrx1 gene
Prrx1 protein
splisoform A
splisoform B'
Zebrafish + Eels
prrx1a gene
Prrx1a protein
splisoform A
Clupeocephala + Eels
prrx1b gene
Prrx1b protein
splisoform A
splisoform B'
Gnathostomes
Prrx2 gene
Prrx2 protein
splisoform A
loss of
spliso-
form B'
TGD

Figure 5

Figure 6


A
prrx2
Loc LG21
Loc LG10
prrx1
B
PRRX2
Hsa9
Hsa1
PRRX1
C
Loc Chromosomes
LG1
LG2
LG3
LG4
LG5
LG6
LG7
LG8
LG9
LG10
LG11
LG12
LG13
LG14
LG15
LG16
LG17
LG18
LG19
LG20
LG21
LG22
LG23
LG24
LG25
LG26
LG27
LG28
LG29
prrx1
prrx2
0
3
6
9
12Mb
Loc LG10 Paralogs
D
Hsa Chromosomes
X
Y
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
PRRX1
PRRX2
120
140Mb
Hsa9 Paralogs

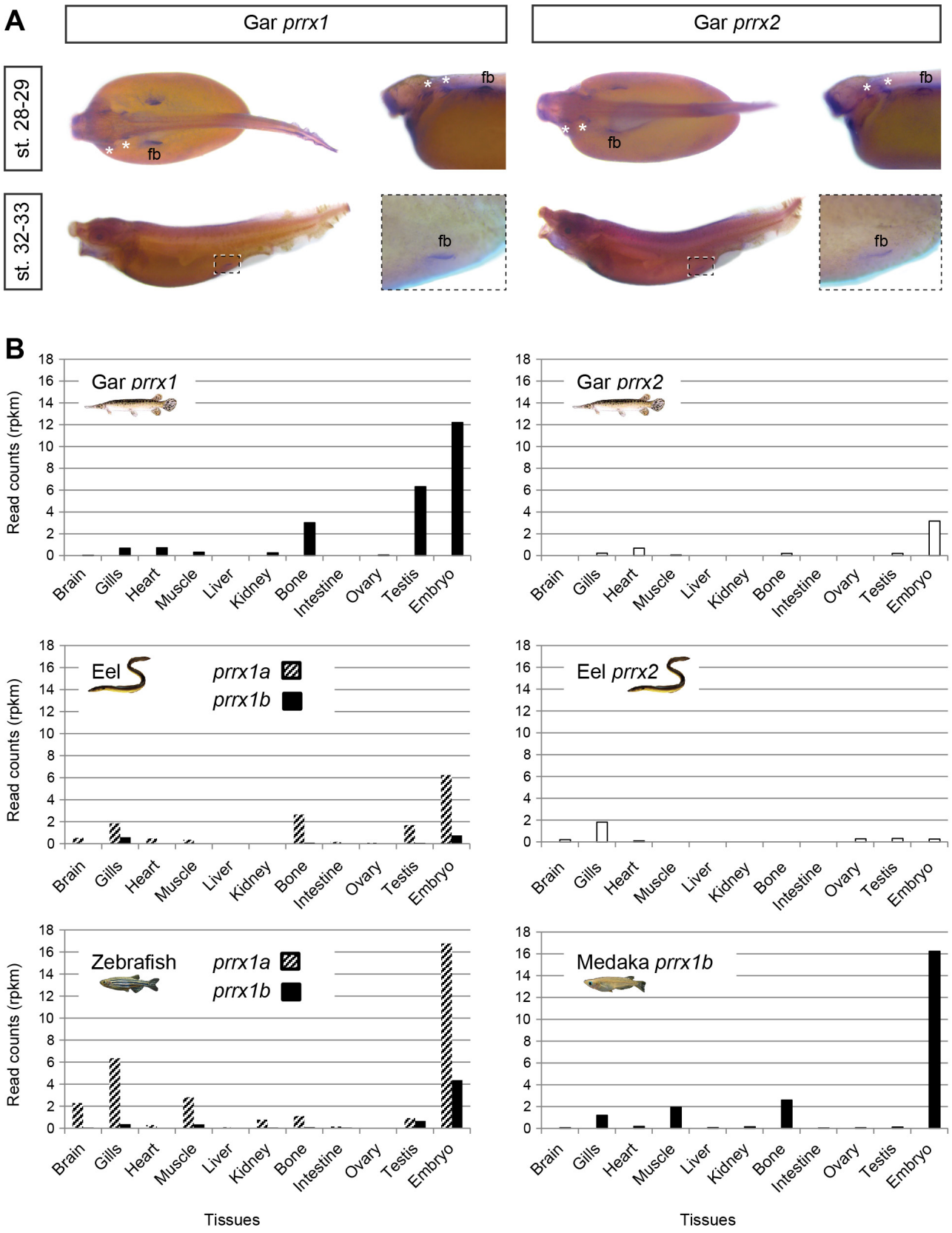

A
Gar prrx1
Gar prrx2
st. 28-29
st. 32-33
fb
B
Gar prrx1
Gar prrx2
Eel
Eel prrx2
Zebrafish
Medaka prrx1b
prrx1a
prrx1b
Read counts (rpkm)
Brain
Gills
Heart
Muscle
Liver
Kidney
Bone
Intestine
Ovary
Testis
Embryo
Tissues

Figure 7

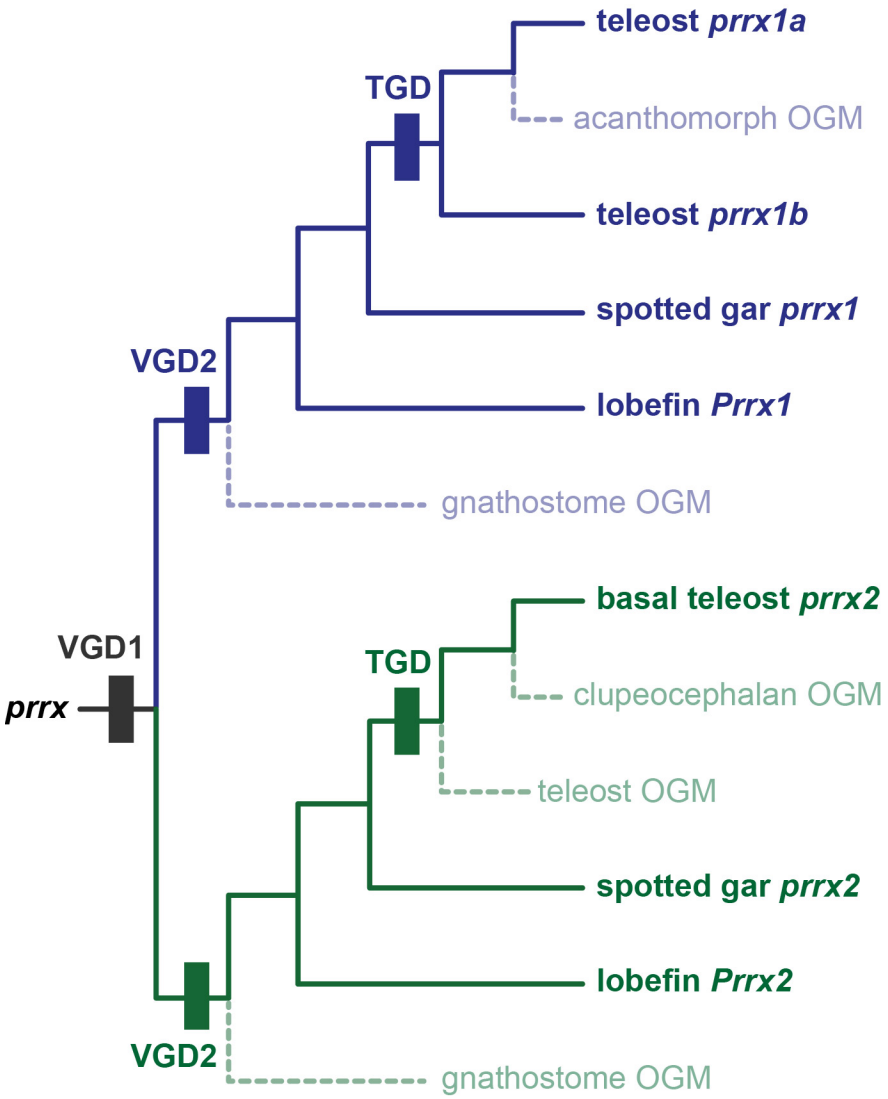

teleost prrx1a
TGD
acanthomorph OGM
teleost prrx1b
spotted gar prrx1
VGD2
lobefin Prrx1
gnathostome OGM
basal teleost prrx2
VGD1
TGD
clupeocephalan OGM
prrx
teleost OGM
spotted gar prrx2
lobefin Prrx2
VGD2
gnathostome OGM

Figure 8