A. Supplementary Note

1. Biology of spotted gar
2. The spotted gar genome
    - 2.1 Genome sequencing and assembly
    - 2.2 Additional genomic resources for spotted gar
3. RNA-seq transcriptomes
    - 3.1 Broad Institute gar transcriptome
    - 3.2 PhyloFish gar transcriptome
    - 3.3 PhyloFish bowfin transcriptome
4. Genome annotation
    - 4.1 MAKER annotation
    - 4.2 Ensembl annotation
5. Transposable elements
6. Phylogenomic analysis
7. Molecular evolutionary rate analysis
8. Evolution of vertebrate genome structure
    - 8.1 The spotted gar karyotype
    - 8.2 Synteny analyses
9. Gene family analyses
    - 9.1 Hox clusters
    - 9.2 ParaHox clusters
    - 9.3 Aldh1a genes
    - 9.4 Clock genes
    - 9.5 Opsin genes
    - 9.6 Immune genes
10. Annotation and expression of gar mineralization genes
11. Spotted gar miRNA genes
    - 11.1 *in silico* analyses of gar miRNAs
    - 11.2 Small RNA sequencing-based analyses
12. Conserved non-coding elements
    - 12.1 Conserved noncoding elements at selected developmental gene loci
    - 12.2 Whole genome alignment-based analysis of CNEs
    - 12.3 Connectivity analysis of human limb enhancers
    - 12.4 HoxD limb enhancer CNS65
13. Analysis gene expression after the teleost genome duplication
    - 13.1 Identification of TGD ohnologs and singletons in zebrafish and medaka
    - 13.2 Comparative RNA-seq expression analysis of gar vs. zebrafish and medaka

B. Supplementary Tables

* separate file

C. Supplementary Figures

Supplementary Fig. 25. Evolutionary relationships among gar and teleost novel immune-type receptors
Supplementary Fig. 26. Conserved synteny of the spotted gar *scpp* gene region on LG2 with teleosts
Supplementary Fig. 27. miRNA genes in non-teleost vertebrates vs. teleosts
Supplementary Fig. 28. miRNA gene annotation in gar based on RNA-seq and orthology searches
Supplementary Fig. 29. Total CNE length of selected gnathostome developmental gene loci
Supplementary Figs. 30-35. VISTA analyses of gnathostome developmental gene loci
Supplementary Fig. 36. phyloFit trees of 13-way whole genome alignments based on 4d sites
Supplementary Fig. 37. Evolution of human limb enhancers

* separate file

D. Supplementary Files

Supplementary File 1. Phylogenomic alignment file in phylip format*

* separate file

E. Source Data Files

Source Data Set 1. Source Data for Figure 1*
Source Data Set 2. Source Data for Figure 2*
Source Data Set 3. Source Data for Figure 3*
Source Data Set 4. Source Data for Figure 6*

* separate file

F. Supplementary References

A. SUPPLEMENTARY NOTE

1. Biology of spotted gar

The ray-finned fish spotted gar *Lepisosteus oculatus* (Winchell 1864) occupies a key node for genome sequencing because it is a representative of an outgroup lineage that diverged from the teleost lineage before the teleost genome duplication (TGD)[1,2]. Gars are well-suited genomic and laboratory models, while other non-teleost ray-finned fish groups (i.e., bichirs, sturgeons, paddlefish, and bowfin) are problematic due to derived morphologies, lineage-specific polyploidizations, and/or difficult husbandry[3].

Spotted gar (Supplementary Fig. 1) is one of seven extant species of the ancient family of Lepisosteidae, which includes two genera: *Lepisosteus* with four species (spotted, Florida, longnose, and shortnose gars) and *Atractosteus* with three species (tropical, Cuban, and alligator gars)[4]. Spotted gar is one of the smaller species of the family, with typical adult sizes of about 100-120 cm total length[5].

Extant gar species are restricted to North and Central America and Cuba. The distribution of spotted gar ranges from southern Canada (Lake Erie and southern Lake Michigan drainage) to the Gulf Coast and northern Mexico[6,7]. Spotted gar samples used in the present study were obtained in Louisiana from the 'core population', which is disjoined from a northern 'peripheral population'[8] (Supplementary Fig. 1d).

The gar fossil record dates back at least 100 million years to the Early Cretaceous[9]. In the *Origin of Species*, Darwin (1859) used ganoid fishes (which include gars, see below) as one defining example to introduce the term 'living fossil'[10]. Among ray-finned fish, gars are characterized by exceptionally low rates of speciation and phenotypic evolution[11]. Our analyses here additionally show a slow molecular evolutionary rate in spotted gar compared to teleosts (see main text and Supplementary Note 7).

The phylogenetic position of gars among ray-finned fish has been controversial (the 'gar-*Amia*-teleost' problem[9]), but is essential for interpreting ray-finned morphological evolution[12] and for identifying the closest living outgroup to the TGD. Two major hypotheses have been discussed: While many morphological analyses favored bowfin (*Amia calva*) as the sister lineage to teleosts to the exclusion of gars (Supplementary Fig. 1e, left), previous molecular phylogenetic studies based on limited data mostly supported a monophyletic clade of Holostei (gars plus *Amia*) as the sister lineage to teleosts (Supplementary Fig. 1e, right)[9,12]. Our phylogenomic analyses here with genome-wide data provide strong support for the latter hypothesis (main text, Fig. 1b, and Supplementary Note 6).

Several characteristics make gars particularly important to understand the evolution of vertebrate body plans, including fin morphologies and fin development that are key to the understanding of vertebrate fin and limb evolution and hence the invasion of the land by vertebrates[13,14] (see also main text, Fig. 5). Gars are facultative air breathers that, in addition to

their gills, use their gas bladder as a respiratory organ (lung), particularly at high water temperatures and low oxygen concentrations[15,16]. Gars possess enamel-bearing teeth as well as ganoid scales (containing ganoin, which has been hypothesized to be a type of enamel)[17,18] and are thus of major importance for the understanding of vertebrate mineralization (see main text and Supplementary Note 10). It has been suggested that gars have the ability for UV light perception[19], which is supported by our analysis of the spotted gar opsin gene repertoire (see Supplementary Note 9.5).

In nature, spotted gar spawns in spring in inundated floodplains or on vegetation, producing brood clutches of hundreds to thousands of adhesive and poisonous eggs[20]. In the laboratory, hormone injections can induce spawning[5] (Supplementary Fig.1b). The development of spotted gar and other gar species has been described[3,21]. Embryos are amenable to gene expression studies such as RNA *in situ* hybridization[22,23] and they can be raised to adulthood in captivity (Supplementary Fig.1a-c).

Taken together, spotted gar is a powerful new model system for studying vertebrate genomics, evolution, and development because of its important phylogenetic position as ray-finned fish outgroup to the TGD as well as its accessibility for developmental studies[2,3,24].

2. The spotted gar genome

2.1 Gar genome sequencing and assembly

Animal work was approved by the University of Oregon Institutional Animal Care and Use Committee (Animal Welfare Assurance Number A-3009-01, IACUC protocol 12-02RA).

A single wild adult female *Lepisosteus oculatus*, collected in Bayou Chevreuil, St. James Parish, Louisiana, USA at the coordinates 29 54'50.56"N 90 47'56.85"W, was sacrificed in the laboratory of John H. Postlethwait (University of Oregon). This spotted gar DNA was sequenced to 90X total coverage by Illumina sequencing technology comprising 45X coverage of 180 bp fragment libraries, 42X coverage of 3kb sheared jumping libraries, 2X coverage of 6-14kb sheared jumping libraries, and 1X coverage of Fosill jumping libraries[25]. The sequence was then assembled into LepOcu1 (Accession number AHAT00000000.1) using ALLPATHS-LG[26]. The draft assembly is 945 Mb in size and consists of 869 Mb of sequence plus gaps between contigs. The spotted gar genome assembly has a contig N50 size of 68.3 kb, a scaffold N50 size of 6.9 Mb, and quality metrics comparable to other Illumina genome assemblies[26].

The ALLPATHS-LG assembly was anchored to a high quality RAD-tag meiotic map[2]. First, mapped marker sequences were aligned to genome sequencing scaffolds using BLAST[27]. Scaffolds aligning logically to markers in the same linkage group were considered anchored. If multiple scaffolds aligned logically to a single linkage group, they were then combined into a single linkage group and represented as such in the resulting agp file. Scaffolds that did not anchor at all, or did not anchor logically, remained unchanged and were appended after the linkage groups in the agp file.

2.2 Additional genomic resources for spotted gar

The following spotted gar genomic libraries are available upon request from the Postlethwait Laboratory (Institute of Neuroscience, University of Oregon):

A **fosmid library** (~4.5X genome coverage) generated from a single unsexed spotted gar juvenile, consisting of ~120,000 fosmid clones (vector pCC2Fos; insert size 40kb).

BAC library VMRC-55 (~15X genome coverage) generated from a single adult **male spotted gar** wild-caught near Thibodeaux, Louisiana, consisting of 99,840 BAC clones (vector pCC1BAC; restriction enzyme *Eco*RI; average insert size ~150kb).

BAC library VMRC-56 (~15X genome coverage) generated from a single adult **female spotted gar** wild-caught near Thibodeaux, Louisiana, consisting of 99,840 BAC clones (vector pCC1BAC; restriction enzyme *Eco*RI; average insert size ~150kb).

For these libraries, fosmid and BAC pool DNAs are available in the Postlethwait Lab to PCR screen for genomic regions of interest.

3. RNA-seq reference transcriptomes and assemblies

3.1 Broad Institute gar transcriptome (SRA accession number: SRP042013)

A panel of 10 spotted gar tissues were RNA-sequenced to aid with genome annotation, including stage 28 embryo[21], 8 day larvae; eye, liver, heart, skin, muscle, kidney, and brain from a single unsexed juvenile; and testis from a single mature adult male wild-caught near Thibodeaux, Louisiana. All RNAs were extracted at the University of Oregon and the RNA-seq libraries were then produced at the Broad Institute's Genomics Platform by the strand-specific dUTP method[28] from Oligo dT polyA-isolated RNA. RNA-seq libraries were sequenced using Illumina Hi-Seq, producing 101bp reads (6-9 Gb of sequence/tissue). All RNA-seq datasets were assembled via the genome-independent RNA-seq assembler Trinity[29] into 1,935,767 transcripts.

3.2 PhyloFish spotted gar transcriptome (SRA accession number: SRP044782)

Adult tissues were collected from wild animals near Thibodeaux, Louisiana. Embryos were grown at the University of Oregon. RNA was extracted from the following tissues: brain, gills, heart, muscle, liver, kidney, bone, and intestine from one adult gar female; ovary from one adult gar female; testis from one adult gar male; developmental stage 27-28[21], pool of three embryos. 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 the 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. The 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.

For each library, raw sequence data in fastq format were filtered to remove unknown nucleotides. The longest subsequences without uncalled bases (Ns) exceeding half of the total read length were extracted. Transcriptome *de novo* assembly was performed using Velvet and Oases[30]. We first performed nine independent assemblies using different k-mers (k-mers for velveth: 25,31,37,43,49,55,61,65,69; parameters for velvetg: -read_trkg yes -min_contig_lgth 100 -cov_cutoff 4; parameters for oases: -cov_cutoff 4). Raw transcripts.fa files were filtered to retain only 1% of transcripts per locus with a modified version of a Perl script developed at Brown University [https://sites.google.com/a/brown.edu/bioinformatics-in-biomed/velvet-and-oases-transcriptome]. Anti-sense chimeras accidentally produced during the assembly step were cut with a homemade script. Then, independent assemblies were pooled and duplicate/similar transcripts built by close k-mers were removed by a cd-hit-est[31] step (parameters: -M 0 -d 0 -c 0.98) and merged by a TGICL[32] step (parameters: -l 60 -p 96 -s 100000). After this assembly process, all input reads were mapped back to the set of transcripts using BWA[33] and the size of the longest open reading frames (ORFs) for each transcript was computed using the getorf EMBOSS tool[34]. Finally, transcripts were filtered using mapping rate and ORF length criteria. Transcripts with ORFs shorter than 200 nt and with fewer than two mapped reads for 1 million overall mapped reads were discarded.

The library-specific assembly was followed by a meta-assembly step. Transcripts from all conditions were pooled. The longest ORF of each transcript was extracted and ORFs were clusterized using cd-hit (parameters: -M 0 -d 0 -c 0.90 -g 1). From each cd-hit cluster, the transcript with the longest ORF or the longest transcript (if more than one transcript had an ORF of the maximum size) was selected. Input reads from all conditions were mapped back to selected transcripts using BWA. Again, transcripts were filtered based on the re-mapping rate. We discarded transcripts with less than 1 mapped read for 1 million overall mapped reads.

A total of 700 million reads were generated with an average number of reads of 70 million reads per library. The final meta-assembly consisted of 41,396 contigs on which 93% of the reads could be re-mapped.

Fish used in all PhyloFish RNA-seq experiments (Supplementary Notes 3.2, 3.3, 13.2) were reared and handled in strict accordance with French and Europeans policies and guidelines of the INRA LPGP Institutional Animal Care and Use Committee (# 25M10), which specifically approved this study.

See also http://phylofish.sigenae.org/ngspipelines/#!/NGSpipelines/Lepisosteus%20oculatus.

3.3 PhyloFish bowfin transcriptome (SRA accession number: SRP044783)

The bowfin (*Amia calva*) transcriptome was assembled as described above for spotted gar. Adult tissues were collected from wild animals near Thibodeaux, Louisiana. RNA was extracted from the following tissues: brain, gills, heart, muscle, liver, kidney, bone, and intestine from one adult bowfin female; ovary from one adult bowfin female; testis from one adult bowfin male.

A total of 700 million reads were generated with an average number of reads of 70 million reads per library. The final meta-assembly consisted of 35,064 contigs on which 94% of the reads could be re-mapped.

See also http://phylofish.sigenae.org/ngspipelines/#!/NGSpipelines/Amia%20calva.

4. Genome annotation

4.1 MAKER annotation

MAKER version 2.29[35] was run on the gar genome sequence using assembled gar RNA-seq data from the Broad Institute and PhyloFish (Supplementary Notes 3.1,3.2), all RefSeq teleost proteins (downloaded July 30, 2013 from http://www.ncbi.nlm.nih.gov), and all Uniprot/swissprot proteins (downloaded July 29, 2013 from ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/complete) as evidence. Repetitive regions were masked using custom repeat library 'Lo-TEs-v3.fa' (see Supplementary Note 5), and a list of known transposable elements[36] provided by MAKER[35]. Additional areas of low complexity were soft masked[37] using Repeatmasker to prevent the seeding of evidence alignments in those regions but still allowing extension of evidence alignments through them[27,38]. Genes were predicted using SNAP[39] and Augustus[40,41] trained for gar using MAKER in an iterative fashion as described by Cantarel et. al.[38], and Genemark trained on the genomic assembly[42].

The spotted gar MAKER annotation is available for download at http://rayfin.uoregon.edu/gar/MAKER_annotation_gar.tar.gz.

4.2 Ensembl annotation

The gar genome assembly (version GCA_000242695.1) was repeat-masked with RepeatMasker [Smit, A.F.A., Hubley, R. & Green, P. RepeatMasker: RepeatMasker Open-3.0. 1996-2010. www.repeatmasker.org] and Dust[43], which identified 26% of the genome as repetitive sequence. Additional low complexity regions were identified using TRF[44].

Protein-coding models were generated by aligning all ray-finned fish (Actinopterygii) protein sequences from UniProt and by aligning other vertebrate protein sequences from UniProt protein existence (PE) levels 1 and 2. These alignments were made to the repeat-masked genome using Genewise[45].

Protein-coding models were also generated using our in-house RNA-seq pipeline[46] using RNASeq data provided by the Broad Institute (Supplementary Note 3.1). These data were aligned to the genome using BWA[33], resulting in just under 500 million reads aligning from ~750 million reads. The alignments were processed by collapsing the transcribed regions into a set of potential exons. Partially aligned reads were re-mapped using Exonerate[47] and this step identified an additional 102 million spliced reads or introns. These introns together with the set of transcribed exons were combined to produce a set of 19,683 transcript models. The longest open reading frame in each of these models were aligned using BLAST[27] against the set of UniProt PE levels 1 and 2 protein sequences in order to classify the models according to their protein-coding potential.

Data from the above two pipelines were filtered to remove poorly supported models. Untranslated regions were added to the coding models using RNA-seq models. The preliminary sets of coding models were combined and redundant models were removed. Additionally we aligned the longest protein coding transcript of each gene from the Zebrafish Ensembl 67 gene set and used these models to fill in gaps in the annotation set resulting in an extra 321 gene models. The remaining unique set of transcript models were clustered into multi-transcript

genes where each transcript in a gene has at least one coding exon that overlaps a coding exon from another transcript within the same gene.

The set of protein-coding gene models was screened for pseudogenes. Short non-coding RNA genes were predicted using annotation from RFAM[48] and miRBase[49]. The final Ensembl gene set consists 18,328 protein coding genes, 42 pseudogenes and 2,595 short non-coding genes.

Further information about the Ensembl gar gene annotation, released in Ensembl74, can be found at http://www.ensembl.org/Lepisosteus_oculatus/Info/Annotation#assembly.

5. Transposable elements

Construction of a spotted gar transposable element library and genome analyses. Transposable elements (TEs) were annotated following the universal classification described by Wicker et al.[50]. The spotted gar TE-specific library was both manually and automatically built. Two different automatic libraries were built using RepeatScout and RepeatModeler with default parameters. In parallel, each known superfamily was searched by tblastn using annotated TE proteins (transposase for DNA transposons and reverse transcriptase for retrotransposons) to better characterize diversity. Specific features such as LTRs (for Long Terminal Repeats) and TIRs (for Terminal Inverted Repeats) were also searched to obtain the most complete sequences in each family. The three different annotations were combined removing redundancies into the custom library 'Lo-TEs-v3.fa'. The genome assembly was then masked using RepeatMasker 3.3.0 [A.F.A. Smit, R. Hubley & P. Green, www.repeatmasker.org] with the custom library and default parameters.

To parse results, total coverage (in base pairs and percentage) and copy numbers were first estimated based on RepeatMasker outfiles (".out" file) using an in house script. Intra-TE insertions were filtered. In a second step, the same statistics were calculated but with filtering sequences smaller than 80 nucleotides and sharing less than 80% of identity with the reference sequence of the TE library.

To estimate the age of TEs in the genome, we assumed that most TEs are silenced by the host genome after insertion and that accumulated mutations in such dead copies are neutral. Therefore, the distance (i.e., the number of mutations) between the individual TE copy in the genome and the corresponding TE consensus sequence in the TE library indicates the potential age of the TE. To correct for multiple mutations at the same site, we used the Kimura distance as the age of the TE. The proportions of transversions q and transitions p were calculated in the alignment file from RepeatMasker. The rates were then transformed into Kimura distances using [$K = -\frac{1}{2}\ln(1 - 2p - q) - \frac{1}{4}\ln(1 - 2q)$]. To detect potential expansion of TEs in the gar genome, our null hypothesis is that if the activities of TEs were uniform during evolution, we would also observe TEs of different ages accumulated uniformly in the genome.

Transcriptome analyses. Potentially active transposable elements in the gar genome were identified by the following steps: First, the assembled RNA-seq transcripts from different tissues (Broad Institute transcriptome, Supplementary Note 3.1) were masked separately using RepeatMasker [http://www.repeatmasker.org/] version 4.0.3 with our custom built repeat library

(Lo-TEs.v3.fa) in the sensitive mode (-s). We kept sequences for further analyses when the TE covered 80% of the sequence length and had a Smith-Waterman score >=220. Second, to exclude potentially exonized TEs, copies of the TE were searched against 'normal' gar protein-coding genes using blastx in the NCBI blast toolkit. A transcript that contained fragments from a TE and the normal gene set (Evalue<=1E-5) was classified as a putative exonized TE. Third, we also excluded from further analysis transcripts that contain fragments from different TE classes (e.g. LINE and DNA transposon).

Results. TEs account for 20% of the gar genome assembly (Supplementary Tab. 1) compared to 46% in human[51], 51% in zebrafish[52], and 16% in medaka[53]. The gar genome harbors a high TE diversity, containing more than 30 superfamilies, which represent almost all previously described TE superfamilies in eukaryotes. Among the four classes (LINE, SINE, LTR and DNA transposons), DNA transposons are the most diverse class with 10 superfamilies (EnSpm, Harbinger, Helitron, Kolobok, MuDr, PiggyBac, Polinton, Sola, TcMariner and hAT; Supplementary Tab. 1), of which the TcMariner is the most abundant superfamily (around 2% of the genome). In contrast, LINE retrotransposons, which only comprise six superfamilies (CR1-like, L1, R2, R4, RTE and Penelope), are the most abundant class (5.8%). Despite this high diversity, LINE retrotransposons are mostly represented by the CR1-like superfamily, which comprises CR1, L2 and Rex1-Babar (4.2% of the genome). Finally, the SINE and LTR retrotransposons have similar content (around 2.6%) in the genome.

Compared to other vertebrate genomes, the TE content of the gar genome is within the range of other bony vertebrates (Supplementary Fig. 2a)[54]. Indeed, the total TE content is close to coelacanth, medaka and cod, higher than birds and turtle, but lower than other tetrapods. Regarding the distribution of the four classes (DNA transposons, LTR, LINE and SINE retrotransposons), the spotted gar profile is most similar to the coelacanth profile (Supplementary Fig. 2b).

Due to its phylogenetic position, the gar genome provides connectivity of TEs in teleost, lobe-finned fish and tetrapod genomes. Comparing gar TE diversity to data from published genomes, we observed a general trend of patchy distribution of different TE families (Supplementary Fig. 2c)[54]. However, some families are completely widespread in vertebrates such as TcMariner, hAT, ERV (Endogenous Retroviruses), L1 and CR1-like (CR1, L2 and Rex1-Babar). We reassessed the lineage-specificity of TE families that have so far been identified 1) only in ray-finned fish and non-tetrapod lobe-finned fish, 2) only in ray-finned fish, and 3) only in teleosts (Supplementary Figs. 2c,3). Two superfamilies, the R2 retrotransposon and the DNA transposon Sola families are specifically present in both lobe- and ray-finned fish. Two DNA transposon superfamilies, MuDr and EnSpm, appear to be specific to ray-finned fish. The Zisupton transposon is a teleost-specific superfamily that has been exapted in tetrapods[55]. Finally, CR1-like is a superfamily composed of CR1 (absent only in teleosts), L2 (absent only in birds) and Rex1-Babar (Rex1 is found only in ray-finned fish; Babar is found in ray-finned fish and frog)[56]. Analysis of the gar genome shows that teleost fish genomes lost CR1, but contain L2, Rex1 and Babar. Spotted gar, in contrast, is the only vertebrate that possesses all four different families of CR1-like (CR1, L2, Rex1 and Babar).

We identified potentially currently active TE superfamilies in gar by searching for TE copies in transcriptome data from eight tissues (brain, eye, heart, kidney, liver, muscle, skin, testis), and

two developmental stages, embryos and larvae. Except for a few superfamilies completely absent from the transcriptome (EnSpm, Helitron, Kolobok, MuDr, Sola, MER6 from DNA transposons, R4, AFC from LINE and SINE), transposons from all classes and superfamilies are active in at least one tissue or developmental stage (Supplementary Tab. 2). The TE transcript content in the different tissues ranges from 2.7% (in muscle) to 3.5% (in brain) (Supplementary Tab. 3).

To evaluate the relationship between TE content in the genome and TE transcription, we plotted the number of TE elements (only superfamilies with more than 100 transcripts) in the transcriptomes against their corresponding genomic copy number (Supplementary Fig. 4). TcMariner is the most highly represented superfamily in the transcriptomes of all tissues examined, followed by CR1 and Rex1-Babar. All TE families were weakly expressed in muscle and heart. In contrast, most TEs were abundantly expressed in brain. Although some TE families were present in high content in the genome (CR1, Ngaro or V), they were not detected in the transcriptome because they may have lost their activity or are expressed in tissues not sampled here. TE copy number in the genome and TE expression are not directly linked. Furthermore, the copy number/expression level disconnection shows that our observations are probably not due to basal transcription.

Two main waves of TE expansion were identified in the gar genome at about Kimura-distance 8 and 25 (Supplementary Fig. 5a). The oldest burst (Kimura-distance 25) is mainly due to the amplification of TcMariner (DNA transposons), CR1 (LINE), Deu, and V (SINE) copy number. Following this burst, a potential increase of Ngaro retrotransposon activity was observed (Kimura-distance 17-18). The second and most recent peak of activity (Kimura-distance 8) corresponds to a general amplification of several superfamilies, especially Rex1-Babar, 5S-SINE and R2 LINE sequences. The graph also shows whether or not TEs are potentially currently active (Kimura-distance close to 0). All TE superfamilies seem to be only weakly active with the exception of TcMariner, which shows more substantial activity. This is concordant with the transcriptome analyses where TcMariner is the most abundant superfamilies (Supplementary Fig. 4).

Using the Kimura distance profiles, comparison of potential TE copy “age” (rather old or recent TE copies) between various vertebrate species might bring interesting clues concerning the rate of TE turnover and possibly DNA elimination in the spotted gar[54]. Supplementary Fig. 5b,c represent the abundance of TE copies according to their divergence with their consensus sequence (Kimura distances; compared to species-specific library consensus). Regarding the profiles of the total TE distribution, the spotted gar genome contains more “ancient” copies than teleost genomes. Indeed, teleosts mostly show recent accumulation of TE sequences, mammals show recent TE accumulation and many older sequences and finally coelacanth presents one large burst of transposition shifted on the left side compared to spotted gar. TE groups (DNA, LTR, LINE and SINE) as well as the most abundant vertebrate superfamilies (CR1-like retrotransposons and TcMariner and hAT DNA transposons) were separately plotted. In most cases, spotted gar peaks are the oldest and located on right side compared to others. These results suggest reduced TE turnover in spotted gar genome, possibly associated with differences in the rate of TE elimination and/or reduced recent TE activity.

Finally, to estimate the abundance of small and/or degenerated TE copies, we filtered our genome-wide TE content analysis and removed TE sequences smaller than 80 nucleotides and

sharing less than 80% of identity with the reference sequence (Supplementary Tab. 1). The TE content strongly decreased when applying this filter. While the total TE content is about 20% of the genome, the filtered TEs only made 9.7%, suggesting that the spotted gar genome contains a significant proportion of small sequences that are either degenerated copies or artefacts of TE library construction or genome assembly. This conclusion is consistent with the ancient bursts of transposition (sequences are probably in the process of elimination) and a recent decrease of TE activity (Supplementary Fig. 5).

6. Phylogenomic analysis

Methods. Analysis of phylogenies including gar is based on the carefully built 251 protein-coding gene alignments of 22 species used for the coelacanth phylogenomic analysis[57] with the addition of gar and several other species, including the orthologous sequences of spotted gar, Western painted turtle[58] (NCBI BioProject PRJNA210179), and bowfin (Supplementary Note 3.3).

Orthology was assigned through HaMStR v13.2.2 software[59] using coelacanth as reference taxon. For spotted gar, orthology assignment was conducted using Ensembl and MAKER gene models independently. Using the Ensembl gene models, we discovered 14 more genes than using MAKER. We further searched the Ensembl orthology relationships for uncertain orthology assignments and excluded questionable genes, keeping 243 of 251 genes. Concerning the excluded genes, seven had a '1-to-many' orthology relationship and one had no candidate gar orthologous sequence that passed HaMStR criteria. Orthologous sequences of painted turtle and bowfin were identified for 241 and 235 genes respectively.

The spotted gar, painted turtle and bowfin sequences were added to the existing alignments[57] with MAFFT v7.050b[60]. Sites with gaps in more than 70% of sequences were deleted with a custom perl script. Finally, the alignments were combined in a supermatrix with FASconCAT v1[61] and input into GBlocks v0.91b[62] to exclude possible misaligned regions. The final alignment consisted of 97,794 amino acid sites and can be found in Supplementary File 1.

Phylogenetic reconstruction was carried out in RAxML 8.0.19[63] with JTT+F+Γ4, which was the model selected through RAxML script *ProteinModelSelection.pl*. Another round of phylogenetic reconstruction was performed using PhyloBayes MPI v1.5[64] with the CAT+GTR+Γ4 model[65] (Fig. 1b). Two chains were run for 10,000 cycles each. The posterior consensus tree was obtained with exclusion of the first 1,000 trees as burn-in and sampling every ten trees from both chains. Node support was evaluated with 100 bootstrap replicates (BS) for RAxML and posterior probability (PP) for Phylobayes run.

To explore further the robustness of any given node of the reconstructed phylogeny, we calculated the internode certainty (IC) of each node and the extended IC (ICA). IC and ICA metrics were developed to inform the certainty of a phylogeny's internode based on a given set of trees[66]. IC is calculated based on the most prevalent conflicting bipartition at the trees set, while for calculating ICA all prevalent conflicting bipartitions are taken into account. The IC calculation was conducted on the JTT+F+Γ4 phylogeny using the individual genes phylogenies as the tree dataset to quantify incongruence[67]. Individual gene phylogenies were estimated in RAxML using the same model as the one used in the main phylogeny. The 154 gene

alignments that included sequence information for all 25 species were included in the calculation of IC values in RAxML (option '-f i').

Results. Morphological analyses favor bowfin (*Amia calva*) as the sister lineage to teleosts with gars diverging basally, but molecular phylogenetics support a monophyletic clade of *Holostei* (gars plus *Amia*) as the sister lineage to teleosts[9,12] (Supplementary Fig. 1e). Our phylogenomic analysis here revealed the overall expected vertebrate phylogenetic patterns and suggested strongly the monophyletic relationship of spotted gar and bowfin, i.e. for the holostean lineage. This is congruent with previous results based on 10-20 nuclear markers[68-70] and ultraconserved elements of uncertain orthology[71].

Maximum likelihood (Supplementary Fig. 6) and bayesian analyses (Fig. 1b) both supported holostean monophyly with maximum support. All branches of the phylogeny are supported with 100 Bootstrap value and Posterior Probability of 1, except for branches defining the position of lungfish and armadillo. The CAT+GTR+Γ_4 analysis fully supports the sister relationship of lungfish to tetrapods, but the JTT+F+Γ_4 analysis did not (BS=67; see ref.[57] for conclusive analysis/discussion on that topic). Regarding the position of armadillo, the difficulty in resolving this question is known from recent phylogenomic analyses that came up with different conclusions[72,73]. In the present analysis, use of the same dataset with different models of evolution supported alternative relationships.

To assess further the support of each internode of the reconstructed phylogeny and evaluate to what extent individual genes conformed to the final result, we estimated the internode certainty values (IC and ICA) for each internode of the JTT+F+Γ_4 reconstructed phylogeny based on a tree dataset produced from the individual genes (Supplementary Fig. 6). Results showed that most of the bipartitions of the main phylogeny did not exhibit remarkable conflict with the individual gene trees. The holostean monophyly was highly supported (IC=0.36), confirming limited conflict on that internode. High conflict was observed in the position of lungfish (IC=0.01) and armadillo (IC=-0.02), showing strong alternative scenarios in both cases. Especially for armadillo, the negative value of IC shows that the most frequent topology within the individual gene trees dataset is not the one recovered in the main phylogeny.

7. Molecular rate analyses

Methods. To study the rate of evolution of the gar genome with respect to other vertebrates, we conducted relative rate analyses at the alignment level with Tajima's tests[74] and at the level of the reconstructed phylogeny with Two-Cluster tests[75]. Tajima's tests were run on the concatenated alignment of all genes with a custom perl script. At the phylogeny level, we performed Two-Cluster tests on both JTT+F+Γ4 and CAT+GTR+Γ4 phylogenetic trees. For the best JTT+F+Γ4 tree, we calculated all pairwise distances between taxa in the main tree and the 100 bootstrap trees using the R package *ape*[76]. From the pair-wise distances, we calculated the mean distance of each monophyletic cluster to the outgroup cluster. Then, for each pair of clusters, we evaluated whether the difference of the mean distance to the outgroup cluster (cartilaginous fishes) was significantly different from 0 with z-statistics. For each comparison, we estimated the variance of the difference from the bootstrapped trees (see ref.[57] for more detailed

description). Similar analyses were conducted for the analysis on the CAT+GTR+Γ4 tree, but here variance was estimated from trees sampled from the two chains (400 trees sampled in total with 1,000 trees burn in and sampling every 45 trees in both chains).

Results. The relative rate analysis included a comparison of species at the sequence alignment level (Tajima's tests) and at the reconstructed phylogeny level (Two-Cluster test). Tajima's tests (Supplementary Tab. 4a) revealed that the spotted gar evolved slower than teleost lineages, but not slower than other species lineages examined. Results showed that the bowfin lineage evolved slower than mammals and at a rate equal to that of anole lizard, but faster than chicken, turtle or coelacanth, which is the slowest evolving species in this analysis.

Two-Cluster tests (Supplementary Tab. 4b) revealed a picture similar to Tajima's tests. They were conducted based on both maximum likelihood- and bayesian-reconstructed phylogenies. Coelacanth had the shortest distance to the outgroup cluster of chondrichthyes, while teleosts had the longest. Holostei had significantly smaller distance to the outgroup compared to all other groups, except coelacanth and birds+turtle. For the latter comparison, the birds+turtle lineage is slower than holostei in the CAT+GTR+Γ_4 tree, but it had no significant difference in the rate of evolution in the JTT+F+Γ_4 analysis. The opposite pattern was observed in the comparisons of holostei versus sarcopterygii (lobe-finned vertebrate lineage) and tetrapods. In both cases, holostei were slower in the JTT+F+Γ_4 analysis, but the difference becomes non-significant at the CAT+GTR+Γ_4 tree. Finally, at a broader scale, results showed that Neopterygii (Holostei+teleosts) evolved faster than sarcopterygii or even tetrapods, driven by the long branches of teleosts.

8. Evolution of vertebrate genome structure

8.1 The spotted gar karyotype

Methods. Four specimens of unidentified sex were used for cell cultures and chromosome preparations according to Amores and Postlethwait (1999)[77] with small modifications. Chromosome spreads were obtained from fibroblast cells growing from caudal fin cultures. Cells were grown at 24°C for up to two weeks before harvesting. Metaphases were prepared by air-drying technique and ten metaphase plates were scored for each individual.

Results. Chromosome counts from all four individuals revealed the same diploid chromosome number of 2N = 58. The karyotype consisted of 17 pairs of metacentric-submetacentric chromosomes, 3 pairs of telocentric-subtelocentric chromosomes and 9 pairs of very small chromosomes (Fig. 2a, Supplementary Fig. 7). The number of chromosome pairs is in agreement with the described number of linkage groups in the spotted gar genetic linkage map[2]. Previous studies in other gar species have described 56 chromosomes in both longnose gar (*Lepisosteus osseus*)[78] and tropical gar (*Atractosteus tropicus*)[79] with similar composition of the karyotypes, where the main differences are due to scoring the chromosomes in different stages of contracting.

The presence of very small chromosomes or microchromosomes is shared with other ancient groups like sturgeons[80], cartilaginous fish[80], coelacanth[81] as well as amphibians and reptiles[82,83], suggesting that the presence of microchromosomes is an ancestral vertebrate condition. In contrast to those species, the holostean sister lineage of gars, the bowfin (*Amia calva*) has no described microchromosomes[80], suggesting that its karyotype is more derived compared to the gar species.

8.2 Synteny analyses

Methods. Circos plots[84] of gar linkage groups *vs.* chromosomes of human (Fig. 2b) and chicken (Fig. 2c) were based on 10,809 and 9,937 orthologs (Homology type: ortholog_one2one; Orthology confidence: 1, high), respectively, obtained from Ensembl75 through Biomart[85]; the Circos plot comparing gar *vs.* medaka (Fig. 2e) was based on the combination of 9,059 one-to-one orthology relations (gar to medaka 'TGD singletons') and 1,279 one-to-two orthology relations (gar to medaka 'TGD ohnologs') obtained as described in Supplementary Note 13.1 below. Genome sizes were scaled to consume half of the Circos plots, which were further optimized in chromosome order and orientation to reduce the number of crossing lines independently for gar *vs.* tetrapods (human/chicken) and for gar *vs.* teleost (medaka) comparisons.

Dotplots of gar linkage groups *vs.* medaka, chicken, and human (Supplementary Fig. 8) were generated with the Synteny Database[86] using Ensembl74 gene annotations.

Comparative synteny maps (Supplementary Fig. 9) were generated as previously described[57,87]. Briefly, the complete set of annotated gar genes was aligned to each of six repeat-masked reference genomes (chicken – galGal4, human – hg19, medaka - oryLat2, puffer fish - fr3, stickleback - gasAcu1 and zebrafish - danRer7: via *tblastn*[27]). Alignments were processed to merge adjacent exons and orthologous relationships were defined using previously described algorithms[88]. For the purpose of this study, a locus from the reference genome is considered to be orthologous to a gar gene if: 1) the alignment bitscore between the gar gene and a given locus from the reference genome is within 90% of the best alignment bitscore for that gene, 2) there are six or fewer paralogs detected in the gar genome, and 3) there are six or fewer paralogs detected the reference genome.

Counts of orthologs on all pairwise combinations of gar linkage groups and reference chromosomes were tabulated and compared to expected values based on random sampling of loci from a set of linkage groups and chromosomes with the same number of loci and loci per chromosome/linkage group. Because these comparisons involve a large number of pairwise combinations, which often possess a small number of putative orthologs (i.e. most cells correspond to non-orthologous chromosomes, especially in comparisons between species that have experienced few rearrangements), the distribution of orthologs was evaluated using a chi-square incorporating Yates' correction for continuity and Bonferroni corrections for multiple testing.

Rates of chromosomal fission/translocation (relative to gar) were estimated by counting the number of statistically-significant syntenic regions in excess of the number of gar linkage groups (N = 29), or 2x the number of gar linkage groups (i.e., N = 58) to account for the teleost genome duplication (TGD). Fusions (relative to gar) are not expected to change the number of

conserved syntenic segments per gar chromosome, whereas fissions, interchromosomal translocations and chromosomal/whole genome duplications will increase the number conserved segments per gar chromosome.

Evolutionary ancestry of vertebrate microchromosomes. Previous studies provided evidence that many chicken microchromosomes represent distinct evolutionarily conserved entities that were present at least as early as the last common tetrapod ancestor[80,89,90]. The gar chromonome provides the opportunity to further test for evolutionary conservation of microchromosomes, and refine our understanding of chromosome structure and counts (karyotype) of the ancestor of all extant bony vertebrates.

Aligning annotated gar genes to the genomes of several vertebrate species (human, chicken and four teleost fish: zebrafish, pufferfish, medaka, and stickleback) revealed conservation of orthologous segments in all species (Fig. 2, Supplementary Fig. 8,9). Strikingly, comparisons between gar and chicken revealed strong conservation of large chromosomal segments and – remarkably – several entire chromosomes (Fig. 2c, Supplementary Fig. 8'',9a). Given the similarity of chicken and gar genomes, it seems likely that both genomes approximate the structure of the ancestral bony vertebrate genome: the two genomes differ by only approximately 17 large fissions, fusions or translocations since the two species shared a common ancestor about 450 million years ago.

Several pairs of chromosomes showed a nearly one-to-one relationship in the gar/chicken comparative map (Fig. 2c, Supplementary Fig. 8'',9a), including two macrochromosomes (gar linkage group 2, Loc2, and chicken chromosome Z, GgaZ, as well as Loc12/Gga7) and twelve smaller chromosomes including microchromosoems (Loc14/Gga9, Loc23/Gga11, Loc13/Gga14, Loc20/Gga15, Loc21/Gga17, Loc22/Gga19, Loc18/Gga20, Loc25/Gga21, Loc26/Gga24, Loc24/Gga25, Loc15/Gga27 and Loc19/Gga28). These patterns would seem to suggest that several avian microchromosomes are directly descended intact from similarly small chromosomes that were already present in the common ancestor of chicken and gar. To further assess the degree of evolutionary conservation of these chromosomes, we compared the relative sizes of orthologous gar and chicken chromosomes. Assembly lengths of orthologous chromosomes are highly correlated (R^2 = 0.97 for all fourteen chromosomes; R^2 = 0.73 for the twelve smaller chromosomes including microchromosomes; Fig. 2d). We interpret the overall similarity in both size and gene content of orthologous gar and chicken chromosome as strong evidence that the karyotype of the bony vertebrate ancestor possessed both macro- and micro-chromosomes, many of which are preserved in the genomes of both gar and chicken.

Karyotype evolution in teleosts after divergence from gar. Comparisons of gar and teleosts reveal the overarching influence of the TGD, with pairs of teleost chromosomes being composed of similar subfractions of gar chromosomes. One notable difference between gar (n = 29) and sequenced teleosts is that teleosts generally possess fewer chromosomes (n~24-25) despite the TGD. Paralogs of two or more gar chromosomes often co-occur on the same pair of teleosts chromosomes (e.g., in medaka, Supplementary Fig. 8'; in stickleback, Supplementary Fig. 9b), such as gar chromsomes Loc2, Loc20, and Loc21 and medaka chromosomes Ola9 and Ola16 (Fig. 2f). See Supplementary Fig. 10 for additional examples. Therefore, differences in chromosome number between gar and other teleosts appear to be largely explained by

chromosome fusion events that occurred in the teleost lineage after it diverged from the gar lineage but before the TGD (as exemplified in Fig. 2f).

The strong evolutionary conservation of the gar genome also provides the opportunity to estimate rates fission and translocation in teleost and gnathostome lineages. Given that the gar genome serves as a reasonable proxy for the ancestral pre-TGD genome, we examined patterns of interchromosomal rearrangements that occurred on branches between chicken and gar and between gar and sequenced teleosts. Comparisons revealed an average fission/translocation rate of $1.11x10^{-3}$ rearrangements per million years for gar/chicken branches and similar rates for comparisons involving stickleback ($1.28x10^{-3}$), pufferfish ($1.42x10^{-3}$) and medaka ($1.33x10^{-3}$), after accounting for the influence of the TGD (Supplementary Fig. 11). Compared to other teleost lineages, anaylses involving zebrafish indicate an increased fission/translocation rate, averaging $2.21x10^{-3}$ rearrangements per million years.

These comparisons indicate that the TGD itself might not underlie changes in rearrangement rates. In contrast, our analyses indicate that the TGD was preceded by a relatively large number of fusions in the ancestral pre-TGD, post-Holostei teleost lineage and reveal that some lineages (e.g. zebrafish) experienced higher rates of fission/translocation over a more recent evolutionary timeframe.

9. Gene family analyses

9.1 Hox clusters

Methods. Hox genes were identified and analyzed as describe in ref.[91].

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

The *hoxA6* and *hoxD2* genes appear to have been lost in the common ancestor of all teleosts following divergence from gar. The loss of *hoxA6* seems to be unique to teleosts, while *hoxD2* has been lost at least twice – once in teleosts as well as in the last common ancestor of lobe-finned fish. Additionally, extensive lineage-specific gene losses of at least one *hox* gene duplicate characterizes the evolution of teleosts following the TGD with the result that most teleosts so far examined do not have significantly greater numbers of *hox* genes than gar (zebrafish 49, stickleback 46), and no teleosts have been found to maintain all 82 *hox* genes thought to be present in their last common ancestor immediately following the TGD. Reconstructed from genes observed in extant teleosts, the last teleost common ancestor (LTCA) had 74 *hox* genes (Supplementary Fig. 12).

The protein-coding gene complement of the spotted gar *hox* clusters is also remarkably similar to the inferred complement of the last bony vertebrate (osteichthyan) common ancestor (LOCA). Only the *hoxA14* and *hoxD14* genes have been lost in the gar lineage since the ancient

divergence between ray-finned (actinopterygian) and lobe-finned (sarcopterygian) fish ca. 450 million years ago. An intact *hoxA14* gene has to date been reported only in coelacanth and thus appears to have been lost at least three times since the last common ancestor of all jawed vertebrates – once in the lineage leading to cartilaginous fish (chondrichthyans), once in the lineage leading to tetrapods, and once in the lineage leading to ray-finned fishes. An intact *hoxD14* gene has been isolated in chondrichthyans including the elephant shark, horn shark, and lesser spotted catshark and has recently been reported in the paddlefish[92], a ray-finned fish diverging basally to spotted gar. No indication of *hoxD14* has been found in any teleost to date, but a discernible *hoxD14* pseudogene is found in the gar genome located in a conserved position between *hoxD13* and *evx2* (Supplementary Fig. 13a), and with three regions of similarity corresponding to the three exons of other vertebrate h*oxD14* genes (Supplementary Fig. 13b). The presence of a pseudogene in gar but not in teleosts suggests that this gene could have been inactivated independently in the two lineages, and more recently in the lineage leadings to gars. Alternatively, the presence of the *hoxD14* pseudogene presence might reflect the slower rate of evolution of gar hox clusters relative to teleosts. No *hoxD14* transcript was detected in either the gar or bowfin transcriptomes. Taken together, this suggests that the *hoxD14* gene was lost independently at least two times in evolution – once in the common ancestor of all lobe-finned vertebrates, and once in the last common ancestor of teleosts and holosteans (gars and bowfin) with slow sequence evolution in gars. Alternatively, *hoxD14* was lost three times, i.e. in lobe-finned vertebrates and independently in teleosts and in gars.

9.2 ParaHox clusters

Methods. ParaHox genes were identified and analyzed as describe in ref.[93].

Results. The ParaHox genes are a developmentally important family of homeodomain-containing transcription factors, with vital roles in the formation of a number of endoderm-derived structures such as the pancreas and intestine[93]. We find a total of seven ParaHox genes in the gar genome, comprising *gsx1* and *gsx2*, *cdx1*, *cdx2* and *cdx4* and *pdx1* and *pdx2* (Supplementary Fig. 14).

The discovery of *pdx2* is particularly significant as it has until now only been known from cartilaginous fish and coelacanths and was proposed to have been lost relatively rapidly following the divergence of the ray-finned and lobe-finned fish lineages[94,95]. The phylogenetic position of spotted gar with respect to the whole genome duplication known to have occurred prior to the evolution of teleost fish[1], coupled with the known effects of this duplication on the composition and organization of ParaHox genes in teleosts[94,96,97] suggests that this gene is more likely to be a common characteristic of other non-teleost ray-finned fish (although it is absent from the bowfin transcriptome). It therefore seems likely that the teleost genome duplication had a far greater effect on the complement and organization of ParaHox genes than was previously suggested, with the loss of both *cdx2* and *pdx2* and the break-up of the canonical ParaHox gene cluster.

VISTA plot analyses reveal the complement of conserved non-coding elements (CNEs) of gar ParaHox clusters (Supplementary Fig. 14, Supplementary Tab. 5) including the conservation of

a non-coding element conserved among gar and coelacanth upstream of *pdx2* (Supplementary Fig. 15), where so far no regulatory element has ever been identified.

9.3 Aldh1a gene family

From conserved synteny analyses using the Synteny Database[86] and Maximum Likelihood phylogenetic analysis, we infer that aldh1a1 is an ohnolog gone missing (OGM) in teleost species (i.e. zebrafish, stickleback, medaka, tetraodon, fugu) (Supplementary Fig. 16).

The phylogeny was generated using PhyML 3.0 subjected to WAG substitution model, a bionj starting tree improved by NNI, and computing aLRT SH-like branch support[98]. Multiple protein sequence alignment was made by MUSCLE implemented in Aliview[99], corrected manually for inconsistencies and trimmed the N- and C-termini (i.e. until positions 24 and 478 of HsAldh1a1) to decrease the total number of gaps. Sequence reference numbers are those used in refs.[100,101], for gar LoAldh1a1, ENSLACP00000021870; LoAldh1a2, ENSLACP00000013047; LoAld1a3, ENSDARP00000055592; and for coelacanth LcAldh1a1, ENSLACP00000021870; LcAldh1a2, ENSLACP00000013047; LcAldh1a3, ENSLACP00000020368.

9.4 Circadian clock genes

The time-keeping mechanisms that drive circadian rhythms and their regulatory genes are highly conserved across taxa[102-104]. To help understand the origin of teleost fish circadian rhythm genes and how they link to the human genome, we analyzed spotted gar circadian clock genes, as described previously[104,105].

Methods. We identified gar circadian clock genes by interrogating the spotted gar draft genome sequence with the sequences of zebrafish circadian clock genes. Phylogenetic trees of 11 families of spotted gar circadian clock genes were then constructed by neighbor-joining (NJ) using MEGA6[106].

Results. Analysis revealed that gar has 25 circadian clock genes from 11 families (Supplementary Tab. 6). Phylogenetic analyses showed that 8 of these 11 families of spotted gar circadian clock genes (including *bmal, clock, period, nr1d, dec, ror, csnk1e,* and *timeless*) have the same number of genes in human and gar (Supplementary Fig. 17). For these genes, we conclude that a single gene copy existed in the last common ancestor of human and ray-finned fish and that the phylogenies are as expected if the gar and human genes are orthologs. Histories of the remaining three families are less straightforward.

***cry* gene family** (Supplementary Fig. 17c): Spotted gar has four copies of the *cryptochrome* gene family. Gar has two *cry1* genes, *cry1a* and *cry1b*, which are closely linked (only four genes between them) and are likely derived from local (tandem) duplication in the ancestor of gar and teleosts; they appear to be co-orthologs of mammalian *Cry1*. In addition, gar has a single ortholog of *cry2*[105] and a *cry3* gene, which is also present in teleosts, amphibians, reptiles, and

birds, but which has gone missing from mammals[105]. Zebrafish contains four *cry1* genes, *cry1aa, cry1ab, cry1ba* and *cry1bb*, that are likely derived from the TGD, as well as *cry2* and *cry3*[105].

***nfil3* (*e4bp4*) gene family** (Supplementary Fig. 17e): Evolution of the *nfil3* (*e4bp4*) gene family was complex. Spotted gar contains at least three paralogs currently annotated as *nfil3* (*e4bp4-1*), *nfil3-2* (*e4bp4-2*) and *nfil3-6* (*e4bp4-3*). Phylogenetic analysis shows that *nfil3* (*e4bp4-1*) is the ortholog of mammalian *Nfil3* (*E4bp4*), and that teleosts have a single copy of this gene. These data along with conserved synteny analysis show that these genes are orthologs.

The other two gar *nfil3* family members, *nfil3-2* and *nfil3-6,* are adjacent to each other and likely result from a tandem duplication event. Humans and other eutherian mammals lack orthologs of these two genes, but because orthologs are present in coelacanth (ENSLACG00000009977 and ENSLACG00000009125) and non-mammalian tetrapods (lizard, ENSACAG00000029030 and frog, ENSXETG00000007922) according to both phylogenies and conserved syntenies, we conclude that the last ancestor of bony vertebrates had the *nfil3-2_nfil3-6* gene pair but that this pair was lost in mammals.

Teleosts have additional *nfil3* family members. Most teleost genomes have a duplicate of *nfil3-2* (currently called *nfil3-5*) and a duplicate of its neighbor *nfil3-6* (currently called *nfil3-3*) that arose after the teleost lineage diverged from the gar lineage. Conserved synteny analysis by the DotPlot function of the Synteny Database[86] support an origin of the tandem neighbor pair *nfil3-2_nfil3-6* and the tandem neighbor pair *nfil3-5_nfil3-3* in the TGD. We propose that the nomenclature of these genes should reflect historical relationships: the old names imply that gar *nfil3-2* is THE ortholog of teleost *nfil3-2,* but this is in error because the teleost genes with the old names of *nfil3-2* and *nfil3-5* are co-orthologs of the gar gene. Accordingly, the two gar tandem duplicates would be renamed as *nfil3-2.1* (old *nfil3-2*) and *nfil3-2.2* (old *nfil3-6*) to reflect that they represent a second member of the *nfil3* family (after the human gene called *NFIL3*) and are tandem duplicates. This assignment makes the four zebrafish genes become *nfil3-2.1a_nfil3-2.2a* (old *nfil3.2_nfil3.6*, respectively) and *nfil3-2.1b_nfil3-2b* (old *nfil3-5_nfil3-3*). In addition, teleosts have an *nfil3* family member currently called *nfil3-4* (ENSDARG00000092346 in zebrafish) that does not appear to have an ortholog in gar or in most lobe-finned vertebrates, although duck and sheep have genes (ENSAPLG00000002454 and ENSOARG00000016478, respectively) more closely related to the teleost *nfil3-4* gene in sequence than to other tetrapod genes.

The most parsimonious explanation of these data is that an ancient *nfil3* gene duplicated in the VGD1 and VGD2 events to make four copies: 1) one copy was lost in all extant lineages; 2) one copy became what is called *NFIL3* in human and *nfil3* in coelacanth, gar, and teleosts (and could be thought of as *nfil3-1*); 3) another copy became *nfil3-4*, which was lost in lineages except for teleosts (and possibly duck and sheep); and 4) the final copy tandemly duplicated before the divergence of lobe-finned and ray-finned vertebrates to produce the tandem duplicates we would call *nfil3-2.1* and *nfil2.2*, which are now present in gar and coelacanth genomes, and then the tandem duplicate was duplicated in the teleost genome duplication to form *nfil3-2.1a_nfil2,2a* and *nfil3-2.1b_nfil2.2b*.

***par* gene family** (Supplementary Fig. 16f): In the *par* gene family, spotted gar has one *tef* and one *hlf* gene like human (both genes with TGD ohnologs). The gar ortholog of *DBP* is present as an unannotated sequence on an unassembled scaffold (JH591448.1: 146,984 to 147,190) in the current version (LepOcu1) of the spotted gar genome (reciprocal best blast hit with coelacanth DBP and nearest four neighbors on one side being orthologs of the four nearest neighbors of zebrafish *dpba*).

In sum, these analyses support the notion that the gar genome retained several circadian clock-related genes that originated in the early vertebrate genome duplication events that subsequently went missing from the human genome, and that in addition, gar lacks gene duplicates that arose in teleosts after the TGD. These attributes make gar a superior experimental model for understanding the genetic nature of circadian rhythm regulation in the last common ancestor of all bony vertebrates.

9.5 Opsin genes

Methods. Gar opsin genes were identified by using tBlastn[27] with amino acid query sequences from Japanese eel (*Anguilla japonica*), zebrafish (*Danio rerio*), barfin flounder (*Verasper moseri*), guppy (*Poecilia reticulata*) and medaka (*Oryzias latipes*) (Supplementary Tab. 7). In addition, we searched for visual opsin pseudogenes by blastn surveys with reduced word size values and reduced mismatch penalty scores, but none were detected. The evolutionary history of opsins was inferred with the Maximum Likelihood method based on the JTT matrix-based model[107] using MEGA6[106]. The analysis involved 232 amino acid sequences. All positions with less than 95% site coverage were eliminated, leaving a total of 196 positions in the final dataset (Supplementary Fig. 18).

Results. A combination of old (shared) and more recent (lineage-specific) duplication, divergence and loss events have generated enormous diversity among fish opsin gene repertoires[108], including spotted gar that shows similarities to both teleost and tetrapod repertoires.

Visual opsins: The spotted gar genome contains seven visual opsin type genes[109] (Supplementary Tab. 7, Supplementary Fig. 18). The gar LWS and SWS2 opsins occur on LG1 and are approximately 10kb apart. RH2 occurs on LG3 and RH1/RHO occurs on LG5. Gar RH1/RHO has no introns, an observation that places its origin by retro-duplication in the common ancestor of gar and teleosts whose RH1 is also intronless[110,111].

Gar-specific SWS1 duplicates occur on LG8 (approximately 5kb apart) and are therefore most likely products of yet another opsin gene tandem duplication (see ref.[108]). To date, the only other known fish species with duplicated SWS1 genes is the Ayu smelt (*Plecoglossus altivelis*)[112]. For each visual opsin, amino acid sites known to influence spectral sensitivity have been characterized[113]. For the gar SWS1 gene pair there are 14 spectral tuning sites, amino acid positions 46, 49, 52, 86, 90, 91, 93, 97, 109, 113, 114, 116, 118, and 298[113,114]. The gar SWS1 gene pair differs at six of these amino acid positions; SWS1A encodes the 14 key-site

haplotype, _ _ _ **F** A_**T**_I_**A**V_ _, and SWS1B encodes the amino acid key-site haplotype, _ _ _ A S_P_V_G S_ _ (where an underscore represents the same residue in both proteins). F86, T93, and A114 (bold/red font above) are associated with UV sensitivity and alternative amino acids at each position tend to shift sensitivity into the violet region of the spectrum[113]. Consistently, UV-sensitive cones have been described for longnose gar (*Lepisosteus osseus*)[19] suggesting UV-sensitivity in *Lepisosteus* gar species.

Non-visual opsins: A total of 17 'non-visual' opsin genes from ten subfamilies are found in gar, including a gar-specific duplication of Teleost Multi-Tissue opsin I (TMT1) (Supplementary Tab. 7, Supplementary Fig. 18). The most fascinating result of this non-visual opsin survey is the discovery of a gar pinopsin gene. Until now pinopsin, which is one of several pineal-expressed photoreceptors that regulate the rhythmic production of melatonin and thereby regulate circadian rhythm, was known only from birds, reptiles and amphibians[115]. Parietopsin and parapinopsin are other pineal-expressed opsins. Interestingly, parietopsin (reported in other fish including zebrafish, but not restricted to teleosts[115]) is the only non-visual opsin not found in gar. As described above, gar has an ortholog of extra-ocular-rhodopsin (or exo-rhodopsin), the RH1/RHO opsin progenitor gene[111]. Exo-rhodopsin was thought to play the same role in ray-finned fish that pinopsin does in non-mammalian tetrapods[116], yet the two genes co-occur in the gar genome. It will be necessary to characterize the expression domains of these genes in gar to determine whether or not exo-rhodopsin is likely to have made pinopsin redundant in other ray-finned fishes. RH1/RHO and exo-rhodopsin are linked in gar, so it seems possible that the derived intronless RH1/RHO gene might have used ancestral regulatory modules immediately following retroduplication. This could explain how it was able to take over the visual role of its intron-containing progenitor.

9.6 Spotted gar and the evolution of vertebrate immunity

MHC (major-histocompatibility complex) genes of two different classes are tightly linked in tetrapods and cartilaginous fish, but MHC class I genes are unlinked to most MHC class II genes in teleost fish[117]; this situation raises the question of whether the organization of MHC genes in teleosts characterizes ray finned fish in general or arose subsequent to the teleost genome duplication (TGD). Phylogenetic analyses of our transcriptomic and genomic data from spotted gar complement recent partial genomic analyses[118,119] by categorizing MHC class I and class II loci into recognizable clades (Supplementary Fig. 19-20). In the gar genome sequence, one MHC class I alpha chain gene is present on LG14, but no MHC class II genes are located nearby in the current assembly (Supplementary Fig. 21). Adjacent to this MHC class I gene lie *ece2* and *psmd2,* the orthologs of which reside on human chromosome 3 (Hsa3), unlinked to MHC genes; note, however, that paralog *PSMB8* is surrounded by MHC class II genes in human, suggesting an ancient linkage before the vertebrate genome duplication events VGD1 and VGD2. Most gar MHC class I and class II genes lie on currently unassembled scaffolds, none of which have a mapped genetic marker (Supplementary Fig. 21), which frustrates a clear answer to the question of whether class I and class II genes are linked in gar. Only one scaffold (JH591501) has an MHC class I alpha chain gene linked to MHC class II genes in gar[118,119]

(Supplementary Fig. 21b), suggesting that such linkages may be ancestral in ray-finned fish and were lost after the TGD. At least two MHC class I alpha chain genes are present on the same scaffold (JH591545) with *psmb8* (Supplementary Fig. 21b), which is present in the human MHC class II region and which encodes a component of the immunoproteasome, a complex that processes MHC class I-restricted T cell epitopes[120]. Although the location of *PSMB8* embedded among MHC class II genes on human chromosome 6 is consistent with the linkage of MHC class I and class II regions in gar, the location of *psmb8* in zebrafish is adjacent to MHC class I genes but not linked to MHC class II genes[121,122]. These lines of available evidence suggest that MHC class I and class II genes may be linked in gar, but a definitive answer awaits the mapping and placement of the other MHC gene-bearing scaffolds. See Supplementary Tab. 8 for the genomic locations of gar genes similar to those of the mammalian MHC region and its paralogous regions[123]. MHC I and class II genes are diverse in gar: gar has some class I genes previously thought to be teleost-specific (Z/P-, L-, and U/S like-lineages (e.g.[122,124])), and class II alpha and beta chain genes, some similar to and some distinct from teleost DA/DB and DE lineages (Supplementary Fig. 19-20). Additional genes encoded within the human MHC region are involved in immune responses (e.g., complement factors like *CFB)*, inflammatory responses (e.g., *TNFSF1*), and processing of class I MHC peptides (e.g., *PSMB8*)[125]. Although few gar orthologs of these MHC region genes are assembled into chromosomes, several (*tapbp, psmb8, pbx2, cfb, tnf*) are on scaffolds with conserved synteny to the human MHC class II or III regions. In gar, complement factor *C2* is on a scaffold with conserved synteny to Hsa11 and at least one of the two gar *C4* genes has neighbors with orthologs also on Hsa11 (Supplementary Tab. 8).

Immunoglobulin (Ig) genes and transcripts in gar generally resemble those of teleosts. Sequences encoding gar IgM and IgD are on LG5 (Supplementary Fig. 22a). Although gar and teleost IgM sequences are similar, gar IgD has 12 non-tandemly duplicated constant domains; in contrast, teleosts generally have fewer (e.g., flounder IgD has 7 constant domains with no duplications; fugu IgD has 13 constant domains, of which 6 are duplicated; salmon, halibut and catfish IgD have 10 constant domains, of which 2, 3 and 4 are duplicated, respectively). Unexpectedly, gar has a second, distinctly different IgM locus on Scaffold JH591415.1 that encodes three constant domains, a single transmembrane domain, and two C_H domains of a lambda light chain-like gene (Supplementary Fig. 22b). The gar genome assembly lacks the IgT (called IgZ in zebrafish) sequences found in the teleost IgM/IgD locus[126,127], suggesting a teleost novelty.

T-cell receptor genes encoding TCRα and TCRδ are tightly linked (on LG24) as they are in mammals. Gar, like *Xenopus*[128], possesses two TCRα and two TCRδ chains, but unlike *Xenopus*, where these sequences are nested within VH genes, gar chain genes are located downstream of V and J segments (Supplementary Fig. 23).

Toll-like receptors (TLRs) are pattern recognition receptors involved in both immune function and development and are conserved from insects to mammals[129]. TLRs share in common the TIR (toll-IL-1 receptor) domain. Six major families of TLRs have been described (TLR1, TLR3, TLR4, TLR5, TLR7 and TLR 11), each of which can contain multiple subfamilies (e.g., TLR7,

TLR8 and TLR9 subfamilies are all part of the TLR7 family)[130]. The human genome contains ten *TLR* genes (*TLR1-TLR10,* lacking TLR11, TLR12, and TLR13) and mouse possesses 12 TLR genes (lacking TLR10 but having TLR11, TLR12 and TLR13). Teleost genomes can possess more than 20 TLR genes[131]. Teleost-specific gene duplications and subsequent divergence of TLRs contribute to this increased number. For example, zebrafish and catfish contain duplicated copies of genes encoding TLR4, TLR5 and TLR8 and at least six other "non-mammalian" TLR genes[131-134].

The gar genome possesses representatives of all six major families of TLRs as predicted by RNA-Seq and Ensembl's automated gene annotation pipeline and confirmed through phylogenetic analysis of Toll/interleukin-1 receptor (TIR) domains (Supplementary Fig. 24). Twelve of 17 gar *tlr* paralogs were annotated using known zebrafish or human TLR-encoding genes, while five genes remain uncharacterized (Supplementary Fig. 24). The TLR1 family has been described as displaying more species-specific adaptations than the other TLR families[130], and the gar genome contains six members of the TLR1 family in contrast to the two to four TLR1 family members in zebrafish, chicken, mouse, and human. As observed in tetrapods, gar possesses a single copy of TLR4 (ENSLOCG00000003751) and TLR5 (ENSLOCG00000018000), whereas teleosts contain tandem copies of these genes. In contrast, TLR8 is duplicated in both teleosts and gar (ENSLOCG00000009990 and ENSLOCG00000009982). Although most gar TLRs share more sequence similarity with teleost TLRs as expected from their shared history, in a phylogenetic analysis, the sequence of gar TLR5 groups more closely with tetrapods than it does to teleosts (Supplementary Fig. 24). Phylogenetic analysis of the TIR domains of TLRs from tetrapods, teleosts, and gar reveals that gar TLRs share evolutionary histories of both teleosts and tetrapods (Supplementary Fig. 24).

NITR (novel immune-type receptor) genes, which function in allorecognition and were thought to be teleost specific[135,136], we find to be present in the gar genome (Supplementary Fig. 25). NITRs typically come from several multi-gene, species-specific families of highly similar sequence that likely arose from lineage-specific tandem gene duplications[137-141]. The 17 *nitr* genes in gar form 15 families (12 genes in a cluster on Loc14 and five genes on unassembled scaffolds) including two families with two genes each, suggesting few recent gene duplication events or rapid sequence divergence after duplication (Supplementary Fig. 25).

10. Annotation and expression of gar mineralization genes

Spotted gar has enamel-bearing teeth and ganoid scales. Gars retain ancestral characteristics in teeth and scales. Their teeth consist of body dentin, covered with enamel on the tooth shaft and enameloid on the tooth apex[17]; in addition, gar has ganoid scales, comprising a basal bony plate and superficial ganoin layers[18]. Such or similar organizations in teeth and scales are found only in gars and bichirs (*Polypterus*) among extant clades, and are in contrast to teleost scales and teeth, which have only enameloid on the tooth[142], and only a bone-like or dentin-like tissue in the scales with no ganoin on the scale surface[143].

The surface enamel, enameloid, and ganoin are similar to bone and dentin in that they form in organic extracellular matrices[144,145]. However, by actively removing organic elements, these surface tissues mature into hypermineralized wear-resistant tissues[142,146]. Among these tissues,

enameloid is different from enamel and ganoin: enameloid forms in a collagen-rich extracellular matrix secreted by both epithelial-derived cells and mesenchyme-derived cells, whereas both enamel and ganoin grow in non-collagenous matrices, secreted solely by epithelial-derived cells[146-148]. Given the similar developmental processes, ganoin has been regarded as enamel despite a paucity of genetic information about ganoin[18].

Methods. We searched for *scpp* genes in the gar genome sequence using tBlastN[27] with amino acid sequences of Scpp proteins from various lobe-finned vertebrates as queries. This analysis revealed two *scpp* gene clusters, one on LG2 and the other on LG4. For these two regions, *scpp* genes were further sought from transcribed sequences (RNASEQT) based on the Broad Institute transcriptomes (Supplementary Note 3.1) as available from the Ensembl genome browser. Exons of *scpp* genes were also searched directly on the genome sequence using the exon-prediction program GENSCAN[149] [http://genes.mit.edu/GENSCAN.html], and manually by exploring splice donor and acceptor sites. From these transcripts and potential exons, previously unknown or highly divergent *scpp* genes and exons we analyzed if the gene or the exon has the characteristic exon-intron structure specific to the members of this gene family[150]. Gar *scpp* gene annotations and orthologies are summarized in Supplementary Tab. 9.

For exons of identified *scpp* genes, PCR primers were designed and used to investigate expression of each gene in growing teeth, jaw, and skin that includes developing ganoin. For this analysis, cDNA libraries (long distance PCR products of SMART cDNA libraries; Clontech) were made from three different unsexed young gars (*Lepisosteus oculatus*) with different total lengths: teeth from a 30 cm individual, jaw from a 24 cm specimen, and scales from a 17 cm long fish. PCR products were also used to confirm the nucleotide sequences of the transcripts. Some *scpp* genes have a large exon that codes for long simple repeats, and most of these repeats fall into a sequence gap in the genome. These regions were amplified by PCR using genomic DNA as the template, and the products were used to determine the actual nucleotide sequence. See Supplementary Tab. 9 for accession numbers of gar *scpp* gene sequences determined in this study.

The PhyloFish gar transcriptome (Supplementary Note 3.2) was further analyzed for *scpp* expression by adding manual annotations of *scpp* genes (confirmed with Spidey[151]) to the Ensembl gene annotations, removing any Ensembl entry that overlapped a manually curated *scpp* gene on the same strand using bedtools v2.23.0[152]. Phylofish RNA-seq reads failing the Illumina Chastity filter were removed. Adapters were clipped from remaining reads using Cutadapt version 1.7.1[153] with default parameters and reads were quality trimmed using Trimmomatic version 0.33[154] with a sliding window of size 5, a minimum average quality score of 10, and a minimum length of 25nt. Only paired end reads were used for all future analyses. Reads were then aligned to the gar genome using GSNAP version 2014-12-28[155] with parameters '-B 5 -M 0 -m .05 -A sam -Q --split-output' and using the splice file created from the updated Ensembl annotation GTF file. Reads in the "concordant_uniq" output file were converted to BAM format and sorted using samtools version 1.1[156]. htseq-count version 0.6.0[157] was used to count reads aligning to annotated exons with stranded set to 'no' and mode set to 'intersection-strict'. Non-protein coding genes were removed, and read counts were converted to Fragments Per Million (FPM).

***Scpp* locus duplications in vertebrates.** *Sparcl1* (*Sparc-like 1*) is thought to be the last common ancestor of *Scpp* genes[158,159], and in the spotted gar genome, this gene is located on LG4 between two different classes of *scpp* genes, each coding for proteins with biased amino acid compositions, proteins either rich in Pro/Glu or acidic amino acids (Fig. 3a). Considering the arrangement of these genes in various bony vertebrate genomes, an ancient configuration of the two classes of *scpp* genes and *sparcl1* appears to be maintained in LG4, whereas *scpp* genes on LG2 were separated from *sparcl1* by chromosomal rearrangements.

In the spotted gar genome, two different *sparc*/*sparcl1*-like genes (*sparcl1-like* and *sparcr1* in Fig. 3a) were additionally identified on LG2 and LG4, adjacent to the *scpp* gene clusters (Supplementary Tab. 9). Furthermore, *sparc* (an ancestral paralog of *sparcl1*) is located on LG6. The arrangement of vertebrate *Sparc* and *Sparcl1*, and *Scpp* genes suggests a complicated duplication history.

The spotted gar *scpp* gene cluster on LG2 shows double conserved synteny to two *scpp* gene regions on zebrafish chromosomes Dre10 and Dre5 and their orthologous regions in stickleback (GacXIV, GacXIII), indicating that these two regions were generated during the teleost genome duplication (TGD) (Supplementary Fig. 26).

Expression of *scpp* genes in gar. Expression data from RT-PCR and transcriptome analyses are summarized in Supplementary Tab. 9. Fourteen genes were found to be expressed in both teeth and scales: *lpq8*, *lpq7*, *enam*, *scpp5*, *scpp7*, *lpq6*, *ambn*, *odam*, *scpp9*, *lpq1*, *scpp1*, *dsppl1*, *lpq16a,* and *lpq14*. Six genes were found to be expressed in bone (vertebral column): *scpp1* and *lpq14* (also expressed in teeth and scales), *ibsp*, *mepe1*, *mepe2*, and *spp1*. No expression of *enam*, *ambn*, *odam*, or *scpp9* was found in bone.

11. Spotted gar miRNA genes

In one set of analyses, we utilized only the gar genome assembly to study miRNA genes *in silico* (Supplementary Note 11.1), and in another series of experiments, we coupled bioinformatic curation informed by small RNA-seq experiments to identify and annotate gar miRNA genes (Supplementary Note 11.2).

11.1 *In silico* analyses of gar miRNAs

(J.H., M.F., S.K., P.F.S.)

Methods. We started with miRNAs from the miRBase database[49,160-162] (release 19.0) as an initial set for the annotation of homologous miRNAs in the spotted gar genome. The set included 2,001 miRNA precursor sequences from species that enclose the phylogenetic position of spotted gar, including teleosts – *Cyprinus carpio*, *Danio rerio*, *Fugu rubripes*, *Hippoglossus hippoglossus*, *Ictalurus punctatus*, *Oryzias latipes*, *Paralichthys olivaceus*, and *Tetraodon nigroviridis* – and tetrapods – *Xenopus tropicalis*, *Anolis carolinensis*, *Gallus gallus*, *Mus musculus* and *Homo sapiens*.

According to miRBase, 1,915 (~95%) miRNA sequences grouped into 128 gene families. The remaining 86 have not yet been assigned to a miRNA family. All 2,001 miRBase pre-miRNAs were used as query to search genomes using NCBI Blast[163] with an e-value of 1e-3; genomes searched were *Branchiostoma floridae* (bfl), *Petromyzon marinus* (pma), *Callorhinchus milii* (cmi), *Astyanax mexicanus* (amx), *Danio rerio* (dre), *Gadus morhua* (Gmo), *Tetraodon nigroviridis* (tni), *Takifugu rubripes* (fru), *Oryzias latipes* (ola), *Xiphophorus maculatus* (xma), *Oreochromis niloticus* (oni), *Gasterosteus aculeatus* (gac), *Lepisosteus oculatus* (loc), *Latimeria chalumnae* (lch), *Xenopus tropicalis* (xtr), *Anolis carolinensis* (aca), *Gallus gallus* (gga), *Mus musculus* (mmu), *Homo sapiens* (hsa). Candidate sequences were taken from the genomes and analyzed for the presence of mature miRNAs and the ability to fold into a stable stem-loop structure using RNAfold[164]. Evolutionary conservation of the precursor's sequence and its secondary structure was verified in a (multiple) sequence alignment with the respective query sequence(s). We constructed a presence/absence map summarizing how many paralogs of each miRNA family were found in each analyzed species (Supplementary Tab. 10).

Results. The results of the bioinformatic, *in silico*, searches are summarized in Supplementary Table 10. Each entry contains the number of paralogs detected for a miRNA family in each species. We list only those miRNA families with at least one copy in Actinopterygii and/or the gnathostome ancestor. Most of these miRNA families (68) appeared to have an origin earlier than the divergence of ray-finned and lobe-finned vertebrates. Gar clearly shows the signatures of a gnathostome that did not undergo the TGD. To analyze the retention of miRNA duplicates after the TGD, we excluded miRNA families that either have been lost completely in the teleosts or that have their origin in teleost species (e.g., *mir7147*). This leaves 102 miRNA families subject to the TGD. Our analysis indicated that 44 miRNA families have members that were lost immediately after the TGD, while 58 families have duplicates that were retained in extant teleost fishes. Fifty miRNA families have one representative outside of teleost fishes, and 21 of those were not retained in duplicate after the TGD. Supplementary Fig. 27 compares the number of copies this bioinformatic analysis found in each miRNA family in teleosts with the number of copies of miRNA genes present in each family in species that did not experience the TGD.

11.2 Small RNA sequencing-based miRNA annotation and analyses

(T.D., M.J.B., P.B., J.S., J.H.P.)

RNA extraction, library preparation and sequencing. Small RNAs were extracted using the Norgen Biotek microRNA purification kit from four different organs (brain, heart, testis and ovary) from reproductive adult spotted gar caught in the wild in Louisiana. From two males, we sampled brain, heart, and testes and made six libraries, three for each individual. From two females, we sampled ovaries and made two libraries, one for each individual. Thus eight tissue-specific sequencing libraries were prepared and barcoded using the BiooScientific NEXTflex small RNA Sequencing Kit, which uses a 3' adenylated adapter that ligates onto miRNAs and other small RNAs with a 3' hydroxyl group, following the manufacturer's instructions. Libraries were subsequently sequenced with Illumina HiSeq2500 and raw single-end 50nt long reads were deposited in the NCBI Short Read Archive under accession number SRP063942. All

animal work was performed according to the University of Oregon IACUC approved protocol (#09–1BRRA). Gar small RNA-seq data were compared to similar small RNA seq data set published for zebrafish "AB" strain (SRP039502[165])

Small RNA processing and microRNA prediction and annotation. Data were processed as described for zebrafish[165]. Briefly, bioinformatic processing involved removing 3' adapter sequences, filtering out reads with any base Q<30 using the FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/), removing reads outside of the targeted size range (<15 and >28), counting the number of times each read occurred, and removing reads with low counts (<30 summed across all eight libraries). Reads were then processed with Prost![166] (see code availability below) using the reference genome assemblies of gar (LepOcu1) implementing blastn[27] by allowing up to five mismatches, one gap, and alignments at up to 20 locations in the genome. Annotation was predicted against mature and hairpin miRNA sequences present in miRBase (Release 20)[49], recently described miRNAs[165], as well as zebrafish noncoding RNA sequences annotated by Ensembl (release 74)[167]. The final annotation was curated based on read presence in our experimental dataset, curation of Ensembl predictions, and orthology searches among species. Secondary structures of microRNA hairpins were computed on the RNAfold web server[168] using default parameters except for changing the calculation of minimum free energy from 37°C to 28°C for zebrafish and 24°C for gar, the temperature at which these animals are reared, respectively[169]. A tentative name was assigned following miRNA gene nomenclature guidelines[170] based on orthology relationships and identity with sequences from zebrafish or other species already deposited in miRBase[49].

Code availability. Code for the miRNA annotation pipline is available within the package Prost!, publicly available at http://dx.doi.org/10.5281/zenodo.35461.

Sequencing results. The sequencing of small RNAs with 3' hydroxyls from several tissues of actively reproductive adult gar yielded a total of 57.5 million reads that passed all filtering criteria, among which 39.7 million (69%) could be directly annotated by sequence identity as putative mature chordate miRNAs from corresponding entries in miRBase[49] and to recently identified zebrafish miRNAs[165]. Among remaining reads sequences mapping perfectly to the gar reference genome at fewer than 20 locations were further studied for putative new miRNA genes and/or for mature strand annotation. Zebrafish data were analyzed as previously described[165].

Orthology relationships. Orthologies between gar and zebrafish miRNA genes were deciphered through conserved synteny analyses. For each gar miRNA gene, we established a window of six genes surrounding the miRNA (three protein-coding genes on each side of the miRNA gene) and searched orthologous regions in the zebrafish genome using the Ensembl orthology database[167], the Synteny Database[86] and Genomicus[171,172]. Regions were considered orthologous if at least one gene out of a six-gene region could seed the discovery of a wider conserved cluster of genes with the Synteny Database or Genomicus webservers. Thus, orthology or co-orthology relationships of miRNA genes between gar and zebrafish were defined based on sequence similarity plus the orthology or co-orthology conservation of the

genomic region. For every gar gene, the TGD duplicated clusters in zebrafish were searched for, and even if the miRNA gene was lost in one ohnologous region, the track of the putative conserved double synteny was reported. For calculating the retention rate of TGD orthologs between gar and zebrafish, only miRNA genes present in both species and for which orthology/co-orthology relationships could be established were taken into account. Additional *cis*-duplicates were not considered in the study.

Gar miRNA annotation. Analysis uncovered 233 gar miRNA genes including 229 genes in 107 miRNA families and four genes not associated with a family (Supplementary Fig. 28). Ensembl annotation predicted 258 genes, but 39 of them were filtered out because they were absent from our miRNA-seq data and lacked a confident homolog (16 hit to doubtfully predicted miRNAs from other species and 23 hit only non-vertebrate metazoans or plants). Sequencing data confirmed the expression of 211 genes, 199 of them predicted in Ensembl and 12 found by orthology conservation with other species but lacking an Ensembl prediction. Manual curation and expertise added 22 genes to the sequencing data list (20 predicted by Ensembl without sequencing data and two predicted only by orthology with other species). For each annotated gar miRNA gene, a full description of its attributes (Ensembl accession number, location, family, mature sequences if available, clustering, secondary structure, and minimum free energy) is given in Supplementary Table 11. 65% of gar miRNA genes are intergenic (152/233), while 35% are intragenic. About half of all gar miRNA genes (115/233) are organized in 49 clusters of two or more genes, similar to zebrafish, mouse, and human[160,173-175].

Zebrafish annotation. We recently published an extended annotation of zebrafish miRNA genes based on sequencing data[165]. Availability of the gar genome allowed us here to annotate an additional six new zebrafish miRNA genes by orthology conservation with gar and other species. For each of the six newly identified zebrafish miRNA genes, a full description of its attributes (location, family, mature sequences if available, clustering, secondary structure, and minimum free energy) is given in Supplementary Table 11. Synteny and sequencing data confirmed the presence and expression of *dre-mir-2187b* even though the sequence in the zebrafish genome is incomplete for the full hairpin.

Approach comparison. The two different approaches (*in silico* prediction, Supplementary Note 11.1, and experimental small RNA-seq, Supplementary Note 11.2) yielded different miRNA gene numbers and miRNA family predictions. The experimental approach predicted 28 more miRNA genes: 14 more genes in families studied in the 11.1 analysis, ten genes from families that were not studied in the 11.1 analysis, and four miRNAs that are not yet in a recognized family. Experimental data permitted the annotation of the gar miRNAOme by providing genomic locations of miRNA precursors as well as the location and sequence of mature miRNA products that were present in the experimental datasets. In addition, experimental data coupled with orthology studies led to some differences from the *in silico* analysis in gene and family gain and loss during actinopterygian evolution. The *in silico* study (11.1) suggested that *mir451* and *mir551* were lost in teleosts, but these sequences appeared in zebrafish small RNA-seq data[165] and/or are referenced in miRBase. Likewise, *mir7147* was present in the gar small RNA-seq

libraries, and so it is not a teleost-specific miRNA. The miRNA section of the main manuscript reports results primarily based on the small RNA-seq analysis (11.2).

12. Conserved non-coding elements

12.1 Conserved noncoding elements at selected developmental gene loci

Methods. To explore the utility of spotted gar for predicting CNEs in human developmental genes and evaluating the effect of the TGD on the evolution of CNEs in teleosts, we first selected developmental gene loci that are rich in conserved *cis*-regulatory elements including HoxA-D clusters, Pax6 and IrxB[176-178], and predicted CNEs in elephant shark, spotted gar, coelacanth, human and representative teleost fishes (zebrafish and stickleback). Repeat-masked genomic sequences encompassing the selected developmental gene loci were extracted from the UCSC genome browser [http://genome-euro.ucsc.edu/]. Elephant shark sequences were obtained from GenBank. Repeat-masked sequences were aligned using the 'glocal' alignment program SLAGAN[179]. CNEs were predicted using a definition of >65% identity and ≥50 bp windows and viewed using VISTA[180]. CNE tables were extracted from the VISTA browser and analyzed using MS Excel.

Results. We predicted 'gnathostome CNEs' (conserved in elephant shark, spotted gar and human) and 'bony vertebrate CNEs' (conserved in spotted gar and human but absent in elephant shark), by using elephant shark and gar as reference sequences, respectively (see Supplementary Tab. 12-13 and Supplementary Fig. 29 for result summaries and Supplementary Tab. 14-19 and Supplementary Fig. 30-35 for locus-specific results).

Spotted gar loci were found to contain a higher number of gnathostome CNEs than those of the two teleost fishes. The inclusion of the spotted gar genome enabled delineation of a substantial number of CNEs recruited in the bony vertebrate ancestor that could not have been predicted by comparing directly only human and teleost fish genomes (Supplementary Tab. 12-13). Such CNEs are likely to represent *cis*-regulatory elements involved in bony vertebrate-specific gene-regulatory networks[181]. In addition, spotted gar-based alignments led to the identification of CNEs that evolved in some loci in the common ancestor of ray-finned fishes (Supplementary Tab. 12). For example, IrxB, HoxC and HoxD loci contain 76 CNEs (average length 116 bp), 12 CNEs (average length 161 bp) and 14 CNEs (average length 133 bp), respectively, that evolved in the ray-finned fish ancestor while very few (≤ 6) such CNEs are found in the HoxA and Pax6.1 loci. CNEs that specifically evolved in the ray-finned fish ancestor are likely to be either compensating for the loss of ancient gnathostome CNEs or to be involved in functions unique to ray-finned fishes.

To determine the role of the TGD in CNE loss, we compared the number of elephant shark-human CNEs that are lost from the unduplicated genome of spotted gar and from the duplicated genomes of zebrafish and stickleback. While spotted gar was found to lack 20 to 41% (average 26%, Supplementary Tab. 12-13) of such CNEs, almost twice that number are lost in the a-paralogs of zebrafish (26 to 76%, average 56%) and stickleback (39 to 82%, average 60%). Moreover, the teleost fish b-paralogs have lost far more elephant shark-human CNEs (77 to 99%) than the a-paralogs.

12.2 Whole genome alignment-based analysis of conserved non-coding elements

Genome preparation. Three different sets of multiple whole genome alignment (WGA) were generated: a gar-centric, a zebrafish-centric, and a human-centric WGA. For each reference species, we generated a 13-way WGA with 12 additional gnathostome vertebrate species. Genome assemblies were downloaded from Ensembl (human, GRCh37), UCSC Genome Browser (mouse, mm10; chicken, galGal4; Anole lizard, anoCar2; Xenopus tropicalis, xenTro3; Tetraodon, tetNig2; stickleback, gasAcu1; zebrafish, danRer7), and NCBI (gar, GCA_000242695.1; coelacanth, GCA_000225785.1; elephant shark, GCA_000165045.2). For platyfish, we used chromosome-level assembly version xma_washu_4.4.2-jhp_0.1[182,183] (based on assembly GCA_000241075.1). Genomes were masked for repeats using RepeatMasker with species-specific repeat libraries (custom libraries: gar, Lo-TEs-v3.fa, Supplementary Note 5; coelacanth, ref.[57]; platyfish, ref.[182]) and Tandem Repeats Finder (TRF)[44].

Whole genome alignments. First, genomes were aligned against the reference species using lastZ[184] with the HoxD55 scoring matrix and the following parameters: H = 2000, Y = 9400 (3400 for 'distant alignments'), L= 3000 (6000), K= 3000 (2200). 'Distant alignments' were defined as non-ray-finned species for the gar-centric as well as the zebrafish-centric alignments, and non-lobe-finned species for the human-centric alignment. Next, MultiZ[185] with roast.v3 was run to generate the three 13way multi-genome alignments using the following tree topology based on known phylogenetic relationships[68]: (((gar, (zebrafish, ((medaka, platyfish), (stickleback, Tetraodon)))), (coelacanth, (Xenopus, ((human, mouse), (chicken, lizard))))), elephant shark).

Generating conserved elements. We used phyloFit[186] (general reversible substitution model "REV") to obtain neutral models for the three species-centric multiple genome alignments from fourfold degenerate (4d) sites obtained from the Ensembl protein-coding gene annotations of the center species (Supplementary Fig. 36). For zebrafish, we excluded 4d sites from the highly repetitive, potential sex chromosome chr4[52,187,188]; for human, we excluded the sex (X/Y) chromosomes as well. To define the most conserved elements in each our 13-way WGA, we then ran phastCons[186] with the following parameters: average length of conserved sequence: 45bp; target coverage of input alignments: 0.3; rho=0.3 of the neutral model.

Masking and filtering conserved elements to obtain conserved non-coding elements. To obtain conserved non-coding elements (CNEs), the most conserved elements of each of the three center species (gar, zebrafish, human) were filtered for genic elements as well as repeat elements. For all genic elements in all three species, the masks were extended by 50bp in both directions into non-exonic space to lessen the likelihood of including conserved splicing motifs.

For spotted gar, we removed all bases from the most conserved elements that overlap exons and UTRs of the MAKER and Ensembl annotations of protein coding genes and noncoding RNAs (Supplementary Note 4) as well as all elements that overlap annotated miRNAs from Ensembl and identified here (Supplementary Note 11). All bases covered by RepeatMasker, TRF, Repeatrunner (from MAKER annotation), and annotated as repeats in Ensembl were removed as well.

For zebrafish, we removed all bases from the most conserved elements that overlap exons and UTRs of the Ensembl annotation of protein coding genes and noncoding RNAs, the UCSC table browser tracks (http://genome.ucsc.edu/cgi-bin/hgTables?command=start) for RefGenes and tRNA genes, as well as all elements that overlap annotated miRNAs from Ensembl, mirBase[49] and our extended zebrafish miRNA annotation[165]. All bases overlapping repeat elements from Ensembl repeat tracks and UCSC tables (tables: interrupted repeats, repeatmasker, simple repeats) were also removed as well as all elements from the heterochromatic, highly repetitive region on zebrafish chr4[52,187,188].

For human, we removed all bases from the most conserved elements that overlap exons and UTRs of the Ensembl annotation of protein coding genes and noncoding RNAs, the UCSC browser tracks for RefGenes, UCSC knownGene, tRNAs, snoRNAs, and miRNAs. Bases overlapping Ensembl repeats and UCSC repeat tables (repeatmasker, microsatellites, simple repeat) were excluded as well as elements not located on chromosomes.

To generate our final CNE complement for the three center species (Supplementary Tab. 20), we filtered for elements that are at least 50bp in length, that have a transformed LOD score at least 333, and that have for at least two species alignments that each cover at least 33% of the element length.

Phylogenetic origin of CNEs. For the three 13-way WGAs, we then determined the phylogenetic age of each CNE by asking for the phylogenetically most distant species among the alignments against the center species covering at least 33% of the alignment length (Supplementary Tab. 20).

In the gar-centric CNE set, we distinguish gnathostome CNEs (GCNEs: alignment of gar and elephant shark), bony vertebrate CNEs (BCNEs: alignment of gar with at least one lobe-finned vertebrate, but no alignment to elephant shark), and ray-finned fish CNEs (RCNEs: alignment of gar to at least two teleosts, but no alignment to any non-ray-finned vertebrate).

In the zebrafish-centric CNE set, we distinguish GCNEs (alignment of zebrafish and elephant shark), BCNEs (alignment of zebrafish to at least one lobe-finned vertebrate, but no alignment to elephant shark), RCNEs (alignment of zebrafish and gar, but no alignment to any lobe-finned vertebrate or to elephant shark), and teleost CNEs (TelCNEs: alignment of zebrafish to two other teleosts, but no alignment to any non-teleost including gar).

In the human-centric CNEs set, we distinguish GCNEs (alignment of human and elephant shark), BCNEs (alignment of human abd at least one ray-finned fish, but no alignment to elephant shark), lobe-finned vertebrate CNEs (LCNEs: alignment of human and coelacanth, but no alignment to any non-lobe-finned vertebrate), tetrapod CNEs (TetraCNEs: alignment of human and Xenopus, but no alignment to any non-tetrapod), and amniote CNEs (ACNEs: alignment of human and at least two amniotes, i.e. chicken, lizard, mouse, but no alignment to any non-amniote). Note that mammalian-specific CNEs (alignment against mouse only) were not defined here because we required at least two species (not including the reference) in the alignment block to call a CNE.

Genome-wide connectivity of human CNEs to zebrafish through gar. About 90% of genetic variants that are associated with human disease in genome-wide association studies (GWAS) are located in non-coding elements[189-191]. A major gap in our understanding is the

mechanisms whereby these disease variants or factors linked to them actually contribute to disease. To investigate such biologically relevant regions in teleost models like zebrafish, the orthologous region(s) in model species' genome must be identified. Results showed that 34,133 CNEs from the human-centric WGA (aligning to at least one lobe-finned vertebrate) were readily connected to zebrafish based on a direct alignment from human to zebrafish, and 54,599 CNEs were directly connected from human to at least one of the five teleost species in the human-centric WGA (Supplementary Tab. 21).

Analysis identified 19,149 human CNEs that were not directly connected to any teleost, but that were connected to gar. We identified the orthologous genomic location in the gar genome for 18,994 of these human CNEs using UCSC's liftOver tool (-minMatch = 0.33)[192]. For these 'human CNEs in gar', we attempted to identify the orthologous region(s) in the zebrafish genome. To this end, we used liftOver (-minMatch = 0.33) on the 'most conserved elements' from the zebrafish-centric WGA to identify their orthologous genomic location in the gar genome. Next, we intersected within the gar genome the 'human CNEs in gar' with the 'zebrafish conserved elements in gar' using BEDtools[152]. The process established connectivity for between 5,761 to 6,839 human CNEs depending on the amount of overlap required (from 1bp to 33% CNE length) within the gar genome (Supplementary Tab. 21). Thus, using the gar genome we could infer hidden orthology from human to zebrafish through gar for more than 30% of human CNEs that were not previously directly connected from human to zebrafish.

Connectivity of human GWAS-SNP locations to zebrafish through gar. Connectivity from human to zebrafish through gar further enables us to identify the location of the zebrafish genomic region orthologous to human CNEs that contain SNPs found in genome-wide association studies (GWAS-SNPs) for human diseases and phenotypes. We downloaded 2,435,071 human GWAS-SNPs from Ensembl75 through BioMart, 25,555 of which are located in 21,564 human CNEs. We then analyzed the number of GWAS-SNPs and the amount of GWAS-SNP-containing CNEs in the connectivity analysis (Supplementary Tab. 21). Connectivity from human to zebrafish through gar was established for 992 to 1,217 GWAS-SNPs from 848 to 1,021 GWAS-SNP-containing CNEs, respectively (depending on the amount of overlap within the gar genome from 1bp to 33% CNE length).

11.3 Connectivity analysis of human limb enhancers

To study the presence/absence of tetrapod limb enhancers in the spotted gar genome, we downloaded a list of all 160 human enhancers positive for limb expression in LacZ transgenic mouse assays at developmental stage e11.5 from the VISTA Enhancer Browser[193] (http://enhancer.lbl.gov/; download date 08/20/2014). In addition, we added 6 human limb enhancer regions from the *BMP7*, *SHH*, and *GLI3* gene regions previously analyzed for presence in coelacanth by Nikaido et al. (2013)[194] as well as 18 known limb enhancer region from the *HoxA* and *HoxD* clusters[176,195]. We then inspected the genomic region of these 183 limb enhancer regions (Supplementary Tab. 22) for presence/absence in spotted gar, teleosts, and other gnathostomes in the 13-way human centric whole genome alignment (Supplementary Note 12.2). A total of 30 enhancers showed conservation only with mouse, suggesting that they

arose within mammals and were not further considered for candidates present in the bony vertebrate ancestor of spotted gar and human, leaving 153 enhancers for further analysis.

For 30 limb enhancer regions that were present in gar but absent in teleosts, we obtained the location of the gar region from the human-centric WGA, checked its orthology in gar by confirming conserved syntenies, and then analyzed the gar-centric whole genome alignment (Supplementary Note 12.2) for an alignment from gar to teleosts. The location of the inferred orthologous region in teleost genomes was further supported by conserved synteny analysis. The results of the connectivity analysis are shown in Supplementary Tab. 22, which led us to a model for the origin of human limb enhancers and their losses among bony vertebrate lineages shown in Supplementary Figure 37 and summarized in Figure 4b of the main text.

12.4 HoxD limb enhancer CNS65

Genomic alignments. Genomic segments of interest were downloaded from the Ensembl and UCSC genome databases and aligned using the mVista (LAGAN)[180,196] program with the following parameters; calc window: 100 bps, Min Cons Width: 100 bps, Cons Identity: 65%. Mouse assembly version 10 (mm10, Dec. 2011) was used in the alignment.

Cloning of gar and zebrafish CNS65. Conserved peaks were amplified by PCR using the following primers: gar- 5'-AAACGATCGCAGTGTTTCAGT-3', 5'-GTCTGGTGGCCTGTGTAAAAA-3' and zebrafish- 5'- CCACTTAAACTGCGCATCAA-3', 5'-TGGATGAACCAGGTATTGCAG-3'. Genomic fragments were gel purified using the NucleoSpin Gel and PCR Clean-up Kit (Macherey-Nagel), and subcloned into PCR8GW/GW/TOPO vector according to the manufacturer's protocols (Invitrogen). The gar and zebrafish CNS65 fragments were shuttled into either the pXIG-cFos-eGFP vector[197] for zebrafish transgenesis, or to the Gateway-Hsp68-LacZ vector (gift of Marcelo Nobrega, University of Chicago) for mouse transgenesis, both using the Gateway system (Invitrogen). Vectors were confirmed by restriction digest and sequencing.

Zebrafish stable line transgenesis: Zebrafish embryos were collected from natural spawnings, and staged according to standard measures[198]. Transposase RNA was synthesized using the mMessage mMachine SP6 kit (Ambion), using the pCS2-zT2TP vector that produces RNA that is codon-optimized for zebrafish[199]. Solutions for injection were prepared according to Fisher et al.[197], and injected into the cytoplasm of 1- or 2-cell wild-type embryos (~100 embryos per construct). Embryos were maintained in egg water at 28^{o}C until visualization at the appropriate stage using a Leica M205FA microscope. About 30 embryos per construct displaying consistent, bright GFP signal were raised to sexual maturity and outcrossed to *AB WT to identify founders in the F1 generation. All founder fish showed very similar expression patterns for these enhancers. We generated at least three independent stable lines for every construct that was injected.

Mouse transgenesis: Sequence-confirmed gar and zebrafish CNS65-Hsp68-LacZ vector were delivered to Cyagen Biosciences for mouse transgenesis (Cyagen Biosciences, Santa

Clara, CA). Briefly, vectors were linearized with SalI, gel purified, microinjected into fertilized mouse oocytes (minimum 150 eggs per construct) and transferred to pseudo-pregnant females. Embryos were collected at e10.5 or e12.5, stained for beta-galactosidase activity, and genotyped using DNA extracted from yolk sac.

Gar CNS65 at e10.5: four embryos were PCR positive for LacZ; one embryo showed no LacZ staining, the other three showed consistent expression in the limb.

Gar CNS65 at e12.5: five embryos were PCR positive for LacZ; all five showed LacZ staining, all of which showed consistent expression in the proximal limb.

Zebrafish CNS65 at e10.5: eight embryos were PCR positive for LacZ; four showed LacZ staining, three of which showed consistent expression in the limb.

Zebrafish and mouse transgenesis experiments were approved by the University of Chicago IACUC committee (ACUP #72074).

13. Analysis of gene expression after the teleost genome duplication

13.1 Identification of TGD ohnologs and singletons in zebrafish and medaka

TGD ohnologs in zebrafish and medaka. To identify zebrafish and medaka TGD co-orthologs of spotted gar protein coding genes, zebrafish and medaka predicted intragenomic paralogs were downloaded along with their spotted gar orthologs from Ensembl74 using Biomart[85]. These ohnologous pairs were filtered for the Biomart-derived duplication ancestor prediction 'Clupeocephala', the most basal duplication point after divergence of sequenced teleosts in Ensembl from gar. Next, each paralog was required to be present only once in the dataset, thereby removing gene duplications that occurred within the lineages leading to zebrafish and medaka after the TGD. Furthermore, each pair was required to have a unique, single gar ortholog to remove paralog pairs for which no gar ortholog was available or for which gene duplication(s) occurred within the gar lineage. Cases of 'split genes', genes that obviously had assembly or annotation errors, were removed as well. This process yielded a total of 1,901 cases of 1:2 gene relations between gar and zebrafish and 1,597 cases of 1:2 gene relations between gar and medaka.

To further filter for zebrafish and medaka paralogs that show the expected pattern of double conserved synteny generated by the TGD, the 1:2 spotted gar vs. zebrafish/medaka gene trios were required to be located on in paralogous clusters defined by the Synteny Database[86] using zebrafish/medaka as source genomes and spotted gar as outgroup genome (sliding window size: 200 genes; membership ≥10 paralogous pairs). After this conserved synteny filtering, 1,606 pairs of zebrafish paralogs and 1,315 pairs of medaka paralogs were retained; we consider these to represent a highly stringent subset of 'TGD ohnologs' having both phylogenetic and synteny support for origin in the TGD (Fig. 6b), but that certainly underestimates the true number of retained TGD ohnolog pairs.

TGD ohnolog pairs were associated to each other based on their single gar ortholog for zebrafish and medaka. Orthology of zebrafish genes to medaka genes was confirmed by patterns of medaka/zebrafish conserved synteny obtained with the Synteny Database[86]. This process defined a total 774 TGD ohnolog pairs shared between zebrafish and medaka (Fig. 6b).

The TGD ohnolog list of zebrafish (1,606 pairs) was randomized with respect to the assignment of one or the other TGD ohnologs of a pair as “Ohnolog1” or “Ohnolog2”. Assignment to “Ohnolog1” or “Ohnolog2” for the 774 TGD ohnologs shared between zebrafish and medaka followed the randomized zebrafish assignment. In other words, it ignored previously assigned ‘*a*’ copy or the ‘*b*’ copy or ‘1 of 2’ and ‘2 of 2’ gene nomenclature designations to avoid bias in analyses. The remaining 541 TGD ohnologs from medaka not shared with zebrafish were further randomized as “Ohnolog1” or “Ohnolog2” (Supplementary Tab. 23).

Singletons in zebrafish and medaka. To identify likely singleton genes with respect to the teleost genome duplication, i.e. cases in which one of the two TGD ohnologs was lost following the TGD in the zebrafish and/or medaka lineage, respectively, we removed genes from the BioMart-derived list of intragenomic paralogs that had an indication for TGD duplication (duplication ancestor ‘Clupeocephala’, see above) as well as for lineage-specific gene duplication after the TGD (e.g. for zebrafish, duplication ancestors ‘Otophysi’ and ‘Danio’). Furthermore, we removed genes with duplication ancestor ‘Neopterygii’ (i.e. the ancestor of gar and teleosts) because these inferred duplication nodes could be artifacts of tree reconstructions in Ensembl and thus potentially include TGD ohnologs or other types of gene duplication that occurred within teleosts. Genes with Ensembl gene names indicative of gene duplication (e.g. ‘1 of 3’) were removed as well. Each zebrafish or medaka singleton gene was required to have a unique, single gar ortholog (‘ortholog-one-to-one’) to remove genes for which no gar ortholog was available or for which gene duplication(s) occurred within the gar lineage. Genes located on unplaced scaffolds or mitochondrial genomes in zebrafish/medaka were removed as well. This survey left us with a list of 10,416 and 9,265 ‘singleton’ genes in zebrafish and medaka, respectively, with a 1:1 relationship to a single gar gene and thus likely to be cases in which one of the TGD ohnologs was lost and one retained (Fig. 6b).

The lists of zebrafish and medaka singleton genes were associated based on their single gar ortholog, identifying a subset of 7,309 genes that are singletons in both zebrafish and medaka, following the parsimonious assumption that the second TGD ohnolog of these genes was lost before the divergence of the zebrafish and medaka lineages, relatively early during teleost evolution within a few tens of millions of years following the TGD[70].

Finally, singleton lists of one teleost species (zebrafish/medaka) were associated with TGD ohnolog lists of the other species (medaka/zebrafish) based on their shared single gar ortholog. This process led to an intersection of 267 zebrafish singletons that intersected with medaka TGD ohnolog pairs, and 518 medaka singletons that merged with zebrafish TGD ohnologs pairs (Fig. 6b). The singleton gene of one species was assigned as orthologous to either “Ohnolog1” or “Ohnolog2” of the other species based on patterns of conserved synteny obtained from the Synteny Database[86] (Supplementary Tab. 23).

13.2. Comparative RNA-seq expression analysis of gar vs. zebrafish and medaka

RNA-seq in zebrafish and medaka. Zebrafish and medaka tissues were collected from individuals grown at the INRA Fish Physiology and Genomics experimental facility.

RNA was extracted from the following tissues from 11 months old medaka: brain, gills, muscle, liver, and intestine from one female; kidney, pooled from two females; and heart and bones, pooled from two females and one male. From 2 months old medaka: ovary from one female; and testis pooled from three males.

In addition, RNA was extracted from pools of zebrafish and medaka embryos when eyes first become pigmented ('eyed stage'; equivalent to gar stage 27-28, Supplementary Note 3.2), i.e., 2 days post fertilization for both species.

RNA-seq and *de novo* transcriptome assemblies were performed as described for gar (Supplementary Note 3.2). RNA-seq data were deposited into SRA under accessions SRP044781 (zebrafish) and SRP044784 (medaka).

See http://phylofish.sigenae.org/ngspipelines/#!/NGSpipelines/Danio%20rerio (zebrafish) and http://phylofish.sigenae.org/ngspipelines/#!/NGSpipelines/Oryzias%20latipes (medaka) for more details.

Transcript expression patterns by mapping and counting of RNA-seq reads. To study the expression patterns and levels of zebrafish, medaka, and spotted gar transcripts, a reference coding sequence (CDS) library was built for each species. Each library was deduced from Ensembl genomic databases for zebrafish (assembly Zv9), medaka (assembly MEDAKA1) and gar (assembly LepOcu1) as follows: for each gene one CDS was retained in the library; when multiple CDS were referenced for a single gene, the longest CDS was arbitrarily retained as representative of the gene product. We then mapped the double stranded RNA-seq reads on the corresponding CDS library using BWA-Bowtie[33,200] with stringent mapping parameters (maximum number of allowed mismatches –aln 2). Mapped reads were counted using SAMtools'[156] idxstat command, with a minimum alignment quality value (–q 30) to discard ambiguous mapping reads. For each species, the numbers of mapped reads were then normalized for each gene across the 11 tissues using DESeq[201].

Evolution of gene expression after the TGD in zebrafish and medaka compared to gar. To compare the expression pattern of genes that were retained as singletons after the TGD to the expression pattern of TGD ohnologs, we created an average expression pattern for each pair of ohnologs calculating the average expression level between the two ohnologs individually for the 11 tissues. This average expression pattern is designated as 'ohnolog pair' (or 'ohno-pair'). Using Pearson's correlation in R[202], we determined the expression pattern correlation between each zebrafish or medaka gene and its gar ortholog. Because values did not meet the assumptions of parametric test (values were not normally distributed and variances were not similar between groups compared), we then performed a multiple two-sided Wilcoxon Mann-Whitney test to compare the mean correlation of singletons, ohnolog-1, ohnolog-2 and ohnolog-pair within and across species. Variances were similar between groups compared.

To study the relative expression levels, we calculated the average expression level of each gene over the 11 tissues. We then calculate the ratio of those average expression levels between each zebrafish or medaka gene and its gar ortholog. Mean expression values meet the assumptions of parametric test (values were normally distributed and variances were similar between groups compared); we therefore performed a multiple two-sided Student t-test to

compare the mean expression level ratio of singletons, ohnolog-1, ohnolog-2, and ohnolog-pair within and across species.

All tests were performed with R[202], and a Bonferroni correction was applied on all multiple tests. Variances were similar between groups compared.

Detection of neo- and sub-functionalization after TGD in zebrafish and medaka. The calculated Pearson's correlation between expression patterns of zebrafish or medaka TGD ohnologs and their gar orthologs were also used to detect automatically neo- and sub-functionalization processes. An arbitrary r value threshold of 0.75 was used to identify correlated expression profiles.

Conditions for detecting conserved neo-functionalization were:

Correlation between zebrafish ohno-1 and medaka ohno-1 >= 0.75,
Correlation between zebrafish ohno-2 and medaka ohno-2 >= 0.75,
Correlation between zebrafish ohno-1 and zebrafish ohno-2 < 0.75,
Correlation between medaka ohno-1 and medaka ohno-2 < 0.75,
Correlation between zebrafish ohno-1 and gar ortholog < 0.75,
Correlation between medaka ohno-1 and gar ortholog < 0.75,
Correlation between zebrafish ohno-2 and gar ortholog >= 0.75,
Correlation between medaka ohno-2 and gar ortholog >= 0.75.

Conditions for detecting conserved sub-functionalization were:

Correlation between zebrafish ohno-1 and medaka ohno-1 >= 0.75,
Correlation between zebrafish ohno-2 and medaka ohno-2 >= 0.75,
Correlation between zebrafish ohno-1 and zebrafish ohno-2 < 0.75,
Correlation between medaka ohno-1 and medaka ohno-2 < 0.75,
Correlation between zebrafish ohno-pair and gar ortholog >= 0.75,
Correlation between medaka ohno-pair and gar ortholog >= 0.75,
Correlation between zebrafish ohno-2 and gar ortholog < Correlation between zebrafish ohno-pair and gar ortholog,
Correlation between medaka ohno-2 and gar ortholog < Correlation between medaka ohno -pair and gar ortholog,
Correlation between zebrafish ohno-2 and gar ortholog < Correlation between zebrafish ohno -pair and gar ortholog,
Correlation between medaka ohno-2 and gar ortholog < Correlation between medaka ohno -pair and gar ortholog.

B. SUPPLEMENTARY TABLES

Supplementary Table 1. Diversity and content of TE superfamilies in the spotted gar genome.

	Statistics concerning repetition in the spotted gar assembly								Statistics after filtering sequences smaller than 80 nt and sharing less than 80% of identity with the reference sequence			
Class/Family	**Total coverage (bp)**	**Coverage (%)**	**Total coverage (TE class)**	**Number of copies**	**% copies length >30%**	**% copies length >50%**	**% copies length >80%**		**Total coverage (bp)**	**Coverage (%)**	**Total Coverage (TE class)**	**Number of copies**
DNA	219084	0.023		1140	62.9	32.5	12.6		19003	0.002		121
DNA/EnSpm	13259	0.001		94	98.9	92.6	84		11604	0.001		79
DNA/Harbinger	324845	0.034		2907	3.7	1.2	0.3		226256	0.024		832
DNA/Helitron	1387	0		10	100	100	80		827	0		6
DNA/Kolobok	116	0		1	100	100	100		116	0		1
DNA/MuDr	406	0		1	100	100	100		406	0		1
DNA/PIF-Harbinger	267478	0.028		816	13.1	7.5	3.4		234192	0.025		597
DNA/PiggyBac	138983	0.015		567	46.2	27.2	10.4		28712	0.003		127
DNA/Polinton	969065	0.101		1632	1.4	0.3	0.1		868563	0.091		1113
DNA/Sola	1833	0		4	75	75	75		838	0		2
DNA/TcMar	278427	0.029		1610	68.6	48.5	3.3		114705	0.012		615
DNA/TcMar-Pogo	243385	0.025		1312	15.4	3.1	0.2		120351	0.013		465
DNA/TcMar-Tc1	18223551	1.909		65097	38.8	21.2	7.1		5474040	0.573		19490
DNA/TcMar-Tigger	1656906	0.174		6462	43.2	27.7	14.3		1228300	0.129		3845
DNA/TcMar-MER6	1532839	0.161		8845	92.6	71.8	37.6		139075	0.015		876
DNA/Mariner	2731980	0.286		7728	33.7	15.1	2.9		509912	0.053		1626
DNA/hAT	2704250	0.283		18884	12	9.3	4.7		1725623	0.181		10912
DNA/hAT-Ac	176016	0.018		590	39.2	21	15.3		173482	0.018		550
DNA/hAT-Buster	653918	0.068		2088	35.8	2.4	0.8		592393	0.062		1153
DNA/hAT-Charlie	3066787	0.321	**DNA**	16683	63.7	45.7	8.8		1291375	0.135	**DNA**	4771
DNA/hAT-Tip100	648362	0.068	**3.544**	2714	30.6	20.6	13.1		421552	0.044	**1.381**	1354
IntegratedVirus/Caulimovirus	23078	0.002		319	100	97.5	54.9		13673	0.001		157
LINE	3301174	0.346		15585	0.1	0	0		3239182	0.339		14831
LINE/CR1	22706510	2.378		90057	33.2	13.2	2.8		5929702	0.621		21756
LINE/L1	2269353	0.238		6838	64.4	25.6	6.9		790936	0.084		1692
LINE/L2	5170501	0.542		24198	26.8	12.5	3.5		2046739	0.214		8382
LINE/Penelope	2642775	0.277		9393	42.2	13.4	5.5		2105160	0.22		5848
LINE/R2	3300758	0.346		15582	0.1	0	0		3238766	0.339		14828
LINE/R4	4185	0		12	75	58.3	41.7		612	0		1
LINE/RTE	1159548	0.121		3910	17.9	11.2	5.5		637913	0.067		1681
LINE/RTE-BovB	915697	0.096		2846	29.7	12.9	3.1		207199	0.022		580
LINE/RTE-X	1136602	0.119		6384	1	0.2	0		948131	0.099		4038
LINE/Rex-Babar	8336055	0.873		33565	29.1	17.5	8		6692485	0.701		20813
LINE/Rex1	3752335	0.393	**LINE**	7608	44.3	25.3	9.4		1656447	0.173	**LINE**	3286
LINE/Vingi	262949	0.028	**5.757**	1704	1.8	0.9	0.5		170856	0.018	**2.897**	999
LTR	1195554	0.125		2280	2.2	0.5	0		805033	0.084		1392
LTR/BEL	1844404	0.193		17715	0.1	0	0		1188293	0.124		9326
LTR/Copia	258542	0.027		420	28.1	20.2	13.1		155860	0.016		182
LTR/ERV1	2603185	0.273		7728	17.6	11.6	5.8		1945586	0.204		5305
LTR/Gypsy	5295012	0.555		13025	19.1	12	5.5		2996814	0.314		6545
LTR/Gypsy-Gmr1	108781	0.011	**LTR**	218	16.5	9.2	1.8		16820	0.002	**LTR**	49
LTR/Ngaro	13075343	1.369	**2.553**	60441	64	45.2	12.2		3793161	0.397	**1.141**	14291
Low_complexity	4523626	0.474		113467	100	100	99.8		243176	0.025		1928
SINE	233879	0.024		2058	90.5	63.8	28.9		42040	0.004		337
SINE/5S	11036055	1.156		34103	75.3	57.5	36.2		10468686	1.096		30209
SINE/AFC	1479	0		21	95.2	95.2	42.9		96	0		1
SINE/Deu	6500648	0.681		31999	50	24.9	5.8		965565	0.101		5368
SINE/HPA	35107	0.004		435	92.4	57.2	35.4		18945	0.002		160
SINE/MIR	2879141	0.302		19453	67.7	44.6	7.6		860120	0.09		4651
SINE/Unclassified	288	0		3	100	100	66.7		172	0		1
SINE/V	5204795	0.545		37440	53.3	24.8	2.2		946090	0.099		6214
SINE/tRNA	21785	0.002	**SINE**	239	89.5	57.3	28		3179	0	**SINE**	32
SINE?	116960	0.012	**2.726**	739	76.7	63.3	50.3		102968	0.011	**1.428**	544
Satellite	503721	0.053		5346	1.3	0.7	0.2		248706	0.026		2063
Simple_repeat	2444774	0.256	**Unknown**	41106	100	100	100		325369	0.034	**Unknown**	1996
Unknown	52882259	5.538	**5.538**	240704	52.7	32.2	14		27575114	2.888	**2.888**	103117
rRNA	88462	0.009		724	87.6	54.6	25.4		56691	0.006		398
Total	**196387439**	**20.568**	**20.118**	**971268**						**9.802**	**9.735**	**341537**

Supplementary Table 2. Transcriptional activity of transposable elements in gar. Number of TE transcripts in the Broad Institute transcriptomes from different tissues and developmental stages.

		Brain	Eye	Heart	Kidney	Liver	Muscle	Skin	Testis	Embryo	Larvae
DNA transposon	DNA	20	23	8	30	17	2	17	21	13	16
	EnSpm	0	0	1	0	0	0	0	0	0	0
	Harbinger	6	17	9	21	14	6	5	7	8	4
	Helitron	0	2	0	0	0	0	0	0	0	0
	Kolobok	0	0	0	0	0	0	0	0	0	0
	MuDr	0	0	0	0	0	0	0	0	0	0
	PIF-Harbinger	23	21	9	15	11	1	11	36	13	29
	PiggyBac	8	19	18	15	10	6	11	19	13	11
	Polinton	149	165	62	138	181	26	127	95	55	146
	Sola	0	0	0	0	0	0	0	0	0	0
	TcMar	12	11	5	9	6	3	2	8	6	3
	TcMar-Pogo	12	10	3	5	4	1	3	1	0	3
	TcMar-Tc1	2065	1746	1093	1868	1366	491	1063	1463	1099	1283
	TcMar-Tigger	221	199	121	204	156	46	110	147	116	129
	TcMar-MER6	0	0	0	0	0	0	0	0	0	0
	Mariner	457	371	275	421	304	101	245	309	274	277
	hAT	123	124	75	150	112	31	76	121	87	104
	hAT-Ac	9	15	7	7	10	6	2	3	8	7
	Buster	37	43	31	44	43	26	39	30	38	38
	hAT-Charlie	191	197	110	194	203	43	130	150	123	146
	hAT-Tip100	55	31	28	38	23	12	14	29	18	15
	Caulimovirus	2	4	2	2	1	1	3	1	1	1
LINE	CR1	1493	1322	786	1410	949	282	751	1080	707	854
	L1	59	53	39	50	36	11	43	56	37	38
	L2	352	345	281	338	263	105	214	340	221	246
	Penelope	264	180	172	203	212	66	143	179	111	114
	R2	33	41	17	31	35	10	17	40	17	22
	R4	0	0	0	0	0	0	0	0	0	0
	RTE	138	117	116	128	96	38	89	153	89	66
	RTE-BovB	43	34	26	46	32	18	23	46	23	33
	RTE-X	26	21	23	39	36	8	26	35	19	11
	Rex-Babar	745	722	483	729	625	178	468	672	376	453
	Rex1	288	308	220	278	236	85	182	309	209	251
	Vingi	7	7	6	13	8	3	6	3	8	7
	LTR	152	102	88	105	86	43	125	135	70	71
LTR	BEL	38	37	28	51	36	11	16	36	19	29
	Copia	21	25	13	14	9	0	7	18	6	28
	ERV1	216	159	157	233	175	98	113	238	100	120
	Gypsy	537	397	216	484	356	80	221	409	190	232
	Gypsy-Gmr1	15	6	0	3	0	0	0	1	0	3
	Ngaro	382	315	262	431	311	103	237	354	265	299
	SINE	10	10	6	2	4	5	3	4	2	6
SINE	5S	178	134	112	190	137	48	130	138	108	119
	AFC	0	0	0	0	0	0	0	0	0	0
	Deu	400	326	220	343	259	96	211	288	222	235
	HPA	1	0	0	3	2	1	1	5	1	2
	MIR	69	60	33	59	35	7	37	53	21	43
	V	155	143	99	123	95	20	97	100	69	91
	SINE?	2	5	4	3	2	1	1	3	0	4
	Unknown	4650	3905	2601	4216	3091	1045	2508	3590	2379	2864

Supplementary Table 3. Percentage of transposable element sequences in different gar transcriptomes.

	Brain	Eye	Heart	Kidney	Liver	Muscle	Skin	Testis
total transcriptome sequence	263982	251429	166566	252078	191325	76383	152004	249425
repeat sequence	9243	8038	5331	8622	6483	2067	5105	7402
% of TEs	**3.50**	**3.20**	**3.20**	**3.42**	**3.39**	**2.71**	**3.36**	**2.97**

	Embryo	Larvae
total transcriptome sequence	155026	177549
repeat sequence	4903	5667
% of TEs	**3.16**	**3.19**

Supplementary Table 4. Molecular rate analyses. a) Tajima's relative rate tests. b) Two-cluster tests on the RaxML and Phylobayes trees. See Supplementary Note 7 for further information.
[separate .xls file]

Supplementary Table 5. Analysis of conserved non-coding elements (CNEs) in ParaHox clusters. Based based on VISTA plots using gar as the base (Suppl. Fig. 14-15). CNE definition: >65% identity, ≥50bp window size.

	Number of CNEs	Average length (bp)	Total length (bp)
CNEs in the ParaHoxA locus (*gsx1, pdx1, cdx2*; Suppl. Fig. 14a)			
Spotted gar - zebrafish a	7	318	2229
Spotted gar - zebrafish b	0	-	-
Spotted gar - stickleback a	4	426	1705
Spotted gar - stickleback b	0	-	-
Spotted gar - coelacanth	14	302	4234
Spotted gar- human	7	311	2180
CNEs in the ParaHoxB locus (*gsx2, pdx2*; Suppl. Fig. 14b)			
Spotted gar - zebrafish a	7	191	1335
Spotted gar - zebrafish b	0	-	-
Spotted gar - stickleback a	4	141	562
Spotted gar - stickleback b	0	-	-
Spotted gar - coelacanth	12	320	9845
Spotted gar - human	11	304	3342
CNEs in the ParaHoxC locus (*cdx1*; Suppl. Fig. 14c)			
Spotted gar - zebrafish a	0	-	-
Spotted gar - zebrafish b	0	-	-
Spotted gar - stickleback a	0	-	-
Spotted gar - stickleback b	0	-	-
Spotted gar - coelacanth	74	124	9160
Spotted gar - human	4	38	150
CNEs in the ParaHoxD locus (*cdx4*; Suppl. Fig. 14d)			
Spotted gar - zebrafish a	11	148	1624
Spotted gar - zebrafish b	0	-	-
Spotted gar - stickleback a	5	144	719
Spotted gar - stickleback b	0	-	-
Spotted gar - coelacanth	7	142	992
Spotted gar - human	1	97	97
CNEs in the gnathostome ParaHoxB locus (*gsx2, pdx2*; Suppl. Fig. 15)			
Spotted gar - coelacanth	12	320	3845
Spotted gar - human	11	304	3342
Spotted gar - skate	6	210	1259

Supplementary Table 6. Spotted gar circadian clock genes.

Gene names	Ensembl gene ID	Protein Length (aa)	Genome location
bmal1	ENSLOCG00000003999	678	Chromosome LG27: 8,317,763-8,344,549
bmal2	ENSLOCG00000015224	639	Chromosome LG8: 3,189,570-3,226,867
clock1	ENSLOCG00000014043	744	Chromosome LG4: 72,323,329-72,339,605
clock2	ENSLOCG00000014750	886	Chromosome LG7: 42,295,248-42,323,335
cry1a	ENSLOCG00000015272	647	Chromosome LG8: 4,053,475-4,074,670
cry1b	ENSLOCG00000011417	675	Chromosome LG3: 32,901,867-32,938,020
cry2	ENSLOCG00000014655	569	Scaffold JH591436.1: 96,765-111,323
cry3	ENSLOCG00000011465	586	Chromosome LG3: 33,014,383-33,053,389
per1	ENSLOCG00000013344	1445	Chromosome LG2: 58,185,728-58,201,428
per2	ENSLOCG00000004441	1385	Chromosome LG14: 7,862,125-7,881,568
per3	ENSLOCG00000002607	1165	Chromosome LG25: 4,785,705-4,800,981
csnk1e	ENSLOCG00000011701	273	Chromosome LG12: 35,099,178-35,103,163
dec1	ENSLOCG00000010962	409	Chromosome LG5: 27,670,723-27,673,540
dec2	ENSLOCG00000015327	422	Chromosome LG8: 4,744,466-4,746,682
nfil3 (*e4bp4-1*)	ENSLOCG00000008217	443	Chromosome LG2: 25,281,929-25,283,356
nfil3-2.1 (*e4bp4-2*)	ENSLOCG00000018299	544	Chromosome LG6: 17,036,778-17,038,412
nfil3-2.2 (*e4bp4-3*)	ENSLOCG00000018298	394	Chromosome LG6: 17,009,772-17,010,956
nr1d1	ENSLOCG00000006223	362	Chromosome LG4: 15,608,213-15,662,480
nr1d2	ENSLOCG00000006818	604	Chromosome LG11: 20,445,844-20,461,290
tef	ENSLOCG00000011595	323	Chromosome LG12: 34,848,929-34,860,886
hlf	ENSLOCG00000012233	298	Chromosome LG10: 33,240,269-33,260,238
rora	ENSLOCG00000014779	519	Chromosome LG3: 53,154,612-53,409,505
rorb	ENSLOCG00000009712	462	Chromosome LG2: 34,273,732-34,319,408
rorc	ENSLOCG00000006502	467	Chromosome LG19: 9,503,786-9,519,701
timeless	ENSLOCG00000004180	1225	Chromosome LG4: 11,892,308-11,918,271

Supplementary Table 7. Opsin genes in the gar genome and in other vertebrate lineages.

Opsin	Spotted gar genes		Mammals	Amphibians	Coelacanth	Teleosts
Visual Opsins						
LWS	1	ENSLOCG00000014714 LG1:2601700:2610864:-1	1 or 2	1	0	1
RH1	2	RH1/RHO: ENSLOCG00000018246 LG5:23736045:23737109:-1 exoRH: ENSLOCG00000013037 LG5:40450222:40456702:-1	1	1	1	2
RH2	1	ENSLOCG00000012404 LG3:36552662:36557163:-1	0	?	1	1
SWS1	2	SWS1A: ENSLOCG00000015553 LG8:10394932:10404708:-1 SWS1B: ENSLOCG00000015552 LG8:10386561:10390426:-1	1	1	0	1
SWS2	1	ENSLOCG00000014721 LG1:2621050:2624669:-1	0	1	1	1 or 2
Non-visual Opsins						
Encephalopsin	1	ENSLOCG00000016574 LG16:15519415:15526427:1	1	1	0	1
Melanopsin	3	Melanopsin X: ENSLOCG00000012921 LG2:55689526:55727565 Melanopsin M1: ENSLOCG00000012496 LG5:35802807:35857709:-1 Melanopsin M2: ENSLOCG00000018136 LG9:37927769:37929211:1	1	2	2	3
Neuropsin	4	Neuropsin A: ENSLOCG00000016365 LG1:31071707:31085860:1 Neuropsin B1: ENSLOCG00000016633 LG1:42368720:42397092:-1 Neuropsin B2: ENSLOCG00000015809 LG1:15171300:15195897:-1 Neuropsin C: ENSLOCG00000016508 LG1:37413809:37476960:1	1	1	1	4
Parapinopsin	1	ENSLOCG00000014167 LG5:48371560:48375442:-1	0	1	0	2
Parietopsin	0	not detected	0	1	0	1
Peropsin	1	ENSLOCG00000011717 LG4:48234503:48246932:1	1	1	1	1
Pinopsin	1	ENSLOCG00000001991 LG22:1260673:1267085:-1	0	1	1	0
RGR	1	ENSLOCG00000005641 LG5:7986725:7995102:1	1	1	1	2
TMT	4	TMT1A: ENSLOCG00000007008 LG14:13234488:13274930:1 TMT1B: ENSLOCG00000004577 LG14:8050700:8061874:1 TMT2: ENSLOCG00000008828 LG17:23068611:23085568:1 TMT3: ENSLOCG00000015121 LG7:46108107:46128367:1	0	3	0	3
VAL	1	ENSLOCG00000007383 LG5:11747776:11761955:-1	0	1	1	1

Supplementary Table 8. Orthologs of human MHC class II and III region genes in spotted gar. [separate .xls file]

Supplementary Table 9. Gar *scpp* gene annotation and expression. [separate .xls file]

Supplementary Table 10. Presence/absence table of miRNAs (*in silico* analysis). [separate .xls file]

Supplementary Table 11. Gar miRNA annotation based on small RNA-seq data and orthology search. [separate .xls file]

Supplementary Table 12. Evolutionary pattern of CNEs in selected gnathostome developmental gene loci.

	Number of CNEs (average size)				
Gene locus/cluster	Gnathostome CNEs[1]	Gnathostome CNEs lost in teleost fishes[2]	CNEs acquired in the bony vertebrate ancestor[3]	Bony vertebrate CNEs lost in teleost fishes[4]	CNEs acquired in the ray-finned fish ancestor[5]
HoxA	55 (146 bp)	27 (120 bp)	1 (153 bp)	1 (153 bp)	-
HoxB	30 (138 bp)	3 (69 bp)	24 (71 bp)	14 (62 bp)	3 (153 bp)
HoxC	15 (142 bp)	1 (53 bp)	19 (71 bp)	5 (62 bp)	12 (161 bp)
HoxD	76 (241 bp)	11 (123 bp)	11 (222 bp)	4 (175 bp)	14 (133 bp)
IrxB	108 (232 bp)	43 (150 bp)	47 (99 bp)	20 (68 bp)	76 (116 bp)
Pax6	106 (221 bp)	57 (160 bp)	42 (103 bp)	28 (85 bp)	6 (163 bp)

[1] CNEs present in elephant shark, human and spotted gar
[2] CNEs present in elephant shark, human and spotted gar but absent in zebrafish and stickleback
[3] CNEs present in spotted gar and human but absent in elephant shark
[4] CNEs present in spotted gar and human but absent in elephant shark and teleost fishes (zebrafish and stickleback)
[5] CNEs present in spotted gar and teleost fishes (zebrafish and stickleback) that are absent in human, coelacanth and elephant shark. The coelacanth IrxB locus is fragmented and is therefore not included in the analysis.

Supplementary Table 13. Elephant shark-human CNEs lost in spotted gar and teleost fishes.

Gene locus/cluster	Elephant shark-human CNEs	Lost in spotted gar	Lost in zebrafish		Lost in stickleback	
			a-paralog	b-paralog	a-paralog	b-paralog
HoxA	77	20 (26%)	56 (73%)	69 (90%)	47 (61%)	69 (90%)
HoxB	39	8 (20%)	10 (26%)	30 (77%)	32 (82%)	33 (85%)
HoxC	27	11 (41%)	14 (52%)	22 (81%)	13 (48%)	-
HoxD	118	33 (28%)	52 (44%)	-	46 (39%)	117 (99%)
IrxB	175	41 (23%)	125 (71%)	-	93 (53%)	-
Pax6	142	28 (20%)	97 (68%)	126 (89%)	111 (78%)	-
Average		23 (26%)	59 (56%)	62 (84%)	57 (60%)	73 (91%)

Supplementary Table 14. CNEs in the *HoxA* locus.

Supplementary Table 14a. CNEs in the *HoxA* locus predicted using elephant shark as the base.

Based on Supplementary Fig. 30a.

	Number of CNEs	Average length (bp)	Total length (bp)
Elephant shark-human	77	120	9234
Elephant shark-spotted gar	69	134	9227
Elephant shark-zebrafish a	27	88	2389
Elephant shark-zebrafish b	14	92	1292
Elephant shark-stickleback a	37	106	3913
Elephant shark-stickleback b	16	75	1197

Summary: A total of 55 CNEs are conserved between the *HoxA* loci of elephant shark, human and spotted gar (gnathostome CNEs). Of these, 27 CNEs are lost from both a- and b-copies of the zebrafish and stickleback *hoxA* loci. Of the CNEs retained in the teleost fish, the majority is present in the *hoxAa* locus.

Supplementary Table 14b. CNEs in the *HoxA* locus predicted using spotted gar as the base.

Based on Supplementary Fig. 30b.

	Number of CNEs	Average length (bp)	Total length (bp)
Spotted gar-elephant shark	42	177	7444
Spotted gar-human	35	179	6277
Spotted gar-zebrafish a	19	147	2792
Spotted gar-zebrafish b	8	165	1322
Spotted gar-stickleback a	30	163	4891
Spotted gar-stickleback b	7	156	1091

Summary: A total of 35 CNEs are conserved between the *HoxA* loci of spotted gar and human. Of these, 19 and 28 CNEs are lost from the zebrafish *hoxAa* and *hoxAb* loci, respectively. On the other hand, 13 and 30 CNEs are lost from the stickleback *hoxAa* and *hoxAb* loci, respectively.

Supplementary Table 15. CNEs in the *HoxB* locus.

Supplementary Table 15a. CNEs in the *HoxB* locus predicted using elephant shark as the base.

Base on Supplementary Fig. 31a.

	Number of CNEs	Average length (bp)	Total length (bp)
Elephant shark-human	39	110	4291
Elephant shark-spotted gar	70	104	7274
Elephant shark-zebrafish a	51	112	5699
Elephant shark-zebrafish b	17	84	1437
Elephant shark-stickleback a	13	82	1062
Elephant shark-stickleback b	17	81	1384

Summary: A total of 30 CNEs are conserved between the *HoxB* loci of elephant shark, human and spotted gar. Of these, 3 CNEs are lost from the a- and b-copies of the zebrafish and stickleback *hoxB* loci, whereas 5 CNEs are lost from the b-copy of both zebrafish and stickleback *hoxB* loci.

Supplementary Table 15b. CNEs in the *HoxB* locus predicted using spotted gar as the base.

Based on Supplementary Fig. 31b.

	Number of CNEs	Average length (bp)	Total length (bp)
Spotted gar-elephant shark	66	106	7030
Spotted gar-human	62	113	6986
Spotted gar-zebrafish a	99	134	13240
Spotted gar-zebrafish b	19	109	2066
Spotted gar-stickleback a	48	90	4341
Spotted gar-stickleback b	20	82	1645

Summary: A total of 62 CNEs are conserved between the *HoxB* loci of spotted gar and human. Of these, 14 and 53 CNEs are lost from the zebrafish *hoxBa* and *hoxBb* loci, respectively. On the other hand, 34 and 55 CNEs are lost from the stickleback *hoxBa* and *hoxBb* loci, respectively.

Supplementary Table 16. CNEs in the *HoxC* locus.

Supplementary Table 16a. CNEs in the *HoxC* locus predicted using elephant shark as the base. Based on Supplementary Fig. 32a.

	Number of CNEs	Average length (bp)	Total length (bp)
Elephant shark-human	27	102	2751
Elephant shark-spotted gar	39	96	3752
Elephant shark-zebrafish a	29	101	2934
Elephant shark-zebrafish b	18	73	1323
Elephant shark-stickleback a	28	96	2689

Summary: A total of 15 CNEs are conserved between the *HoxC* loci of elephant shark, human and spotted gar. Of these, 1 CNE is lost from the zebrafish *hoxCa* and *hoxCb* loci and the stickleback *hoxCa* locus. Of the CNEs retained in the teleost fish, the majority is present in the *hoxCa* locus.

Supplementary Tab. 16b. CNEs in the *HoxC* locus predicted using spotted gar as the base.

Based on Supplementary Fig. 32b.

	Number of CNEs	Average length (bp)	Total length (bp)
Spotted gar-elephant shark	37	97	3586
Spotted gar-human	37	110	4056
Spotted gar-zebrafish a	84	121	10136
Spotted gar-zebrafish b	29	114	3304
Spotted gar-stickleback a	72	125	8983

Summary: A total of 37 CNEs are conserved between the *HoxC* loci of spotted gar and human. Of these, 6 and 29 CNEs are lost from the zebrafish *hoxCa* and *hoxCb* loci, respectively. Of the 37 gar-human CNEs, 12 are lost from the stickleback *hoxCa* locus (stickleback has lost the *hoxCb* locus).

Supplementary Table 17. CNEs in the *HoxD* locus.

Supplementary Table 17a. CNEs in the *HoxD* locus predicted using elephant shark as the base.

Based on Supplementary Fig. 33a.

	Number of CNEs	Average length (bp)	Total length (bp)
Elephant shark-human	118	177	20926
Elephant shark-spotted gar	118	196	23084
Elephant shark-zebrafish a	98	123	12037
Elephant shark-stickleback a	95	156	14800
Elephant shark-stickleback b	17	56	960

Summary: A total of 76 CNEs are conserved between the *HoxD* loci of elephant shark, human and spotted gar. Of these, 11 CNEs are lost from the zebrafish *hoxDa* locus and stickleback *hoxDa* and *hoxDb* loci. Of the CNEs retained in the teleost fish, the majority is present in the *hoxDa* locus. The zebrafish *hoxDb* locus, which has lost all *hox* genes, has lost all associated CNEs as well (not shown in VISTA plot).

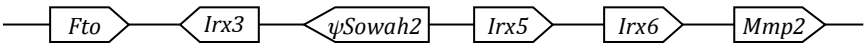
Supplementary Table 17b. CNEs in the *HoxD* locus predicted using spotted gar as the base.

Based on Supplementary Fig. 33b.

	Number of CNEs	Average length (bp)	Total length (bp)
Spotted gar-elephant shark	83	264	21871
Spotted gar-human	74	297	21970
Spotted gar-zebrafish a	100	196	19579
Spotted gar-stickleback a	106	245	25992
Spotted gar-stickleback b	-	-	-

Summary: A total of 75 CNEs are conserved between the *HoxD* loci of spotted gar and human. Of these, 17 CNEs are lost from the zebrafish *hoxDa* locus, whereas 14 are lost from the stickleback *hoxDa* locus. Additionally, none of these 75 CNEs are present in the stickleback *hoxDb* locus. The zebrafish *hoxDb* locus, which has lost all *hox* genes, has lost all associated CNEs as well (not shown in VISTA plot).

Supplementary Table 18. CNEs in the *IrxB* locus.



The *IrxB* (*Iroquois B*) gene cluster is composed of genes *Irx3*, *Irx5* and *Irx6*. For CNE prediction, we used the sequence spanning the three *IrxB* genes plus the immediate flanking genes *Fto* (*fat mass and obesity-associated protein*) and *Mmp2* (*matrix metalloproteinase 2*). Teleost fishes have two *irxB* loci (*irxBa* and *irxBb*) as a result of the TGD. In the zebrafish *irxBa* locus (located on chromosome 7), *irx3a* is separated from *irx5a* and *irx6a* by ~15 intervening genes; *irx3a* is therefore not included in the CNE analysis. The zebrafish *irxBb* locus on the other hand lacks *irx6b*; in addition, *irx3b* and *irx5b* are separated by more than 10 genes (~300 kb). The *irxBb* locus is therefore excluded from the CNE analysis. Stickleback and medaka possess only the *irxBa* locus, as the *irxBb* locus has been lost secondarily.

Supplementary Table 18a. CNEs in the *IrxB* locus predicted using elephant shark as the base.

Based on Supplementary Fig. 34a.

	Number of CNEs	Average length (bp)	Total length (bp)
Elephant shark-human	175	166	29032
Elephant shark-spotted gar	267	175	46817
Elephant shark-zebrafish a	122	111	13494
Elephant shark-stickleback a	175	126	21975

Summary: A total of 108 CNEs are conserved between the *IrxB* loci of elephant shark, human and spotted gar (gnathostome CNEs). Of these, 68 CNEs are lost from the zebrafish *irxBa* locus whereas 46 CNEs are lost from the stickleback *irxBa* locus.

Supplementary Table 18b. CNEs in the *IrxB* locus predicted using spotted gar as the base.

Based on Supplementary Fig. 34b.

	Number of CNEs	Average length (bp)	Total length (bp)
Spotted gar-elephant shark	268	175	46789
Spotted gar-human	180	175	31512
Spotted gar-zebrafish a	280	120	33568
Spotted gar-stickleback a	436	136	59170

Summary: A total of 180 CNEs are conserved between the *irxB* loci of spotted gar and human. Of these, 105 CNEs are lost from the zebrafish *irxBa* locus whereas 54 CNEs are lost from the stickleback *irxBa* locus.

Supplementary Table 19. CNEs in the *Pax6* locus.

The greater *Pax6* (*Paired box 6*) locus comprises the *Pax6* gene and its flanking genes *Wt1* (*Wilms tumor 1*), *Rcn1* (*reticulocalbin 1*), *Elp4* (*Elongation protein 4*). For analysis of CNEs, we used an extended region for alignment spanning all the way from *Eif3m* (*Eukaryotic translation initiation factor 3, subunit M*) to *Dcdc1* (*Doublecortin domain containing 1*). In elephant shark the canonical form of *Pax6* is known as *Pax6.1*. Amongst teleost fishes, zebrafish retains duplicate copies of the pax6 locus (*pax6a* and *pax6b*) whereas stickleback possesses a single *pax6* locus.

Supplementary Table 19a. CNEs in the *Pax6* locus predicted using elephant shark as the base. Based on Supplementary Fig. 35a.

	Number of CNEs	Average length (bp)	Total length (bp)
Elephant shark-human	142	178	25215
Elephant shark-spotted gar	193	171	33039
Elephant shark-stickleback a	49	102	4988
Elephant shark-zebrafish a	95	119	11358
Elephant shark-zebrafish b	33	88	2911

Summary: A total of 106 CNEs are conserved between the *Pax6* loci of elephant shark, human and spotted gar (gnathostome CNEs). Of these, 80 CNEs are lost from the single *pax6* locus in stickleback, whereas 65 and 90 CNEs are lost from the duplicate zebrafish *pax6a* and *pax6b* loci, respectively.

Supplementary Table 19b. CNEs in the *Pax6* locus predicted using spotted gar as the base.

Based on Supplementary Fig. 35b.

	Number of CNEs	Average length (bp)	Total length (bp)
Spotted gar-elephant shark	201	168	33774
Spotted gar-human	144	201	28933
Spotted gar-stickleback a	109	115	12558
Spotted gar-zebrafish a	178	144	25604
Spotted gar-zebrafish b	62	99	6127

Summary: A total of 144 CNEs are conserved between the *Pax6* loci of spotted gar and human. Of these, 101 CNEs are lost from the single *pax6* locus in stickleback, whereas 75 and 119 CNEs are lost from the duplicate zebrafish *pax6a* and *pax6b* loci, respectively.

Supplementary Table 20. Summary statistics of three whole genome alignments (WGAs).

	# CNEs	fraction of CNEs	average length	total length	fraction of CNE length	fraction of genome
Gar-centric WGA						
total	156,087	100.0%	240bp	37,480,218bp	100.0%	3.96%
GCNEs	43,765	28.0%	235bp	10,288,819bp	27.5%	1.09%
BCNEs	78,889	50.5%	242bp	19,081,754bp	50.9%	2.02%
RCNEs	33,433	21.4%	243bp	8,109,645bp	21.6%	0.86%
Zebrafish-centric WGA						
total	239,485	100.0%	175bp	41,838,831bp	100.0%	2.96%
GCNE	41,572	17.4%	165bp	6,859,877bp	16.4%	0.49%
BCNE	110,694	46.2%	170bp	18,781,769bp	44.9%	1.33%
RCNE	41,207	17.2%	185bp	7,637,038bp	18.3%	0.54%
TelCNE	46,011	19.2%	186bp	8,560,147bp	20.5%	0.61%
Human-centric WGA						
total	146,988	100.0%	193bp	28,309,482bp	100.0%	0.91%
GCNE	30,578	20.8%	203bp	6,219,485bp	22.0%	0.20%
BCNE	52,540	35.7%	172bp	9,059,225bp	32.0%	0.29%
LCNE	13,804	9.4%	191bp	2,638,662bp	9.3%	0.09%
TetraCNE	8,093	5.5%	204bp	1,647,732bp	5.8%	0.05%
ACNE	41,973	28.6%	208bp	8,744,378bp	30.9%	0.28%

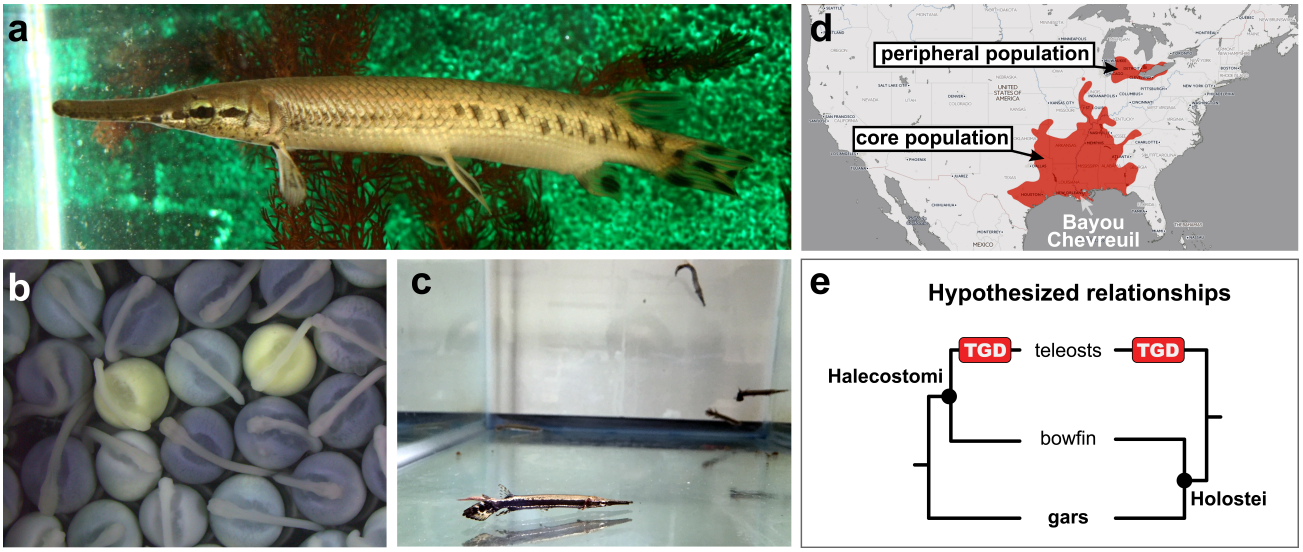
Supplementary Table 21. CNE and GWAS-SNP connectivity from human to zebrafish through gar.

		# CNEs	CNE fraction	GWAS SNPs	GWAS-SNP-containing CNEs
Human-centric CNEs (aligning ≥1 lobe-finned vertebrate)					
total		143,525	100.0%	25,555	21,564
connected to gar		39,964	27.8%	6,770	5,661
connected to ≥1 teleost		54,599	38.0%	7,770	6,650
connected to zebrafish		34,133	23.8%	5,314	4,458
connected to gar, not to any teleost		19,149	13.3%	3,501	2,932
Human CNEs (aligning ≥1 lobe-finned vertebrate) in gar with no connection to teleost					
total		18,994	100.0%	3,480	2,915
zebrafish connectivity:					
	1bp intersect in gar	6,838	36.0%	1,217	1,021
	10% intersect in gar	6,602	34.8%	1,164	980
	20% intersect in gar	6,310	33.2%	1,117	939
	33% intersect in gar	5,761	30.3%	992	848

Supplementary Table 22. Analysis of human limb enhancer evolution informed by gar. [separate .xls file]

Supplementary Table 23. TGD ohnologs and singletons in zebrafish and medaka and their gar ortholog. [separate .xls file]

C. SUPPLEMENTARY FIGURES

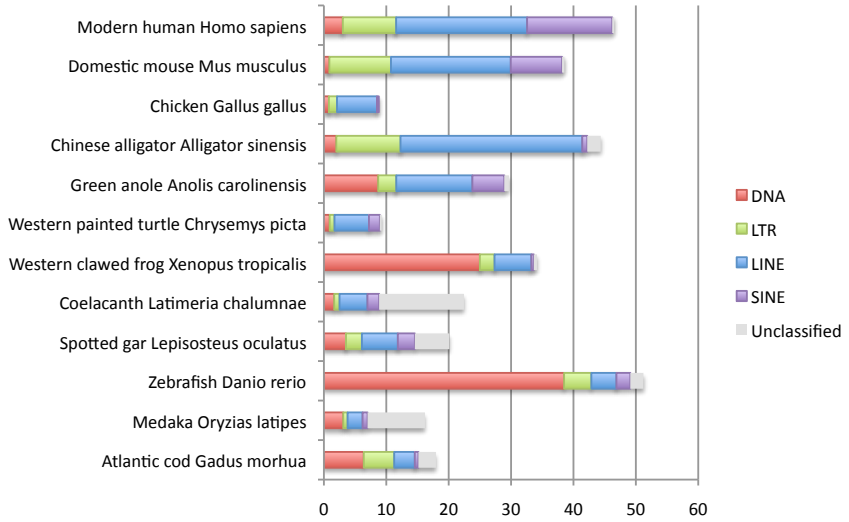


Supplementary Figure 1. Spotted gar (*Lepisosteus oculatus*). a) A laboratory grown adult spotted gar male ('Garfield'). b) Laboratory spawned spotted gar embryos (2.5 days post fertilization; dorsal view). c) Laboratory grown spotted gar juveniles. d) Biogeography of spotted gar in two main populations: core (southern US) and peripheral (northern US and south central Canada). Specimens for the present study were caught in Bayou Chevreuil, Louisiana. Species map was obtained from MAP OF LIFE [http://www.mol.org/] using data by Page and Burr (2011)[6]. e) Alternative hypotheses for the phylogenetic relationships of neopterygian ray-finned fishes. Monophyly of Halecostomi (teleosts + bowfin, *Amia calva*, left) is mostly supported by morphological analyses, while molecular phylogenetics tend to support the monophyly of Holostei (gars + bowfin, right); after Grande (2010)[9]. Photo credit: a, T. D.; b, c, I. B.

A

Species	DNA	LTR	LINE	SINE	Unclassified	Total TEs
Atlantic cod *Gadus morhua*	6.39	4.88	3.3	0.58	2.81	17.96
Medaka *Oryzias latipes*	3.1	0.7	2.4	0.8	9.2	16.2
Zebrafish *Danio rerio*	38.5	4.37	4.02	2.22	2.1	51.21
Spotted gar *Lepisosteus oculatus*	3.54	2.55	5.76	2.73	5.54	20.12
Coelacanth *Latimeria chalumnae*	1.65	0.86	4.47	1.87	13.6	22.45
Western clawed frog *Xenopus tropicalis*	25	2.35	5.9	0.4	0.5	34.15
Western painted turtle Chrysemys picta	0.93	0.77	5.52	1.75	0.14	9.11
Green anole *Anolis carolinensis*	8.7	2.9	12.2	5.11	0.75	29.66
Chinese alligator *Alligator sinensis*	1.99	10.29	29.13	0.8	2.16	44.37
Chicken *Gallus gallus*	0.8	1.3	6.4	0.1	-	8.6
Domestic mouse *Mus musculus*	0.88	9.87	19.2	8.22	0.38	38.55
Modern human *Homo sapiens*	3.03	8.55	20.99	13.64	0.15	46.36

B

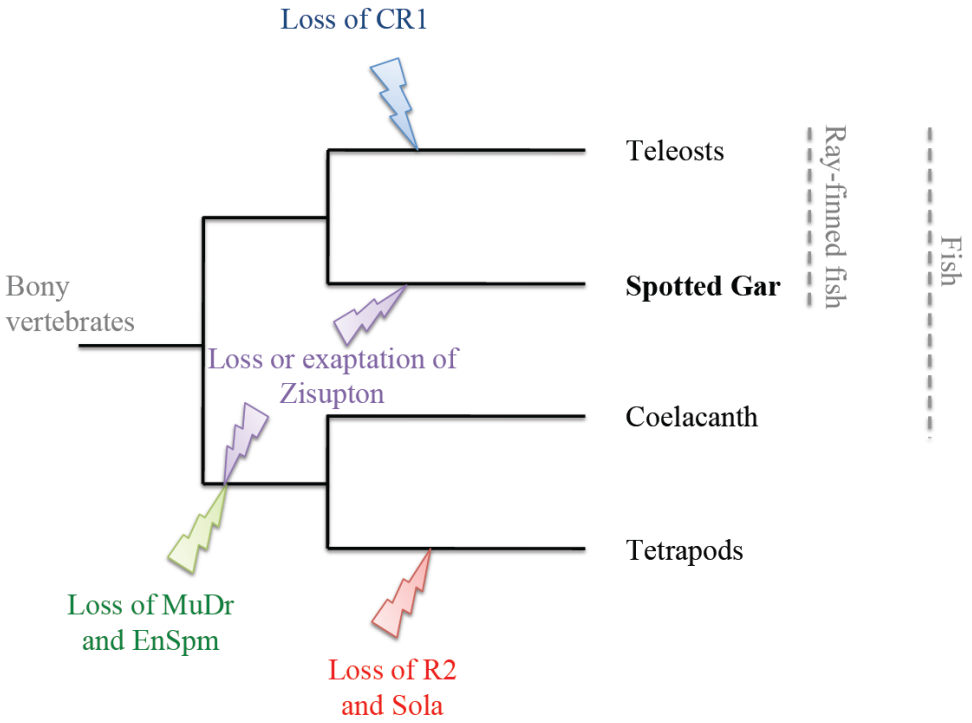


C

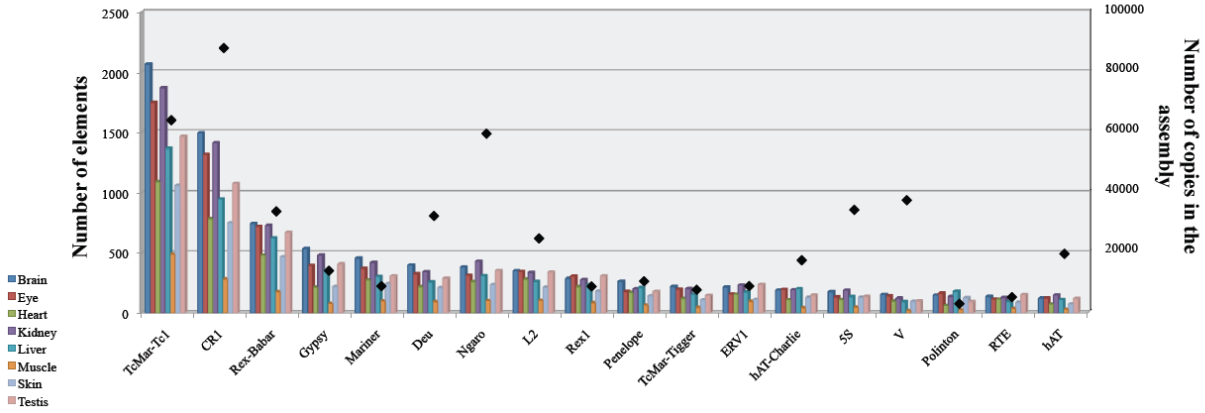
Species	DNA											LTR					LINE								
	Helitron	Polinton	Zisupton	Tc-Mariner	hAT	Harbinger	PiggyBac	Sola	MuDr	EnSpm	Merlin	Kolobok	Gypsy	BEL	Copia	ERV	DIRS	Penelope	R4	R2	LINE1/Tx1	RTE/Rex3	CR1/LINE3	LINE2/Maui	Rex1/Babar	Jockey/Nimb
Fugu pufferfish *Takifugu rubripes*			x	x	x	x	x						x	x	x	x	x	x	x	x	x	x		x	x	x
Platyfish *Xiphophorus maculatus*	x	x	x	x	x	x	x		x				x	x		x	x	x	x	x	x	x		x	x	x
Zebrafish *Danio rerio*	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x		x	x	x
Spotted gar *Lepisosteus oculatus*	x	x		x	x	x	x	x	x	x		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Coelacanth *Latimeria chalumnae*	x	x		x	x	x	x	x					x		x	x	x	x	x	x	x	x	x	x		x
Western clawed frog *Xenopus tropicalis*	x	x		x	x	x	x				x	x	x	x	x	x	x	x			x		x	x	x	
Western painted turtle Chrysemys picta				x	x	x	x						x			x	x	x			x	x	x	x	x	x
Green anole *Anolis carolinensis*	x	x		x	x								x	x	x	x	x	x	x		x	x	x	x		x
Chicken *Gallus gallus*				x	x		x									x		x			x		x			
Domestic mouse *Mus musculus*				x	x											x					x		x	x		
Modern human *Homo sapiens*				x	x		x									x					x		x	x		

TGD

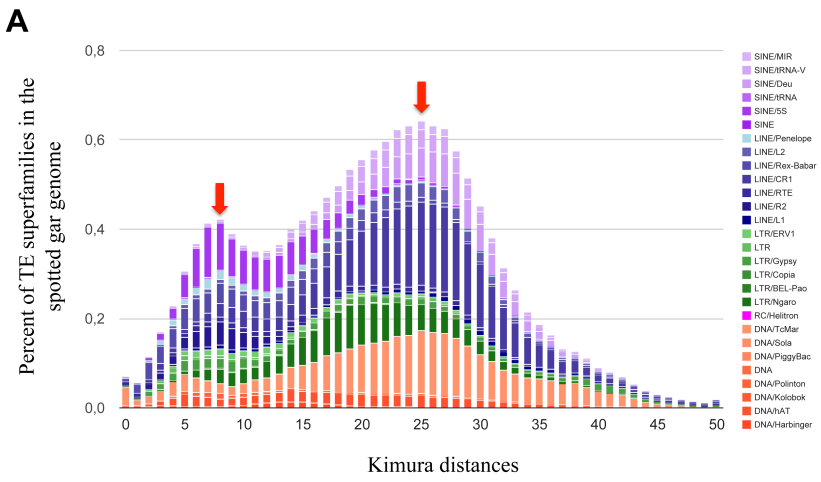
Supplementary Figure 2. TE class abundance in gar compared to other bony vertebrates. Comparison of the abundance of the four classes of TE (DNA, LTR, LINE and SINE) in different vertebrate genomes including spotted gar. Values of the classes and diversity (presence/absence) of other elements are based on literature when available (Atlantic cod[203]; medaka[53]; zebrafish[52]; Coelacanth[57]; Western clawed frog[204]; Western painted turtle[58]; green anole[205]; Chinese alligator[206]; chicken[207]; mouse[208]; human[51]; fugu[209]; platyfish[182]). a) Percentage of genome covered by the four TE classes. b) Histogram of TE content profile. c) Presence of autonomous TE superfamilies (x) in gar and other species (data from literature, see citations above). The red arrow indicates the teleost genome duplication event (TGD).

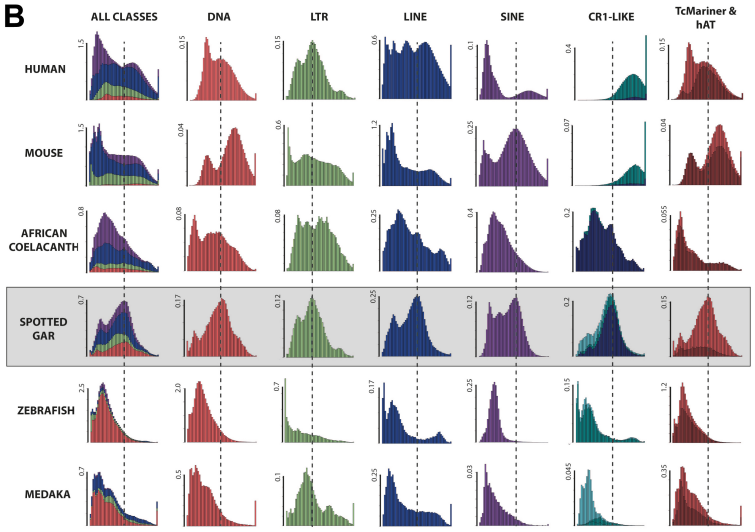


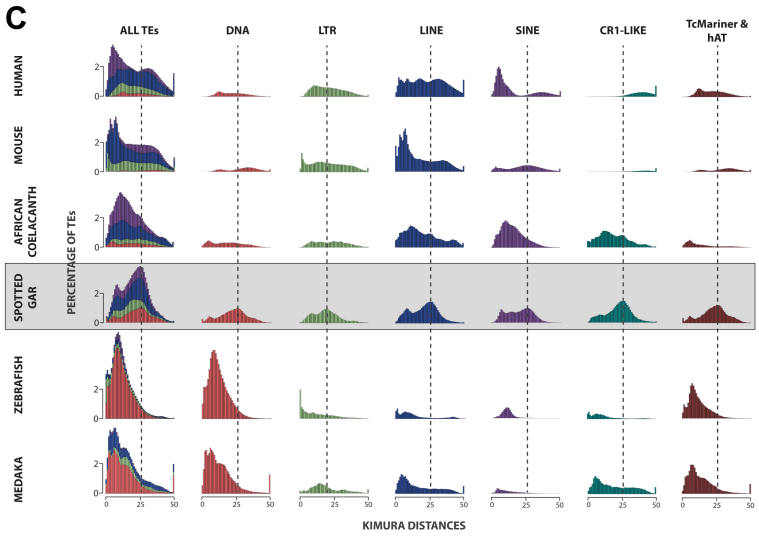
Supplementary Figure 3. Gar informs TE superfamily losses in bony vertebrates. Lineage-specific TE superfamily losses based on the hypothesis that these superfamilies were present in the ancestral bony vertebrate genome and were vertically transmitted.



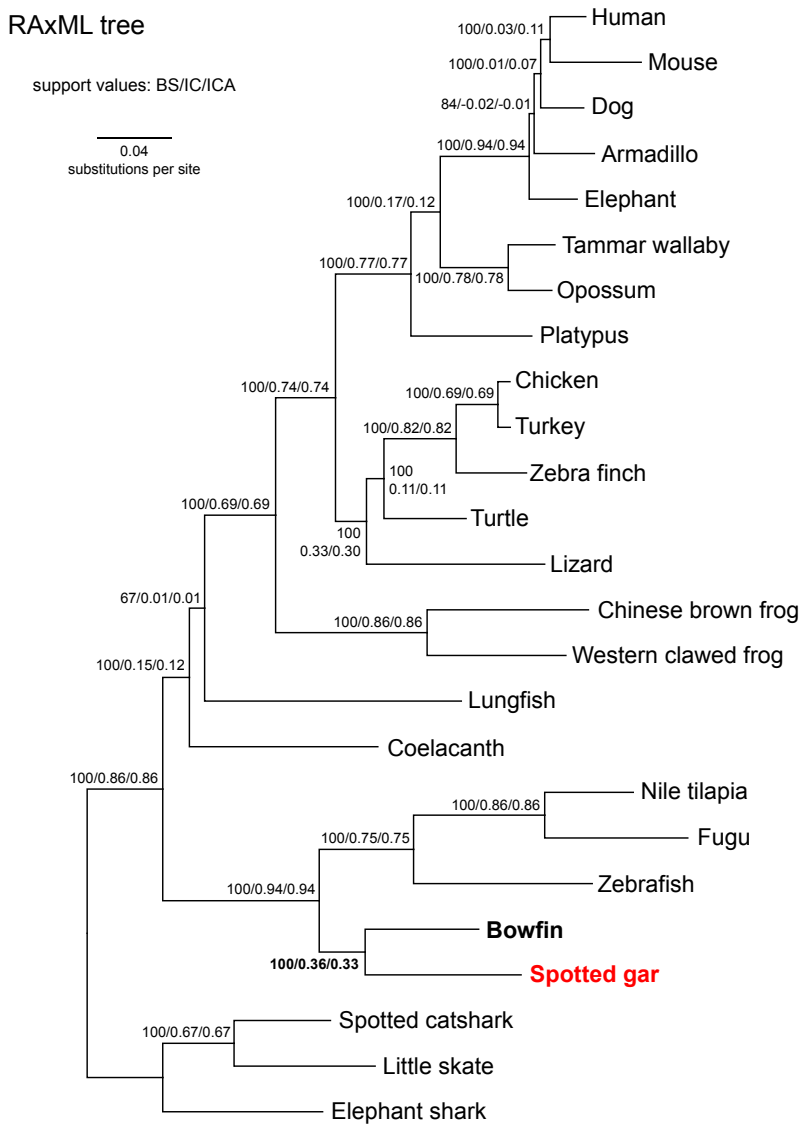
Supplementary Figure 4. Transcriptional activity of gar transposable elements. Number of TEs in transcriptomes and genomic copies of the most represented superfamilies in the transcriptome.


A
Percent of TE superfamilies in the spotted gar genome
0,8
0,6
0,4
0,2
0,0
0
5
10
15
20
25
30
35
40
45
50
Kimura distances
SINE/MIR
SINE/tRNA-V
SINE/Deu
SINE/tRNA
SINE/5S
SINE
LINE/Penelope
LINE/L2
LINE/Rex-Babar
LINE/CR1
LINE/RTE
LINE/R2
LINE/L1
LTR/ERV1
LTR
LTR/Gypsy
LTR/Copia
LTR/BEL-Pao
LTR/Ngaro
RC/Helitron
DNA/TcMar
DNA/Sola
DNA/PiggyBac
DNA
DNA/Polinton
DNA/Kolobok
DNA/hAT
DNA/Harbinger

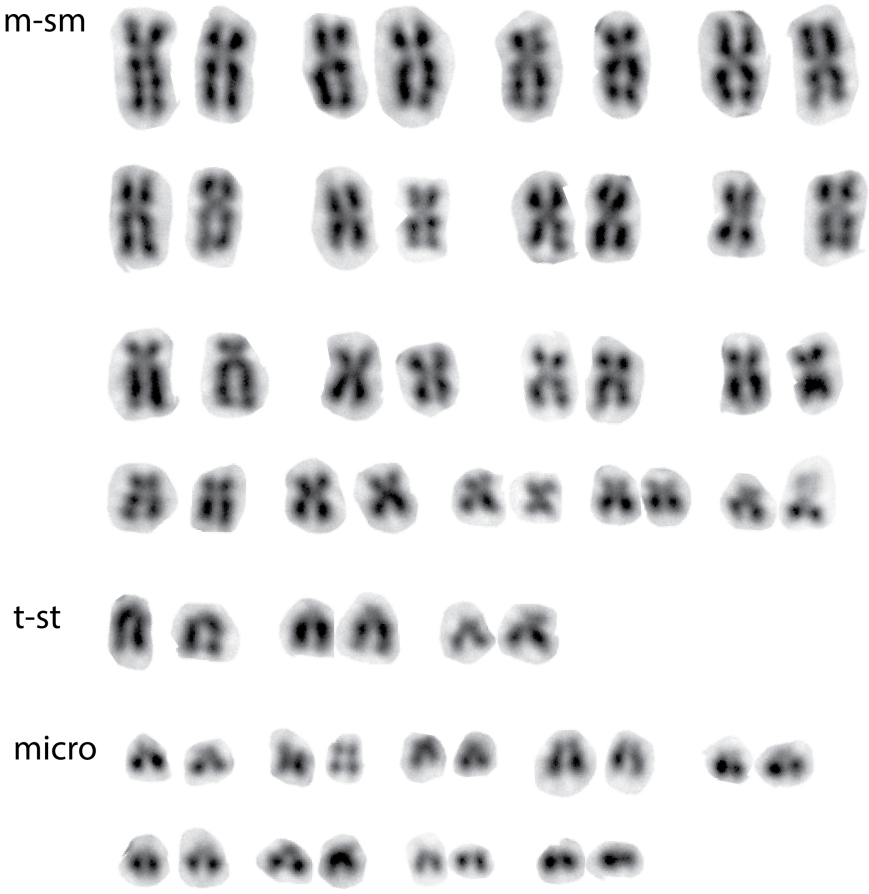

B
ALL CLASSES
DNA
LTR
LINE
SINE
CR1-LIKE
TcMariner & hAT
HUMAN
MOUSE
AFRICAN COELACANTH
SPOTTED GAR
ZEBRAFISH
MEDAKA


C
ALL TEs
DNA
LTR
LINE
SINE
CR1-LIKE
TcMariner & hAT
HUMAN
MOUSE
AFRICAN COELACANTH
SPOTTED GAR
ZEBRAFISH
MEDAKA
PERCENTAGE OF TEs
KIMURA DISTANCES

Supplementary Figure 5. Age profile of TE superfamilies in the spotted gar genome. a) Two major peaks of TE activity at Kimura distances around 8 and 25 (red arrows). b) Distribution of transposable elements (expressed as % of the genome) based on Kimura-distance analyses in six vertebrate species. Graphs represent genome percentage constituted by TEs (Y-axis) in six vertebrate genomes (Human, Mouse, African coelacanth, spotted gar, zebrafish and medaka), for several TE groups (DNA, LTR, LINE and SINE) and TE superfamilies (CR1-like that contains CR1, L2 and Rex1-Babar retrotransposons; TcMariner and hAT DNA transposons) clustered according to Kimura distances (from 0 to 50; X-axis). CR1-like superfamily as well as TcMariner and hAT data have been extracted from LINE and DNA groups, respectively. To obtain a better view of the peaks, each graph presents its own Y-axis scale. The black dashed line highlights the spotted gar major peak for comparison. c) Distribution of transposable elements (expressed as % of total TEs) based on Kimura-distance analyses in six vertebrate species. Graphs represent the proportion of TE groups and superfamilies within total TE content (excluding "Unknown" elements) (Y-axis) in each six vertebrate genomes (Human, Mouse, African coelacanth, spotted gar, zebrafish and medaka), according to Kimura distances (from 0 to 50; X-axis). The black dashed line highlights the spotted gar oldest peak for each group and superfamily comparison.

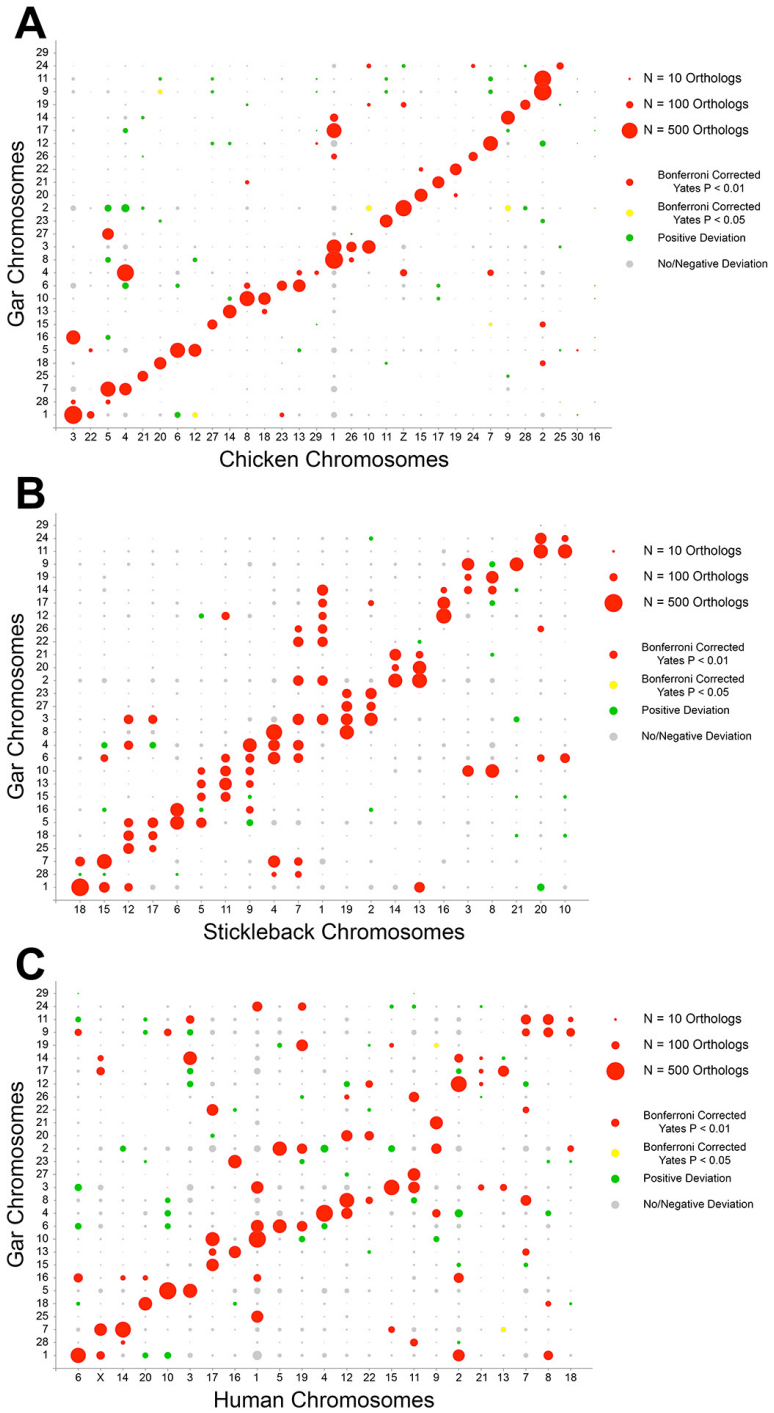


Supplementary Figure 6. Maximum likelihood analysis of spotted gar phylogenetic relationships. RAxML phylogeny with model JTT+F+Γ4. Support values on nodes include percent support from 100 bootstrap (BS) replicates / internode certainty (IC) of each node / and the extended IC (ICA). The tree strongly supports the monophyly of Holostei (gar+bowfin) as the sister lineage of teleosts (see also Supplementary Note 6).

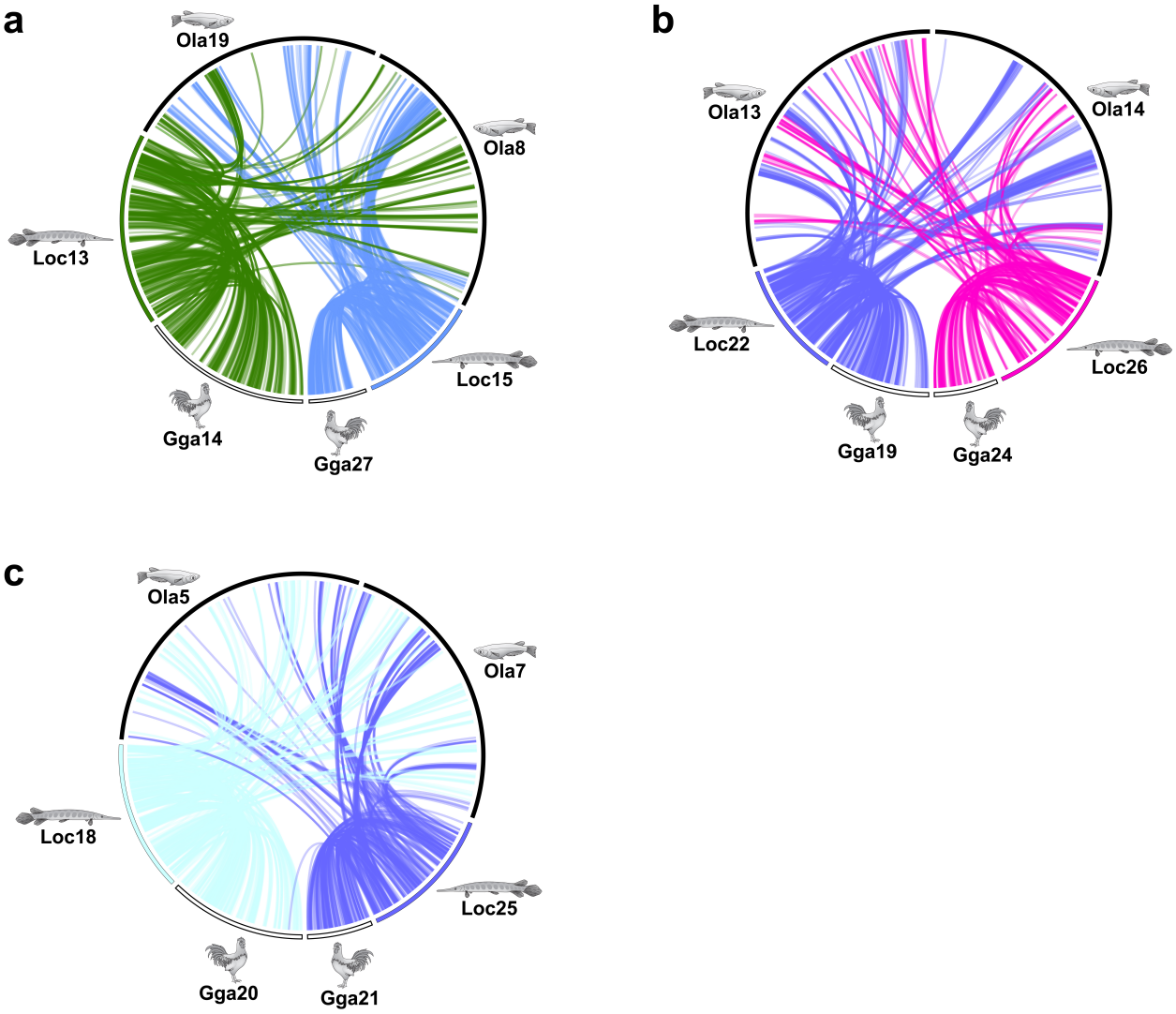


Supplementary Figure 7. Karyotype of the spotted gar. The spotted gar genome (2N = 58) consists of 29 pairs of chromosomes: 17 pairs of metacentric-submetacentric (m-sm) chromosomes, three pairs of telocentric-subtelocentric chromosomes (t-st), and nine pairs of microchromosomes (micro).

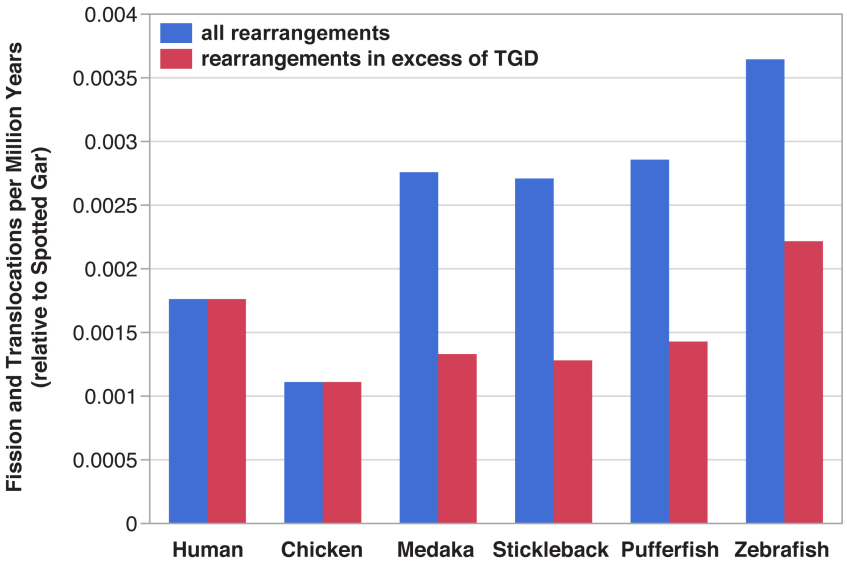
Supplementary Figure 8. Synteny dotplots of gar linkage groups against medaka, chicken, and human. Dotplots of gar (Loc) vs. chromosomes from medaka (Ola,‘), chicken (Gga,’’), and human (Hsa,’’’) generated with the Synteny Database[86]. Red crosses indicate (co-) orthologous genes. Gene orders follow the order in the gar genome assembly. The very small linkage group 29, uninformative for synteny analysis, was not included.
[single separate .pdf file]



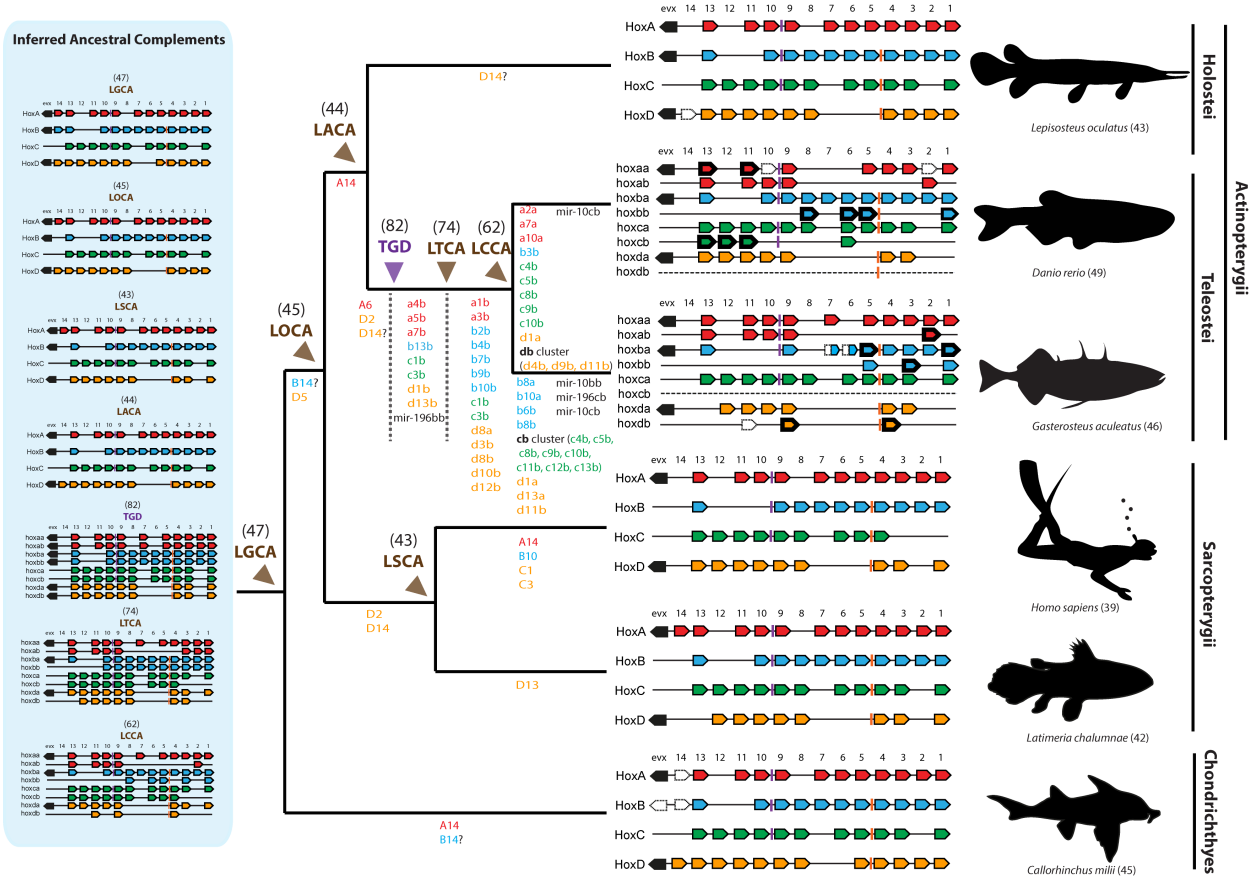
Supplementary Figure 9. Distribution of conserved syntenic regions in gar vs. three bony vertebrates. a) Comparison of gar and chicken genomes. b) Comparison of genomes of gar and stickleback, a percomorph teleost derived from the TGD event. c) Comparison of gar and human genomes. The size of each circle is proportional to the number of orthologous genes located on the corresponding gar linkage group and reference chromosome. The color of each circle represents the degree to which the number of observed orthologs deviates from null expectations under a uniform distribution across an identical number of linkage groups, chromosomes and genes per linkage group and chromosome.



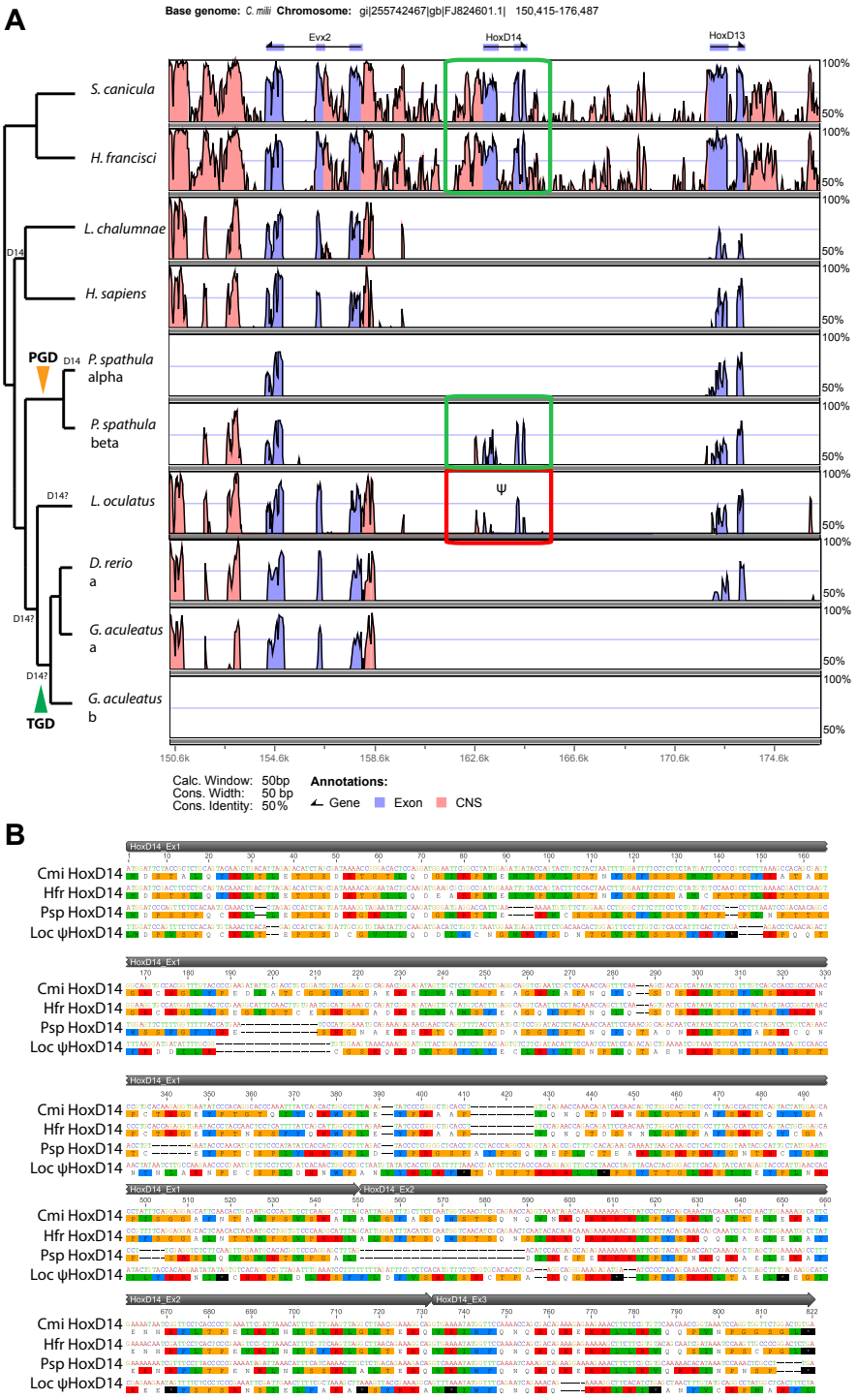
Supplementary Figure 10. Pre-TGD chromosome fusions in the teleost lineage. Circos plots of pairs of microchromosomes that show 1:1 conserved synteny between gar (Loc) and chicken (Gga) and double conserved synteny to the same pair of medaka (Ola) chromosomes. Examples include a) Loc13/Gga14 and Loc15/Gga27 to Ola19 and Ola8; b) Loc22/Gga19 and Loc26/Gga24 to Ola13 and Ola14, and c) Loc18/Gga20 and Loc25/Gga21 to Ola5 and Ola7. These patterns provide evidence for chromosome fusions in the teleost lineage after divergence from gar and that these fusions were followed by the TGD, similar to the case shown in Figure 2f of the main text.



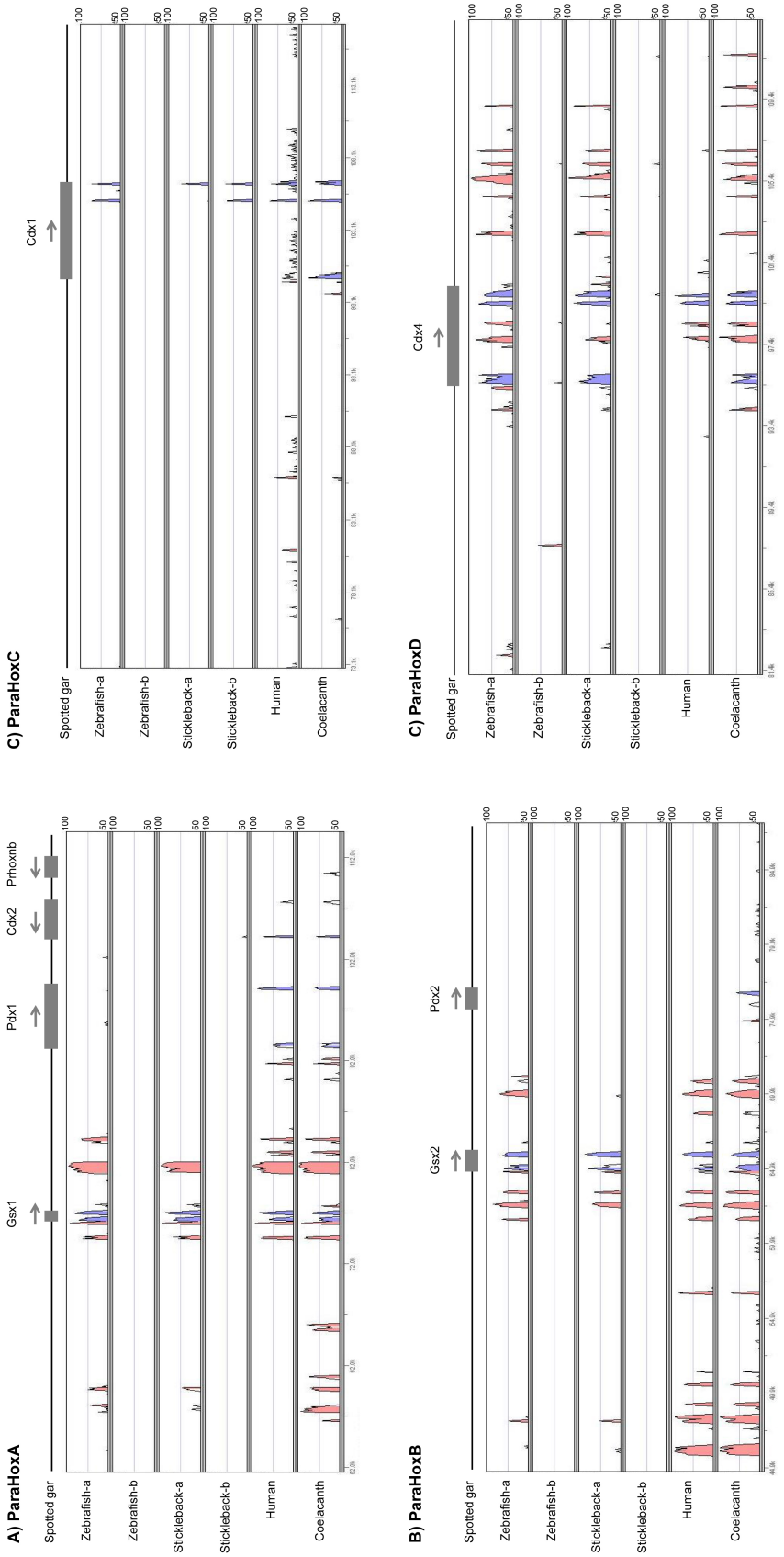
Supplementary Figure 11. Chromosomal rearrangement rates between gar and six bony vertebrates. Rearrangement rates relative to gar were determined including (blue) and excluding (red) the effects of the teleost genome duplication (TGD).



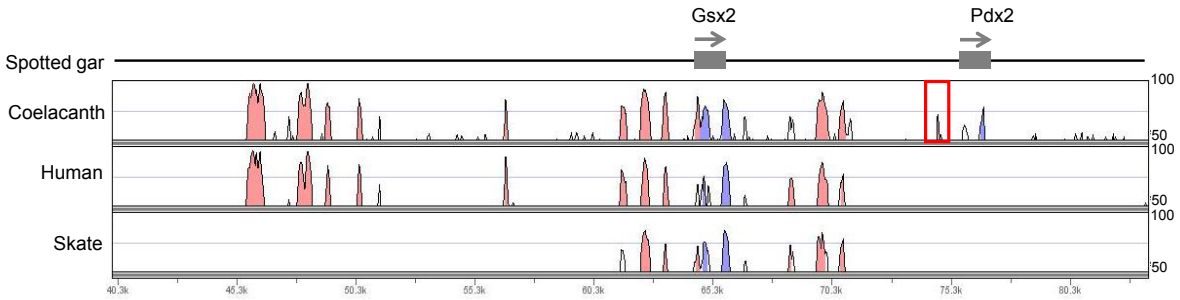
Supplementary Figure 12. Spotted gar Hox gene clusters compared to other vertebrate lineages. Gene losses (colored gene names) are mapped onto the phylogeny. Inferred Hox gene complements of the last common ancestors of gnathostomes (LGCA), osteichthyans (LOCA), sarcopterygians (LSCA), actinopteryigans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



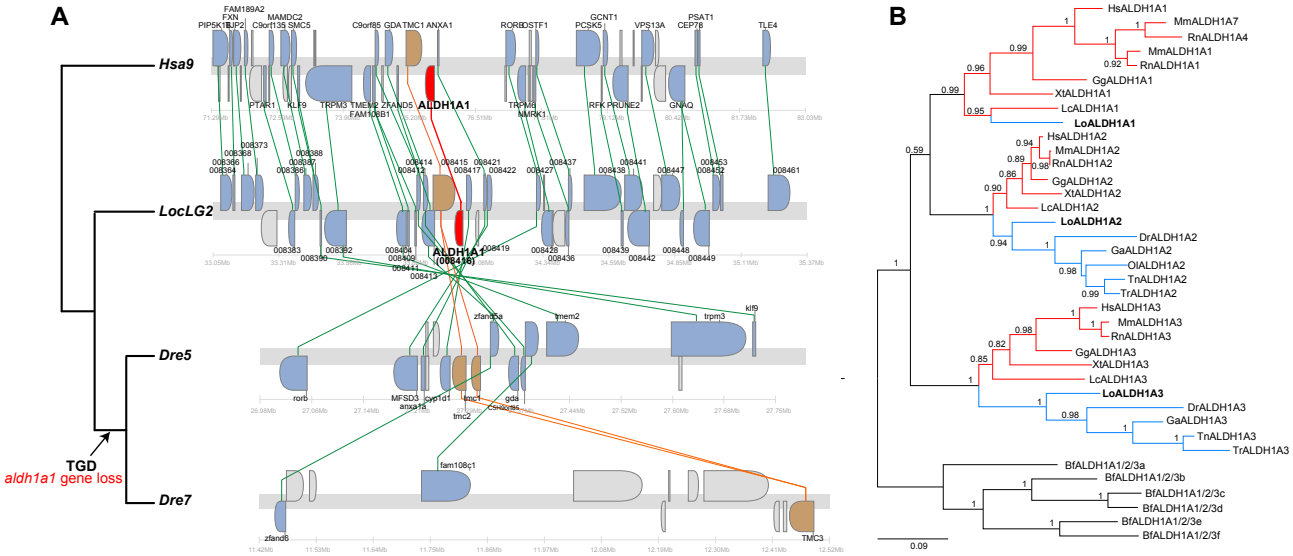
Supplementary Figure 13. The spotted gar *hoxD14* pseudogene. a) VISTA plot of the *Evx2-HoxD14-HoxD13* region in elephant shark (*C. milii*) against the (co)-orthologous regions from lesser spotted catshark (*S. canicula*), horn shark (*H. francisci*), coelacanth (*L. chalumnae*), human (*H. sapiens*),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



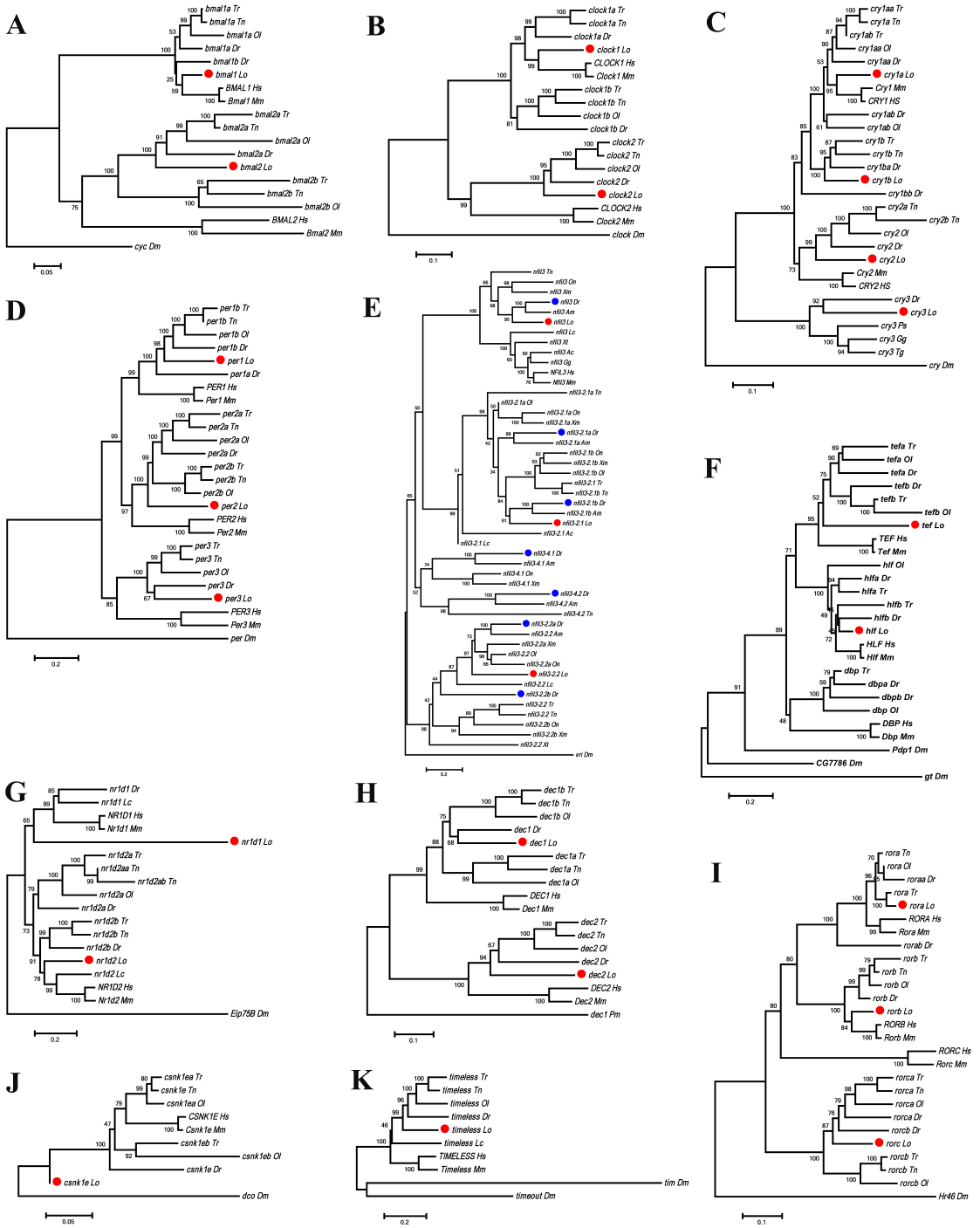
Supplementary Figure 14. VISTA plot of the ParaHox loci using spotted gar as the base. a) ParaHoxA locus. b) ParaHoxB locus. c) ParaHoxC locus. d) ParaHoxD locus. SLAGAN alignment; CNE definition: >65% identity, ≥50bp window size. Blue peaks, exons; pink peaks, CNEs.



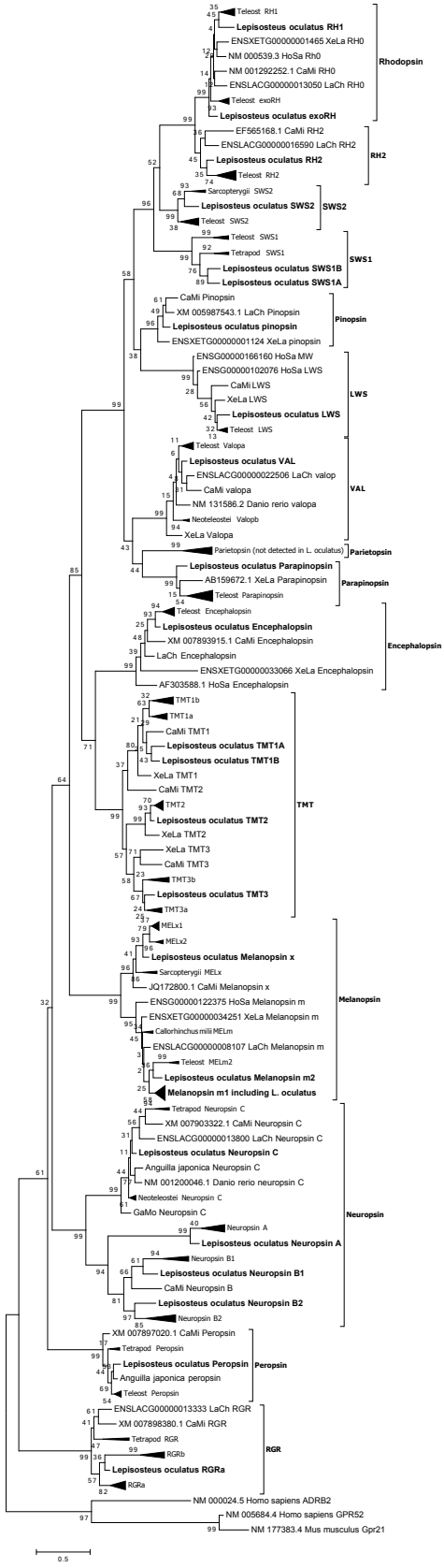
Supplementary Figure 15. VISTA plot of the gnathostome ParaHoxB locus using spotted gar as the base. A CNE conserved between gar and coelacanth is the first putative regulatory element of a gnathostome *Pdx2* gene identified (red box). SLAGAN alignment; CNE definition: >65% identity, ≥50bp window size. Blue peaks, exons; pink peaks, CNEs.



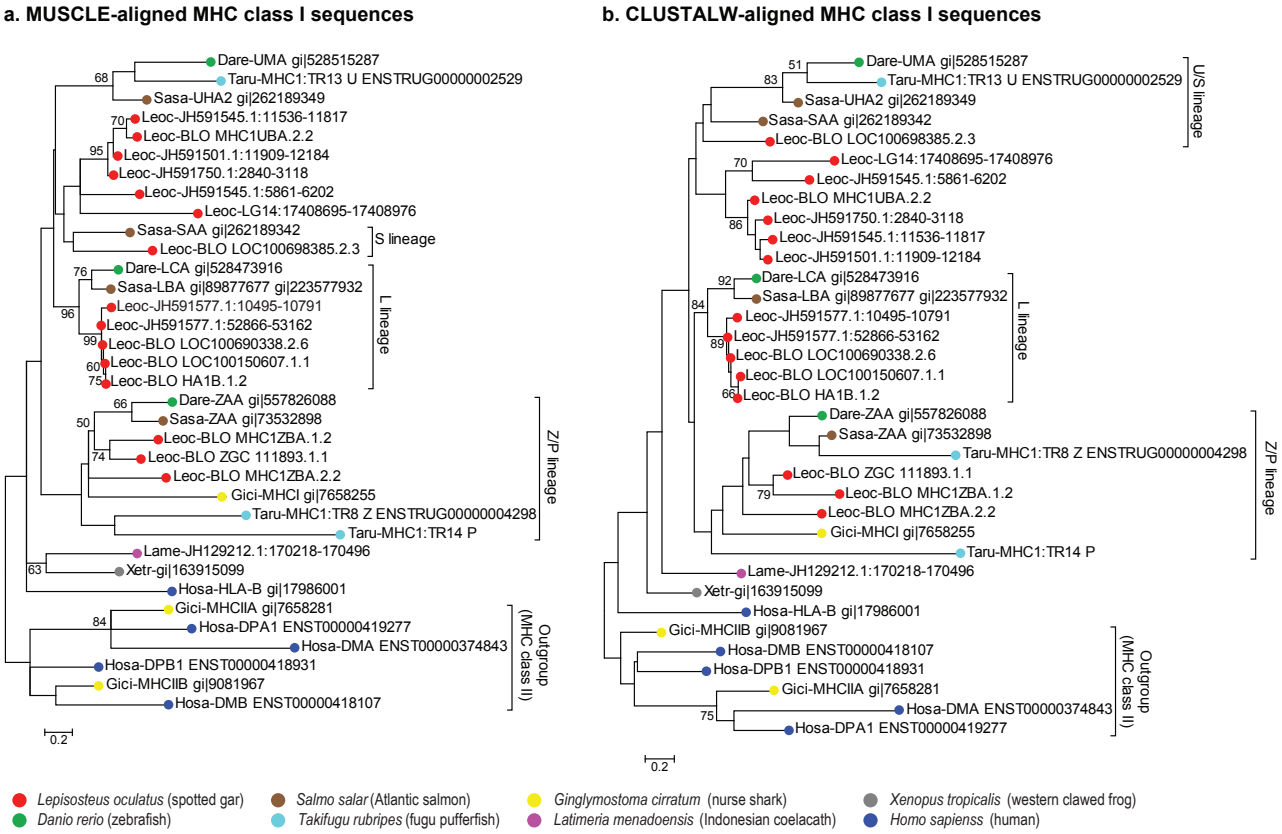
Supplementary Figure 16. Aldh1a1 evolution and identification of ohnologs gone missing in bony vertebrates. a) Conserved synteny analysis reveals the presence of gar *aldh1a1* on gar linkage group 2 (LocLG2) in conserved synteny with human *ALDH1A1* on chromosome Hsa9 and loss of *aldh1a1* in teleosts (although TGD ohnologons that formerly would have had *aldh1a1* can still be identified for instance in zebrafish on chromosomes Dre5 and Dre7, in which *aldh1a1* ohnologs should have been located). b) Maximum likelihood phylogeny of bony vertebrate Aldh1a proteins showing aLRT SH-like branch support at nodes. Ray-finned fish (blue branches): Lo; spotted gar; Dr, zebrafish; Ol, medaka; Tn, Tetraodon; Tr, fugu. Lobe-finned vertebrates (red branches): Lc, coelacanth; Hs, human; Mm, mouse; Rn, rat; Gg, chicken; Xt, frog. Outgroup sequences used for rooting: Bf, amphioxus.



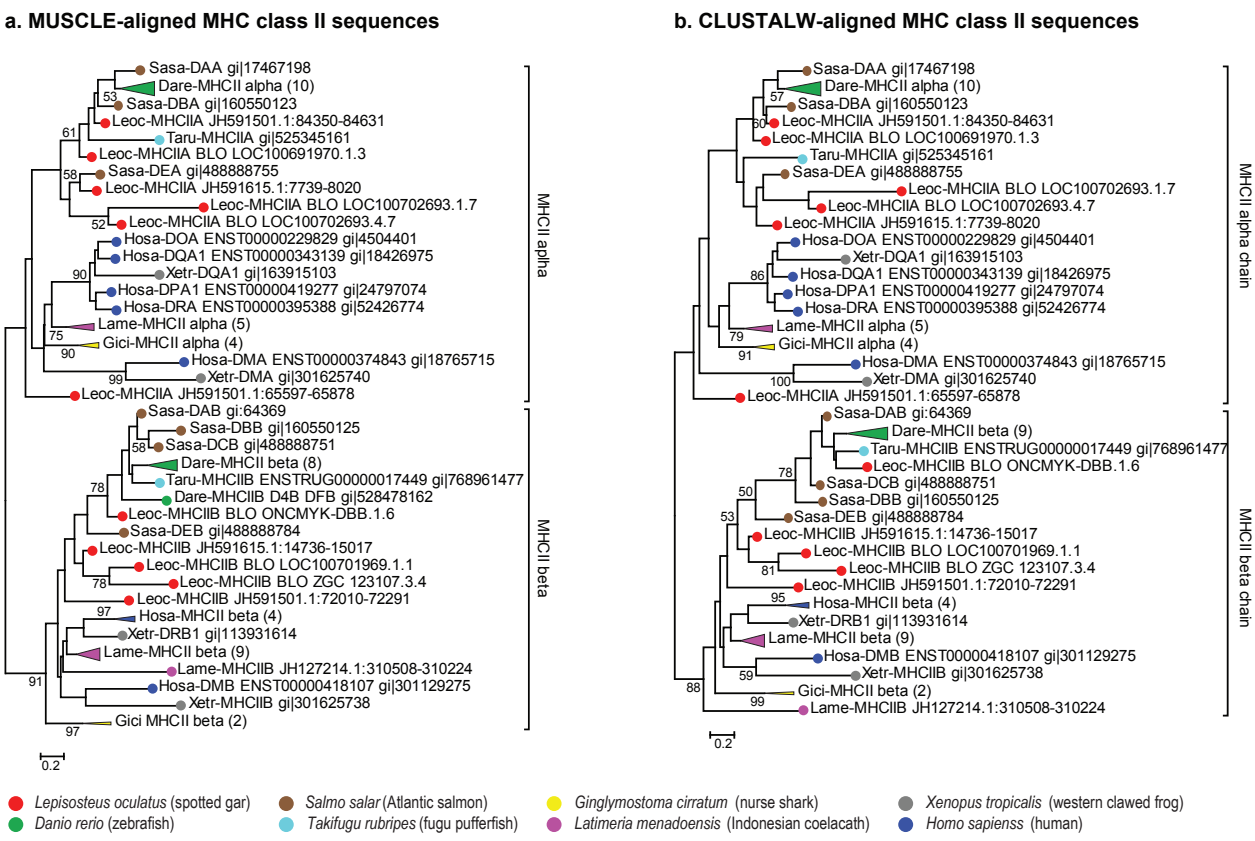
Supplementary Figure 17. Phylogenetic trees of circadian clock genes. a) *bmal* genes; b) *clock* genes; c) *cry* genes; d) *per* genes; e) *nfil3* (*e4bp4*) genes; f) *par* family genes; g) *nr1d* genes; h) *dec* genes; i) *ror* genes; j) *csnk1e* genes; and k) *timeless* genes. The trees were constructed by neighbor-joining with MEGA6[106]. Numbers on the nodes are percent support values based on 1,000 bootstrap replicates. Ac, *Anolis carolinensis*; Am, *Astyanax mexicanus*; Dm, *Drosophila melanogaster*; Dr, *Danio rerio*; Gg, *Gallus gallus*; Hs, *Homo sapiens*; Lc, *Latimeria chalumnae*; Lo, *Lepisosteus oculatus* (red dots); Mu, *Mus musculus*; Oa, *Ornithorhynchus anatinus*; Ol, *Oryzias latipes*; On, *Oreochromis niloticus*; Pm, *Petromyzon marinus*; Ps, *Pelodiscus sinensis*; Tg, *Taeniopygia guttata*; Tn, *Tetraodon nigroviridis*; Tr, *Takifugu rubripes*; Xm, *Xiphophorus maculatus*; Xt, and *Xenopus tropicalis*.



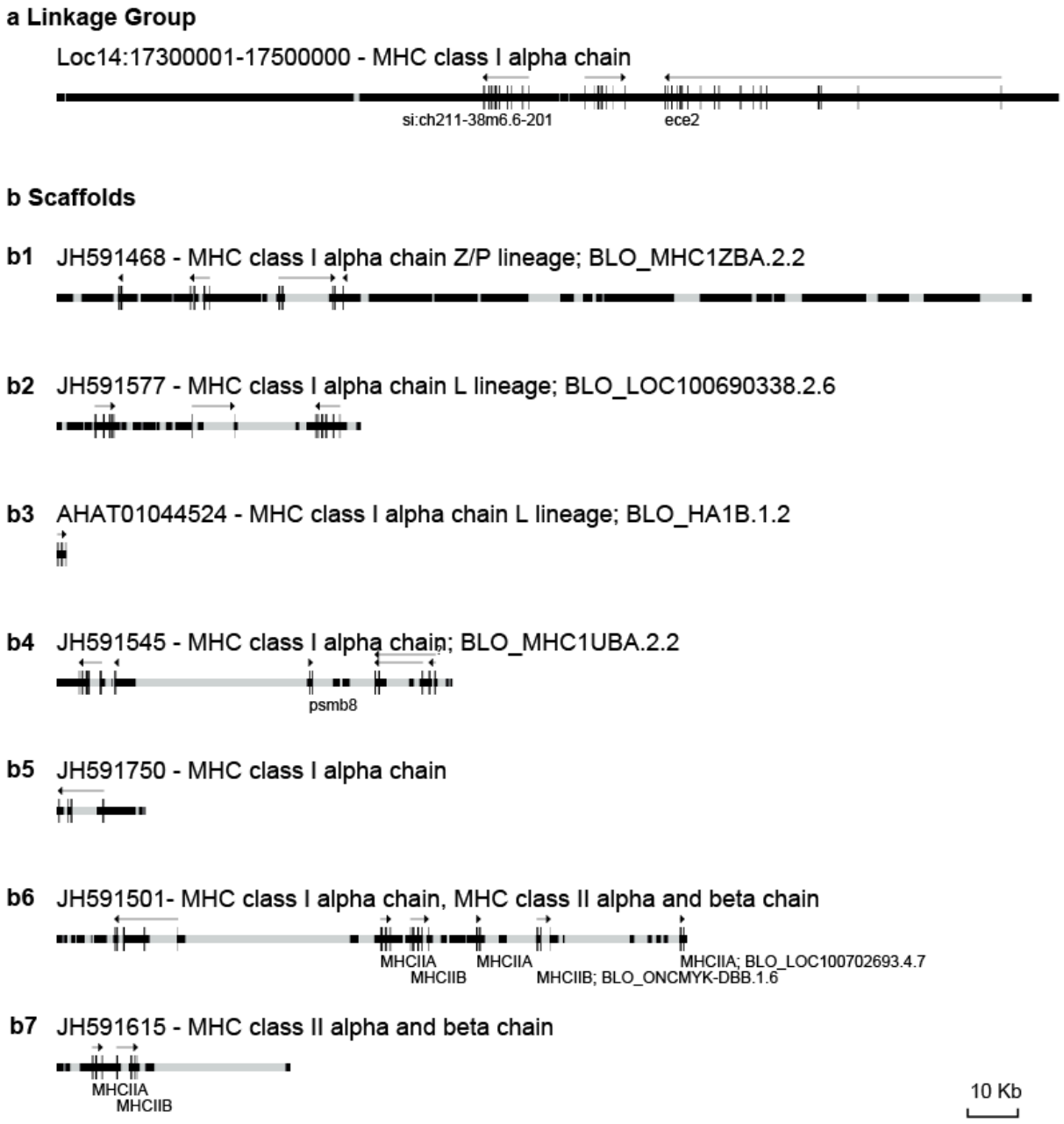
Supplementary Figures 18. Maximum likelihood phylogeny of vertebrate opsin proteins. Gar sequences are indicated in bold and groups are defined to the right.



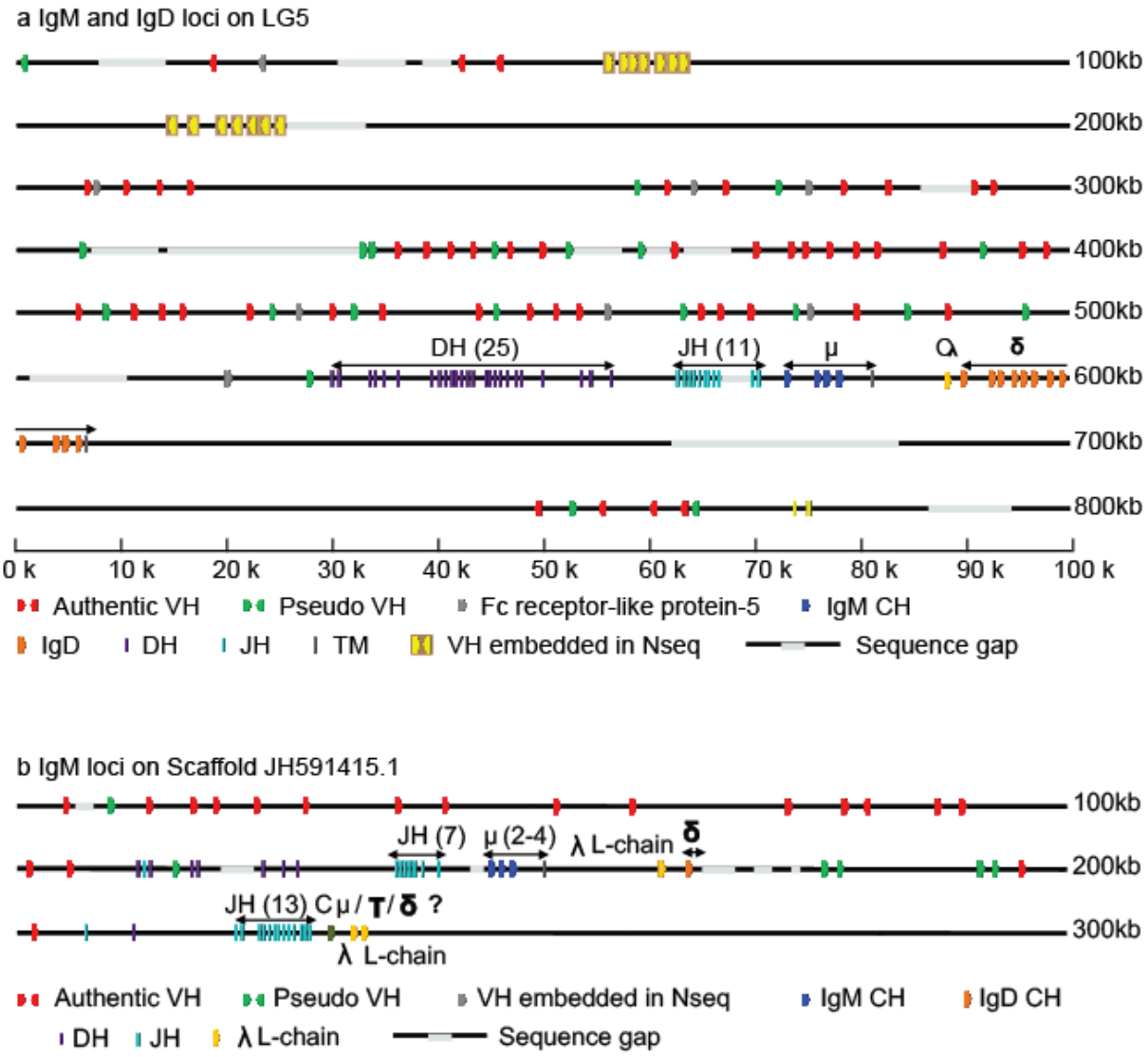
Supplementary Figure 19. Phylogenetic analysis of MHC class I alpha-3 domains. MHC class I alpha-3 domains have rather simple evolutionary relationships[210]. Amino acid sequences for previously characterized bony fish MHC class I genes were taken from GENBANK and Ensembl[119,121,122,124,211,212]. Sequences aligned by MUSCLE were first analyzed by NJ with p distance to provide a general picture of clustering within a species and to select representative amino acid sequences (data not shown). Selected sequences were then aligned by MUSCLE (a) or CLUSTALW (b) programs after removing vaguely aligned regions and analyzed by MEGA6[106]. Maximum likelihood trees were inferred based on the best model estimated. (a) WAG + G (parameter = 2.2027) + I (1.5197% sites) model was used for MUSCLE aligned sequences and (b) WAG + G (parameter = 3.4661) model was used for CLUSTALW aligned sequences. Bootstrap number >50% appears next to the branches. Branch lengths measured by number of substitutions per site. Leoc: Spotted gar, Dare: Zebrafish, Sasa: Atlantic salmon, Taru: Fugu, Gici: Nurse shark, Lame: Indonesian coelacanth, Xetr: African clawed frog, Hosa: Human.



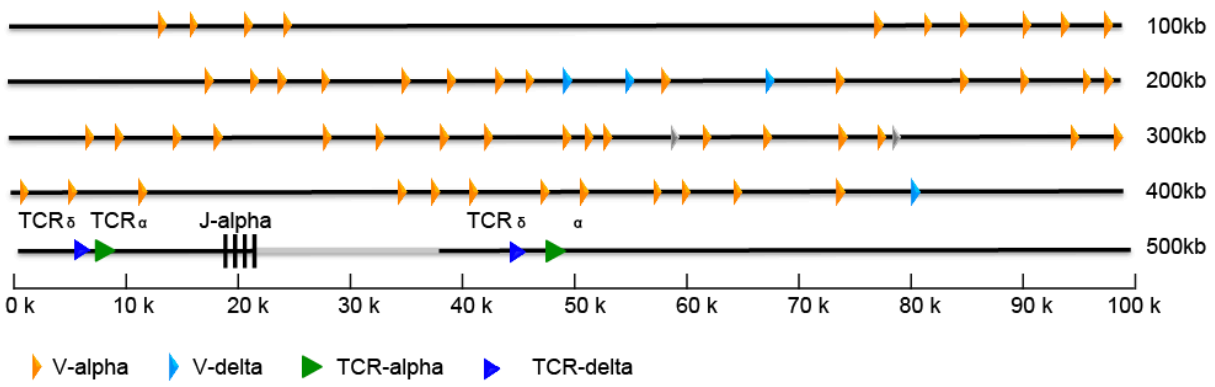
Supplementary Figure 20. Phylogenetic analysis of MHC class II alpha-2 and beta-2 domains. Amino acid sequences came from GENBANK and Ensembl along with previously characterized MHC class II genes[118,212]. Amino acid sequences were aligned by MUSCLE or CLUSTALW programs. Maximum likelihood trees were inferred with the best estimated model. a) MUSCLE aligned sequences using the WAG + G (parameter = 2.9553) model. b) CLUSTALW aligned sequences used the WAG + G (parameter = 2.9130) model. Where a species had a cluster of MHC genes, the cluster was compressed and sequences used do not appear in the figure. Leoc: Spotted gar, Dare: Zebrafish, Sasa: Atlantic salmon, Taru: Fugu, Gici: Nurse shark, Lame: Indonesian coelacanth, Xetr: African clawed frog, Hosa: Human.



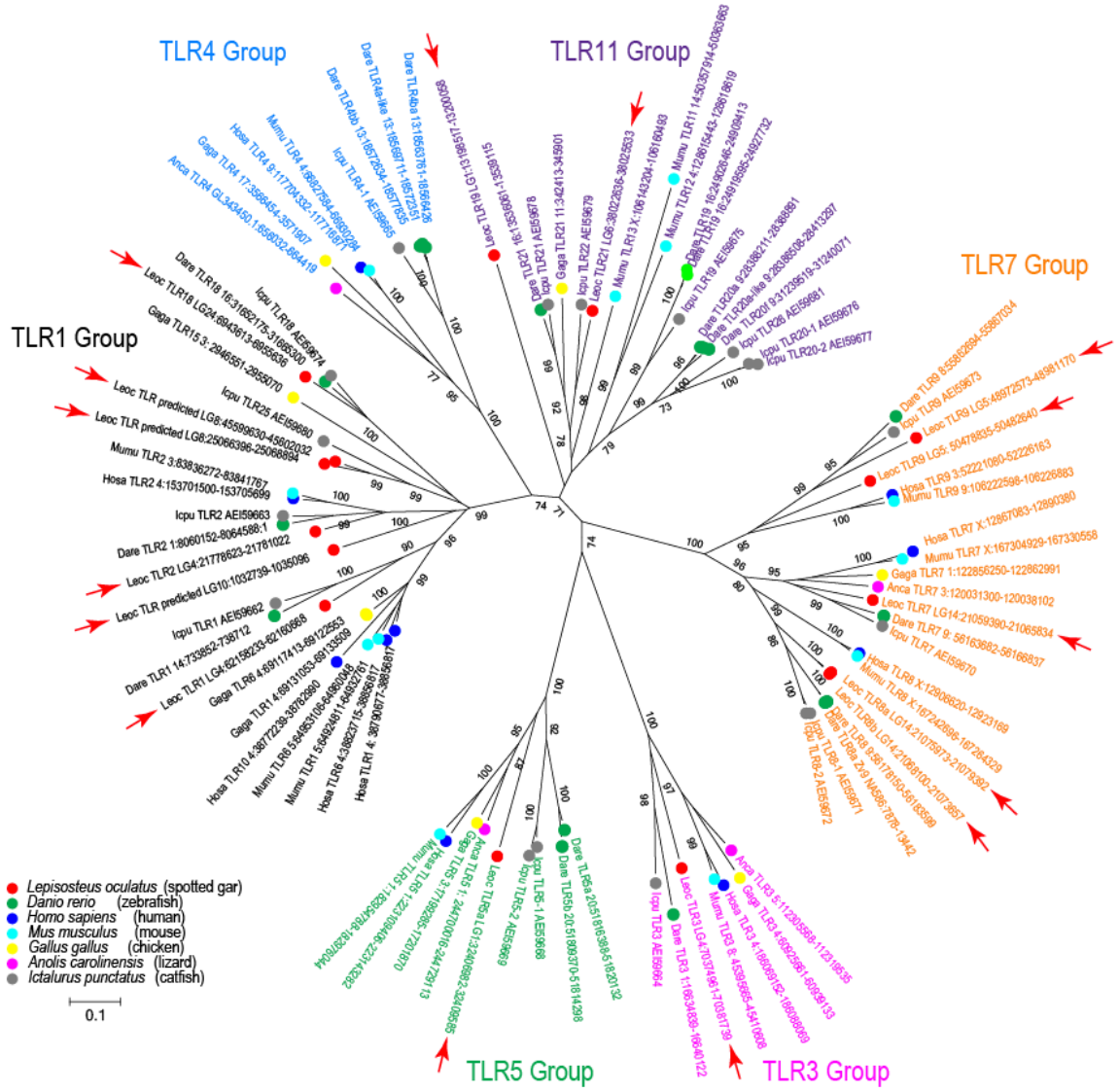
Supplementary Figure 21. Genomic arrangement of gar MHC genes. a) Only one MHC gene is assembled into a chromosome, Loc14. b) Several unassembled scaffolds contain MHC genes.



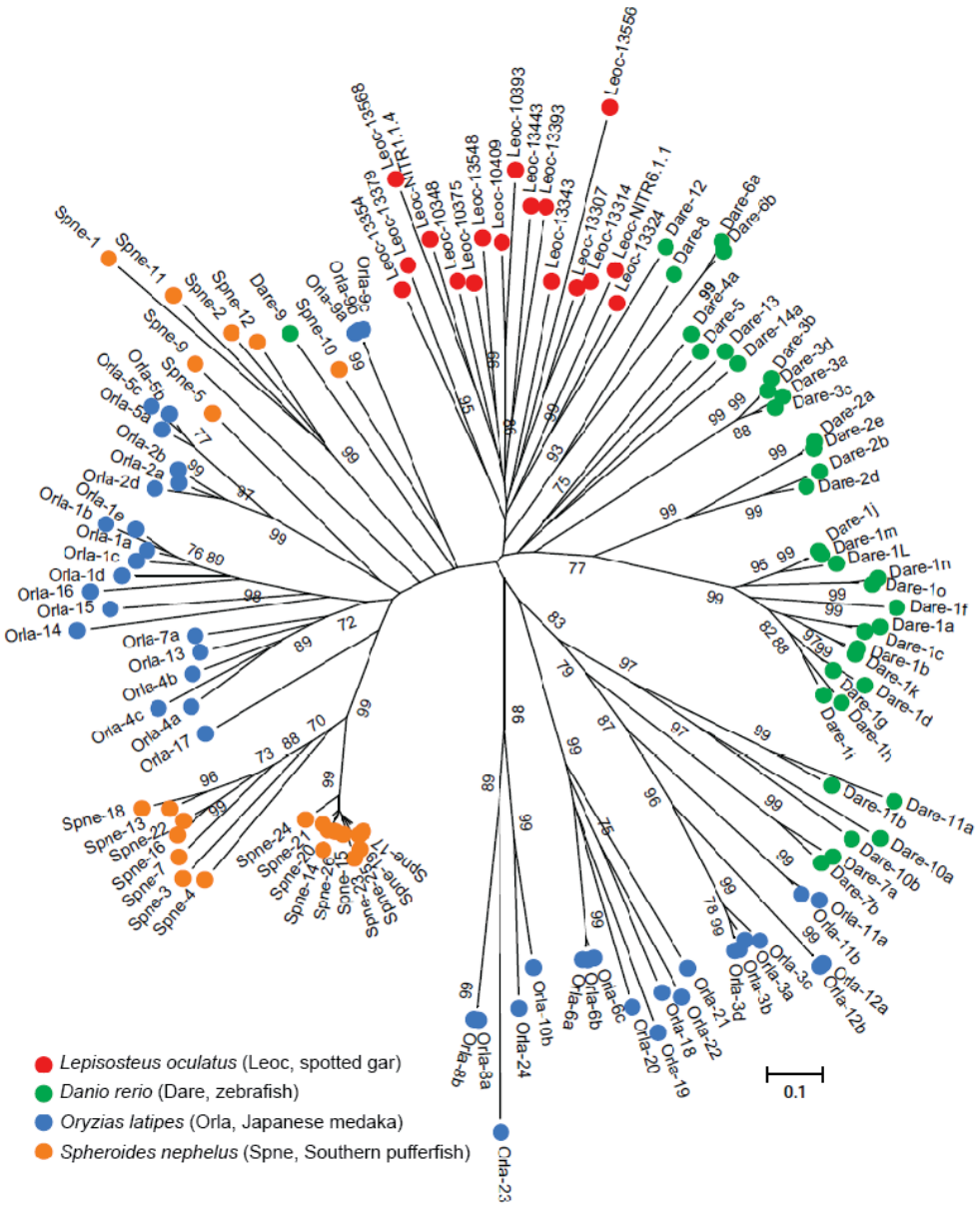
Supplementary Figure 22. IgH chain genome organization in spotted gar. a) BLAST searches using as query an IgM sequence (GenBank accession: U12455) from longnose gar (*Lepisosteus osseus*, a congener of spotted gar), identified linkage group Loc5 with several IgM-related sequences. We imported the first 800kb of the chromosome containing these sequences into VectorNTI (Invitrogen, USA) for further analysis. VH, JH and CH elements for both IgM and IgD were identified via motif searches and DH elements were identified by inference using the position flanking RSS sequences. Different domains are represented with different color codes as shown below each panel. This Ig-locus consists of 46 functional VH gene segments, 19 pseudo VH gene segments, and 8 VH gene segments that are partially embedded in sequence gaps. Although teleost homologs of both IgM and IgD are located on Loc5, neither bioinformatic searches no manual annotation of additional sequence within the locus identified a gene that encodes a third teleost isotype (IgZ/IgT). The figure shows the positions of modules within the locus in scale but icons indicating module positions are not in scale. b) IgH chain genome organization in spotted gar (Ig locus-2, Scaffold JH591415.1). This locus was manually annotated as described in a.



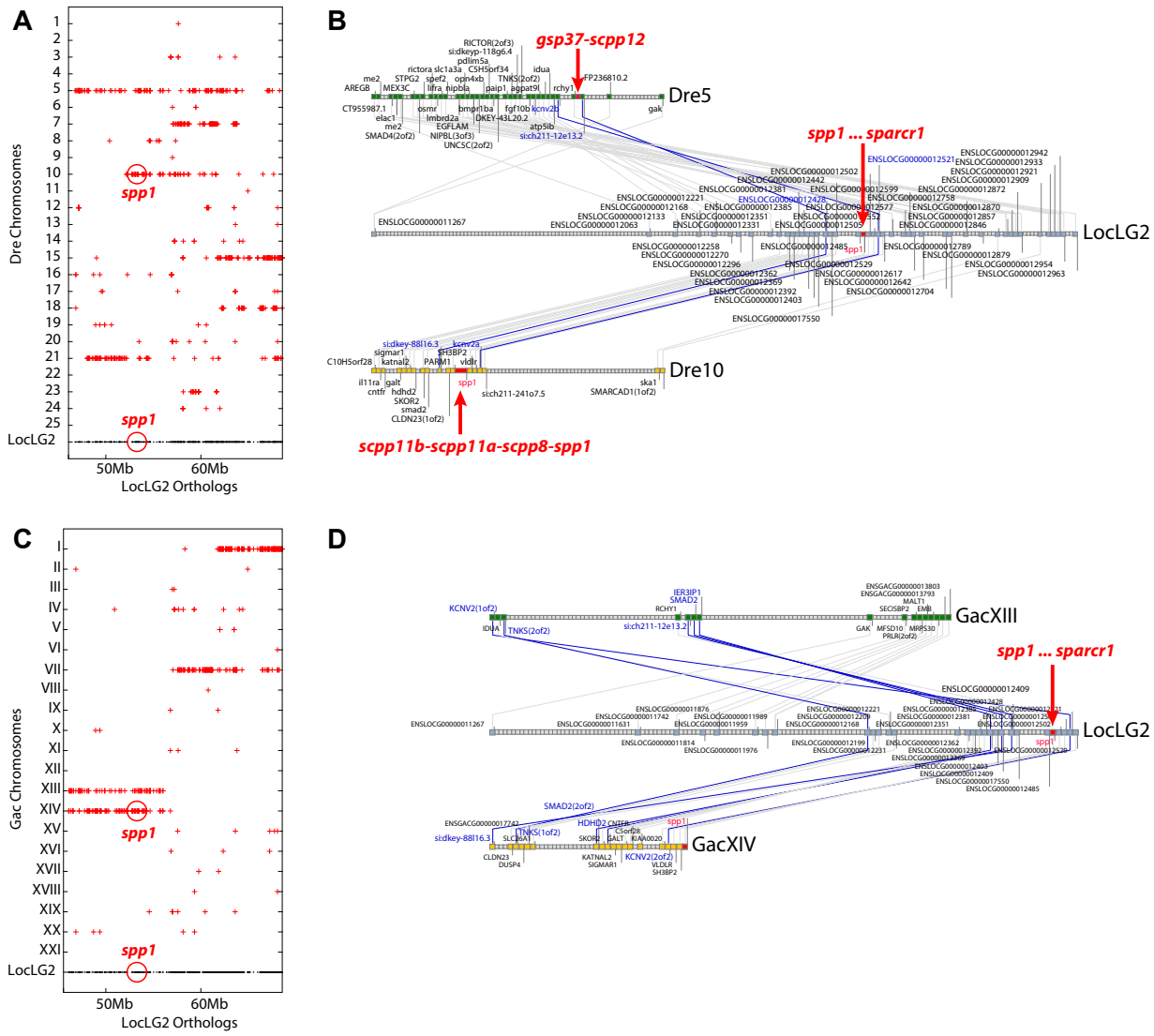
Supplementary Figure 23. Physical map and annotation of T-cell receptor α/δ locus. The TCR α/δ locus was annotated manually in VectorNTI sequence analysis software. Both TCRα and TCRδ are located on LG24 between 4200000 to 4628560 bp and are organized in an arrangement specific for spotted gar. Transcriptional orientation is shown by an arrowhead for each segment. VH genes for TCRα and TCRδ are shown by different colors. The position of modules within the locus is in scale but icons indicating module positions are not in scale.



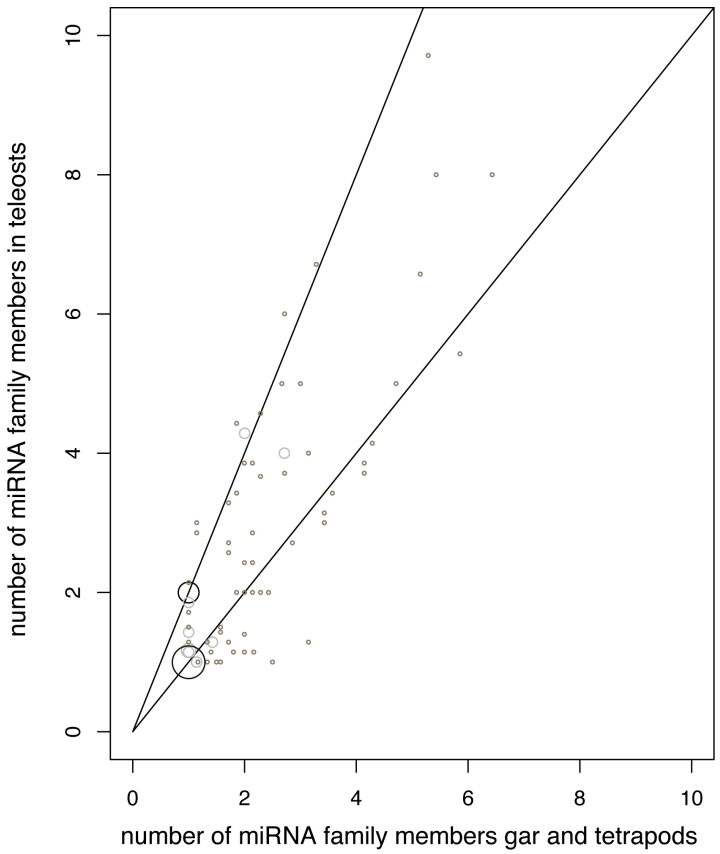
Supplementary Figure 24. Phylogenetic analysis of TLR-TIR domains. Evolutionary relationships among gar, zebrafish, human, mouse, chicken, catfish, and lizard TLR-TIR domains. TIR domains were predicted by SMART software[213]. Evolutionary history was inferred using the Neighbor-Joining method[214]. The image shows the optimal tree with the sum of branch length = 16.03559248. The number next to the branches indicates the percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (2,000 replicates)[215]. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. Evolutionary distances were computed using the Poisson correction method[216] in units of the number of amino acid substitutions per site. Analysis involved 86 amino acid sequences with a total of 121 positions in the final dataset. All positions containing gaps and missing data were eliminated. Evolutionary analyses were conducted in MEGA6[106]. Gar TLR4 (ENSLOCG00000003751) was not included in this analysis because its predicted sequence lacks a TIR domain. The TLR-related sequence, ENSLOCG00000017625, was omitted from this analysis because its predicted sequence lacks leucine rich repeats, a defining feature of TLRs.



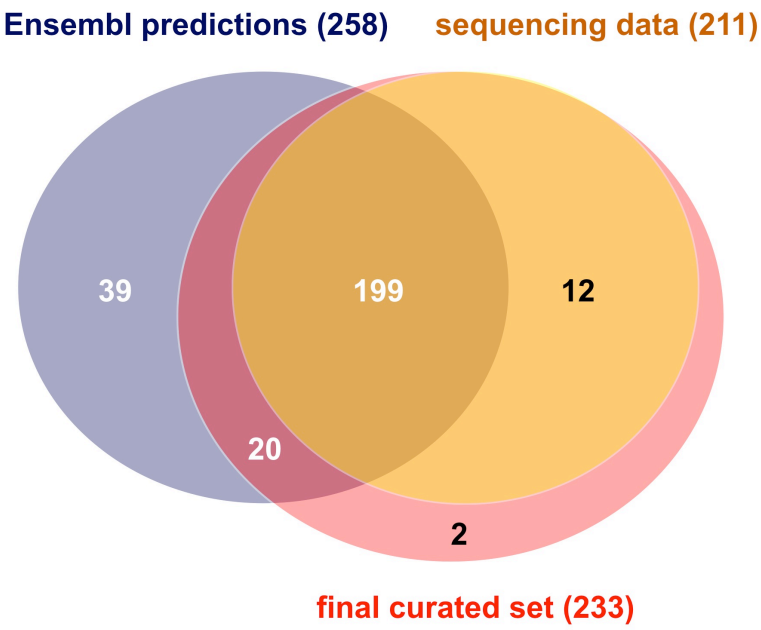
Supplementary Figure 25. Evolutionary relationships among gar and teleost novel immune-type receptors. Novel immune-type receptor (NITR) genes were initially identified on gar Loc14 with BLAST searches employing zebrafish NITRs as queries. Additional NITRs were identified on unplaced scaffold JH591499.1 by BLAST searches using gar NITRs from LG14 as queries. Two NITR sequences from transcriptomic analyses were identified that did not map to the reference genome ("BLO_NITR1.1.4" and "BLO_NITR6.1.1"). The evolutionary history of NITR V-type immunoglobulin domains[136] was inferred using the Neighbor-Joining method[214]. Bootstrap support was inferred from 2000 replicates and displayed on branches reproduced in greater than 70% of replicates[215]. The figure shows the optimal tree with the sum of branch length = 31.86. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. Evolutionary distances were computed using the Poisson correction method[216] and are in units of the number of amino acid substitutions per site. Analysis involved 122 NITR V domain amino acid sequences. All positions containing gaps and missing data were eliminated. A total of 89 positions contributed to the final dataset. Evolutionary analyses were conducted in MEGA6[106].



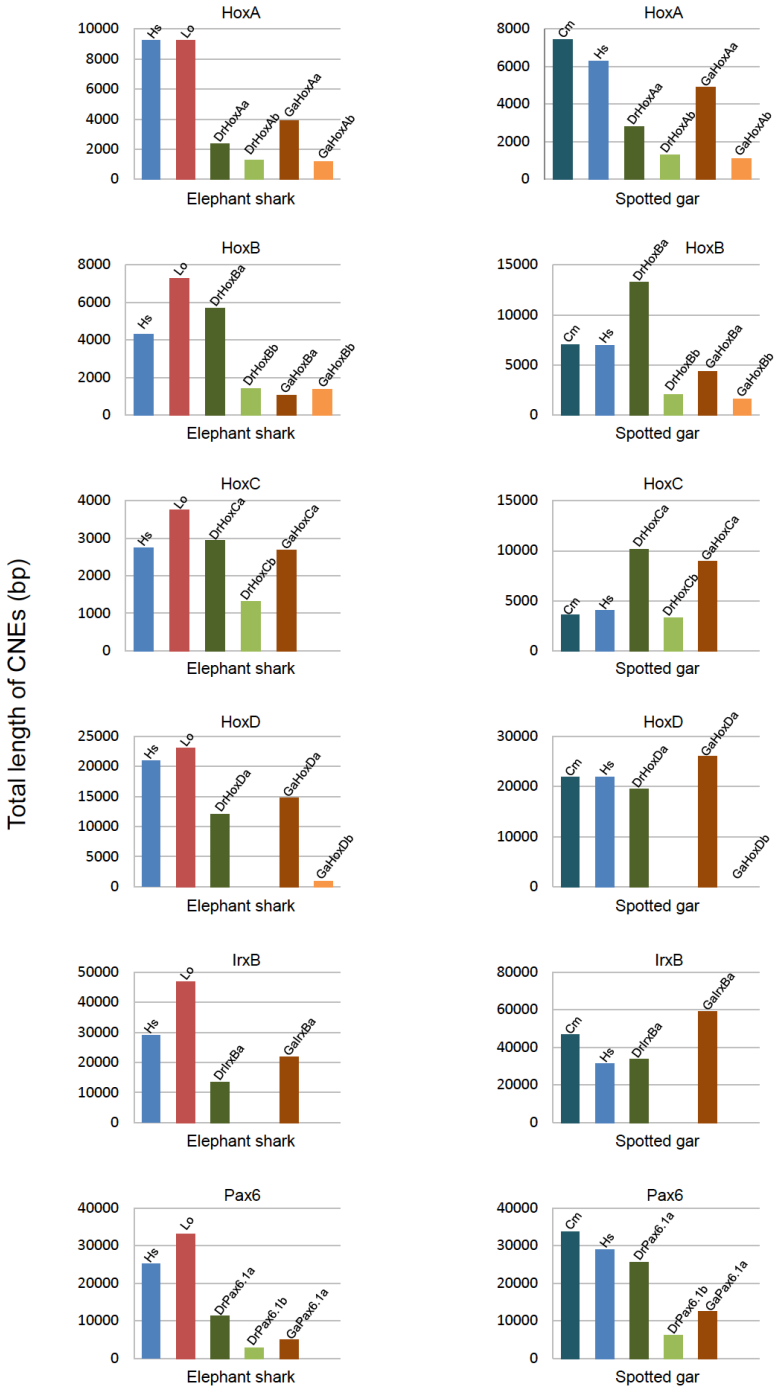
Supplementary Figure 26. Conserved synteny of the spotted gar *scpp* gene region on LG2 with teleosts. Dotplots (a, c) and composite clusters (b, d) generated with the Synteny Database[86] show double conserved synteny of gar LG2 to *scpp* regions on zebrafish chromosomes Dre5 and Dre10 (a, b) and with the *scpp* region on stickleback chromosomes GacXIV and GacXIII (c, d), providing evidence for the origin of these ohnologous genomic regions in the teleost genome duplication (TGD).



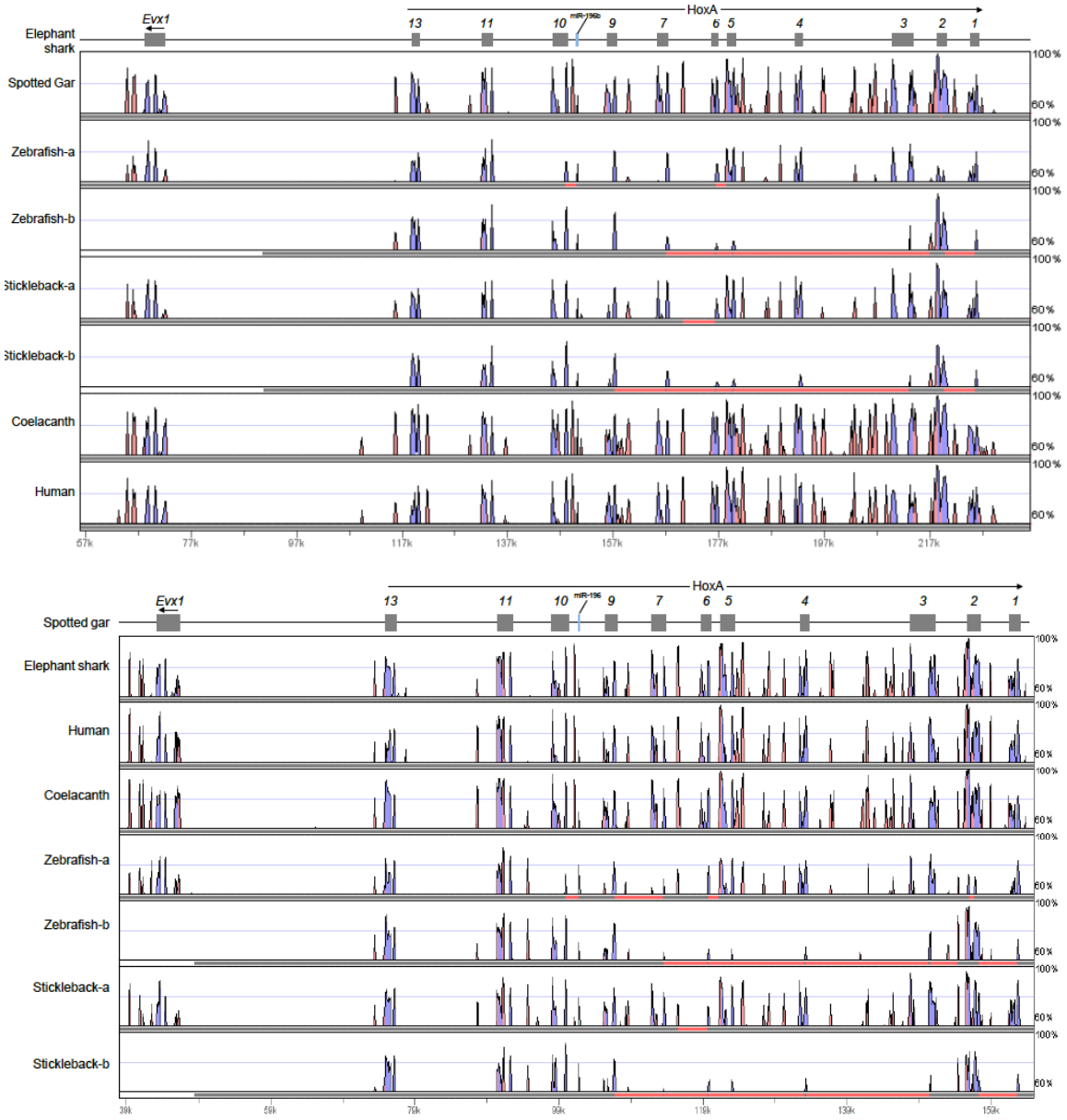
Supplementary Figure 27. miRNA genes in non-teleost vertebrates vs. teleosts. The number of genes in miRNA families in species that did not experience the TGD (including gar, horizontal axis) is plotted against the number of genes in corresponding families in teleosts (vertical axis). Data are based on the *in silico* miRNA analysis, Supplementary Note 11.1). For families on the diagonal, the numbers are about the same, and so gene loss after the TGD brought the total number of microRNA genes back to singletons. The steeper line indicates miRNA families that tended to retain both copies of most family genes after the TGD.



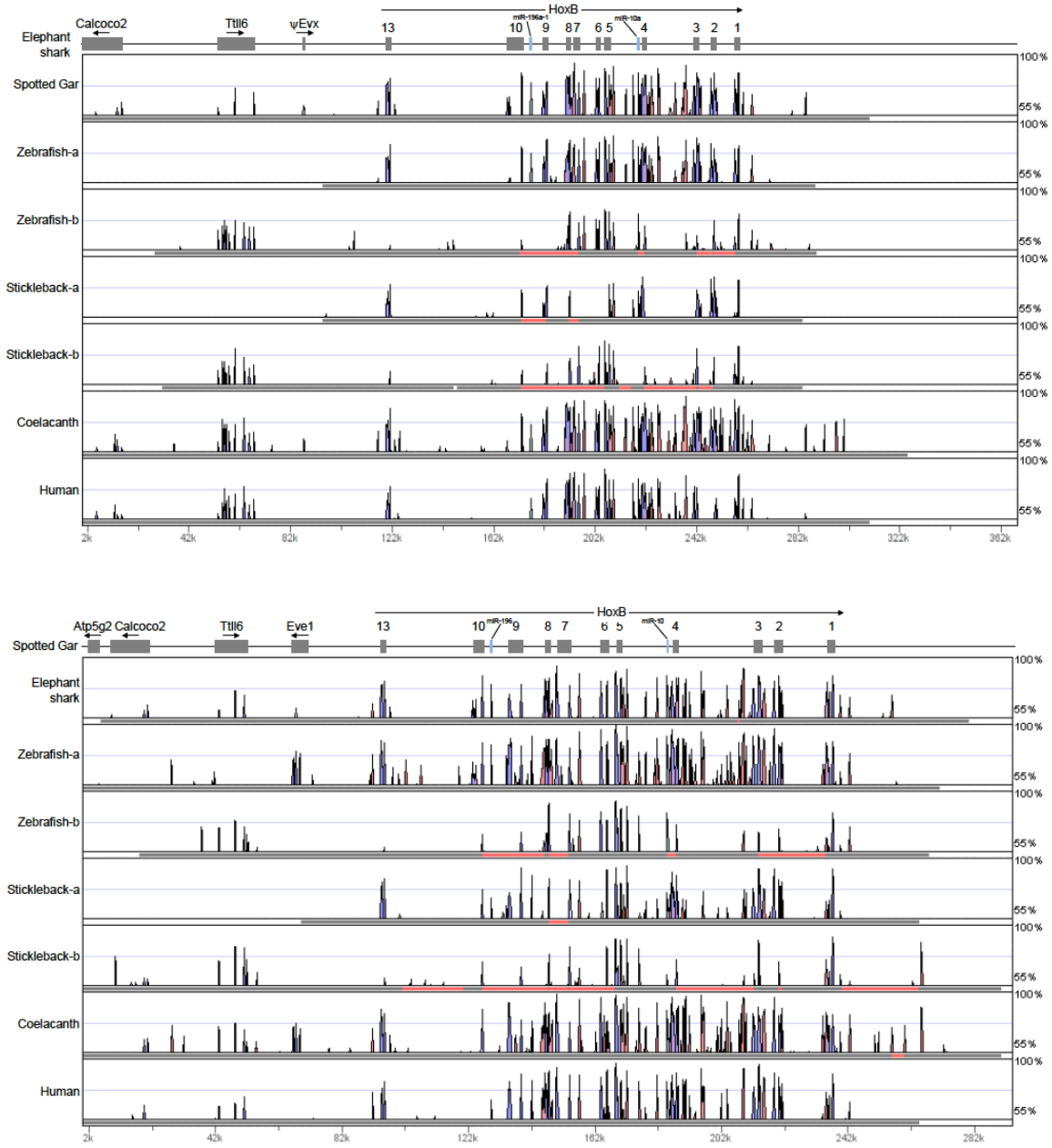
Supplementary Figure 28. miRNA gene annotation in gar based on RNA-seq and orthology searches. A final set of 233 miRNA genes (red disc) was annotated in gar using small RNA-seq data and orthology searches (Supplementary Note 11.2), including 211 genes identified from our sequencing data (yellow disc), among which 199 overlap with Ensembl annotations (blue disc). Additionally, 20 genes predicted by Ensembl and two more genes were annotated following orthology verification with other species. See Supplementary Table 11 for information on individual gar miRNA annotation information.



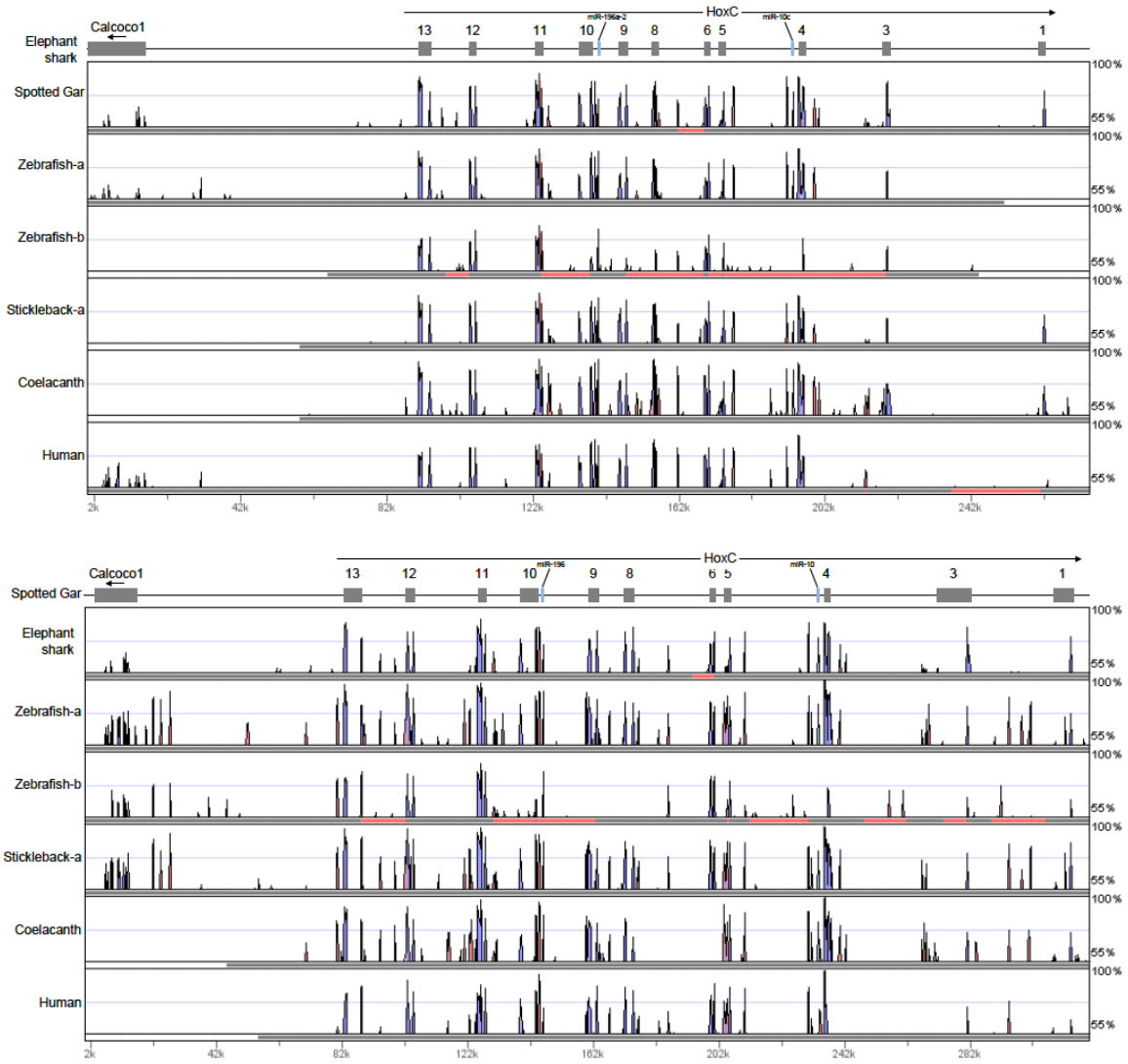
Supplementary Figure 29. Total CNE length of selected gnathostome developmental gene loci. The figure depicts data using elephant shark (left column) or spotted gar (right column) as the base sequence from Supplementary Tables 14-19. Sequences from human (Hs), spotted gar (Lo), zebrafish (Dr), stickleback (Ga) and elephant shark (Cm) were aligned using SLAGAN and CNEs were predicted using VISTA. The Y-axis lists total length of CNEs in base pairs and the gene locus is shown above the graph.



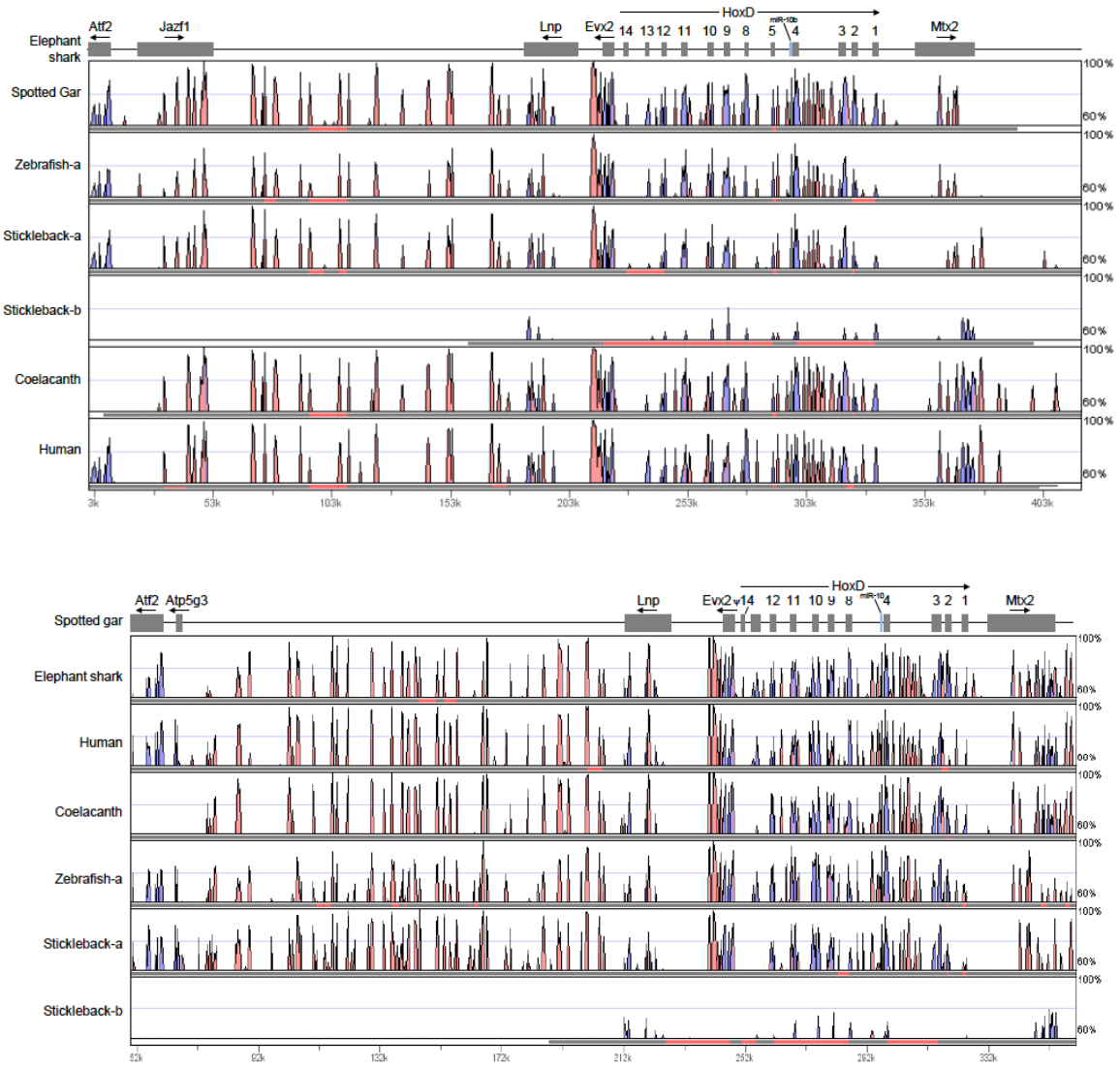
Supplementary Figure 30. CNEs in the *HoxA* locus of spotted gar, human, coelacanth, elephant shark and teleost fishes (zebrafish and stickleback) predicted using elephant shark (top) or spotted gar (bottom) as the base sequence. Sequences were aligned using SLAGAN and CNEs were predicted using VISTA. The CNE definition used was: >65% identity and ≥50 bp window size. Blue peaks represent conserved exons whereas the pink ones denote CNEs.



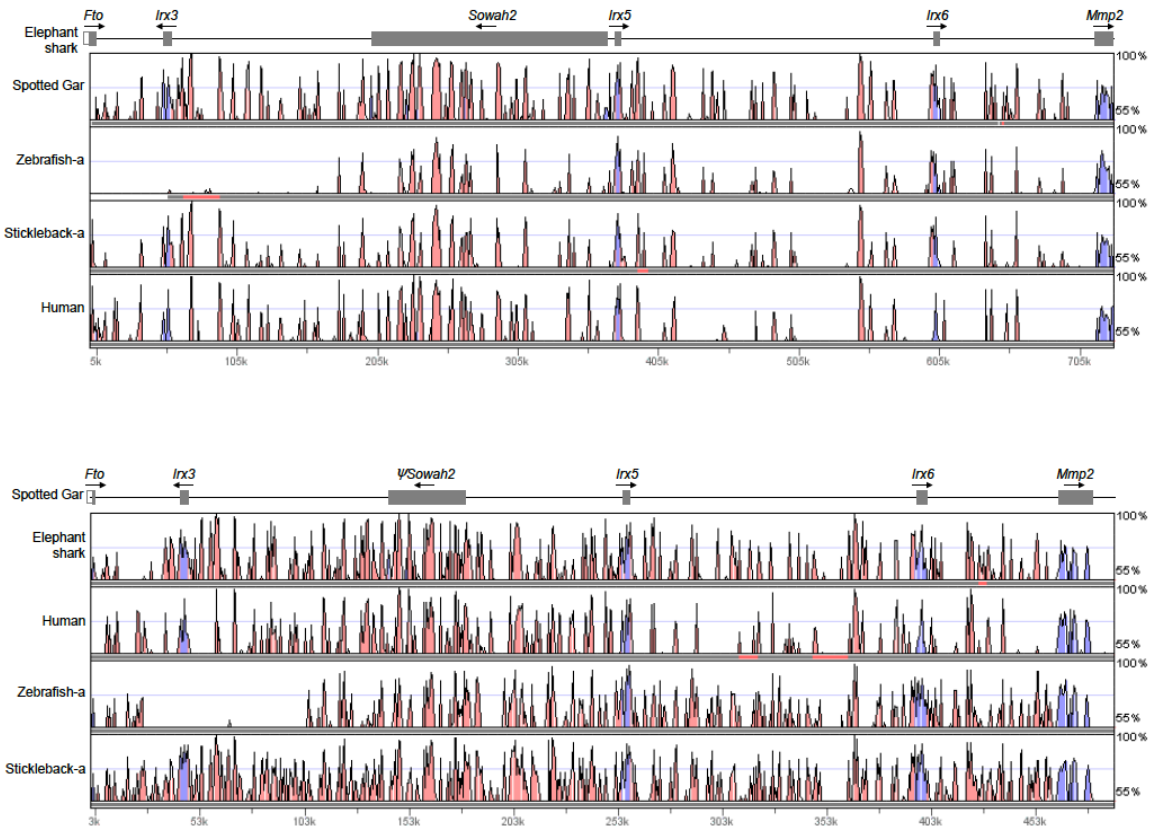
Supplementary Figure 31. CNEs in the *HoxB* locus of spotted gar, human, coelacanth, elephant shark and teleost fishes (zebrafish and stickleback) predicted using elephant shark (top) or spotted gar (bottom) as the base sequence. Sequences were aligned using SLAGAN and CNEs were predicted using VISTA. The CNE definition used was: >65% identity and ≥50 bp window size. Blue peaks represent conserved exons whereas the pink ones denote CNEs. There is a pseudo-*Evx* gene in the elephant shark locus and an intact *eve1* gene in the spotted gar locus.



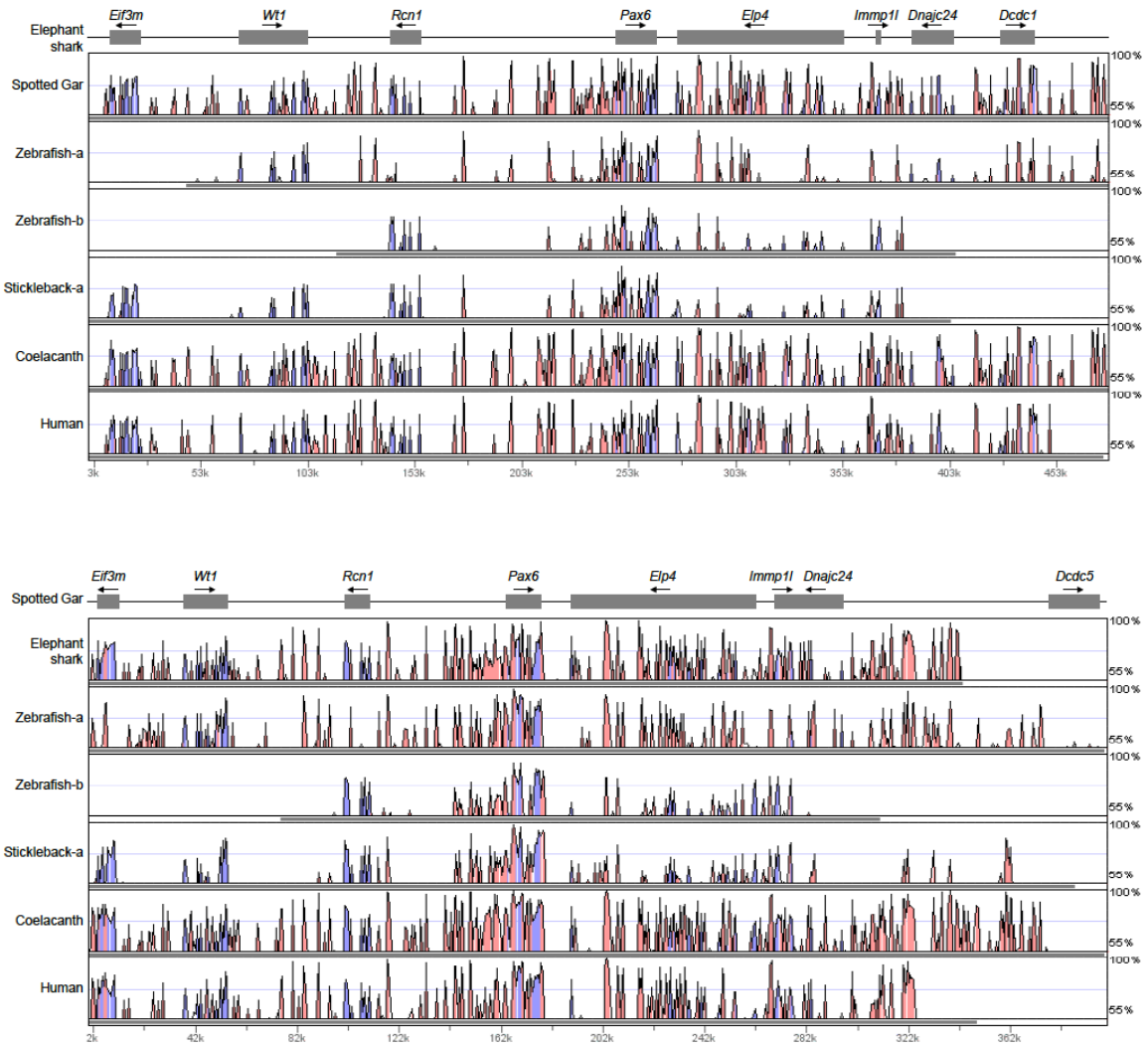
Supplementary Figure 32. CNEs in the *HoxC* locus of spotted gar, human, coelacanth, elephant shark and teleost fishes (zebrafish and stickleback) predicted using elephant shark (top) or spotted gar (bottom) as the base sequence. Sequences were aligned using SLAGAN and CNEs were predicted using VISTA. The CNE definition used was: >65% identity and ≥50 bp window size. Blue peaks represent conserved exons whereas the pink ones denote CNEs. Stickleback retains only the a-paralog of the *hoxC* locus.



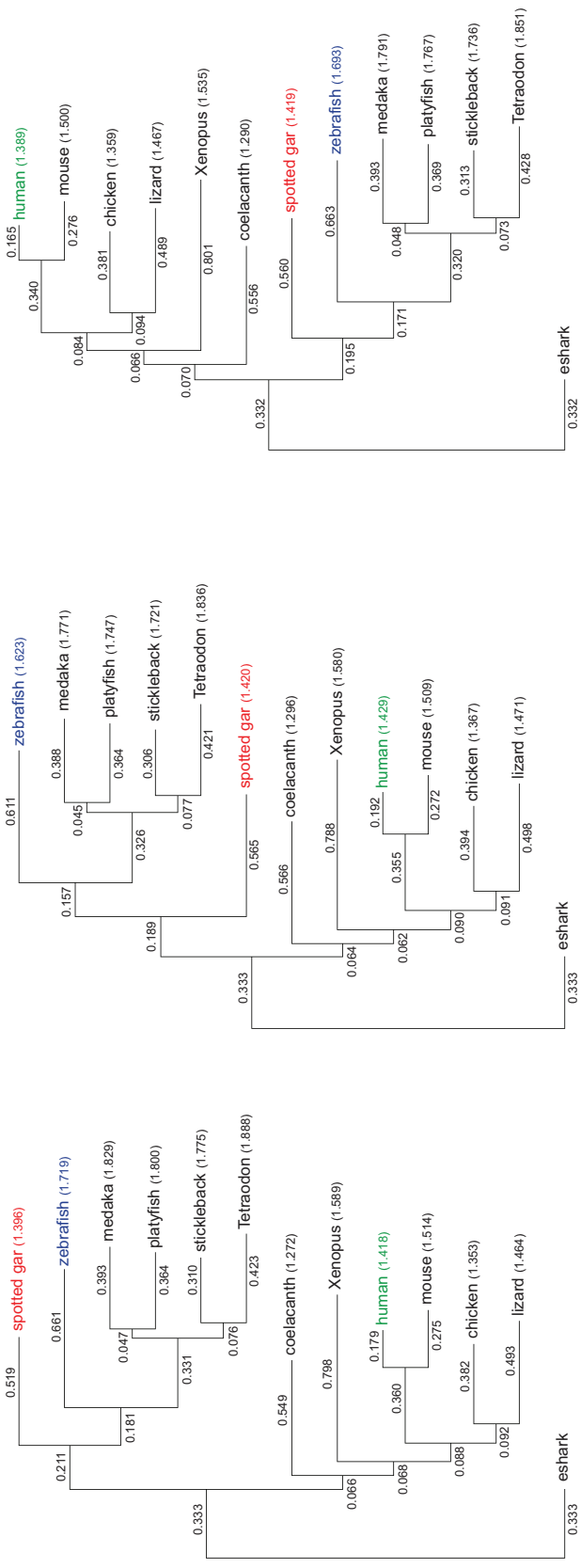
Supplementary Figure 33. CNEs in the *HoxD* locus of spotted gar, human, coelacanth, elephant shark and teleost fishes (zebrafish and stickleback) predicted using elephant shark (top) or spotted gar (bottom) as the base sequence. Sequences were aligned using SLAGAN and CNEs were predicted using VISTA. The CNE definition used was: >65% identity and ≥50 bp window size. Blue peaks represent conserved exons whereas the pink ones denote CNEs. Spotted gar possesses a *hoxD14* pseudogene in its *HoxD* locus.



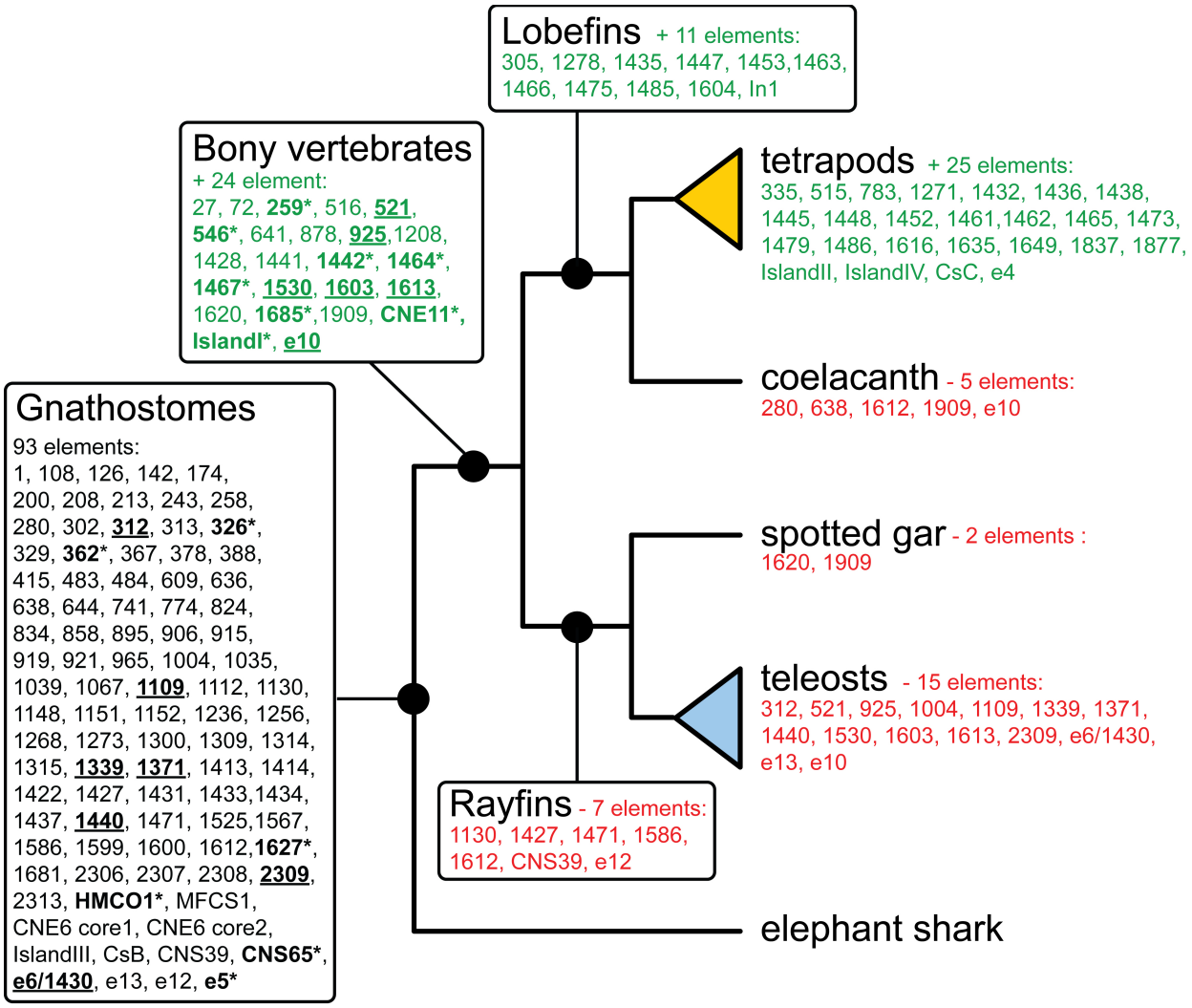
Supplementary Figure 34. CNEs in the *IrxB* locus of spotted gar, human, coelacanth, elephant shark and teleost fishes (zebrafish and stickleback) predicted using elephant shark (top) or spotted gar (bottom) as the base sequence. Sequences were aligned using SLAGAN and CNEs were predicted using VISTA. The CNE definition used was: >65% identity and ≥50 bp window size. Blue peaks represent conserved exons whereas the pink ones denote CNEs. The *irx3a* gene is not included in the CNE analysis as it is separated from *irx5a* and *irx6a* by ~15 intervening genes. The zebrafish *irxBb* locus is excluded from the analysis as it lacks *irx6b*; in addition, *irx3b* and *irx5b* are separated by more than 10 genes (~300 kb). Stickleback and medaka possess only *irxBa* locus as the *irxBb* locus has been lost secondarily.



Supplementary Figure 35. CNEs in the *Pax6* locus of spotted gar, human, coelacanth, elephant shark and teleost fishes (zebrafish and stickleback) predicted using elephant shark (top) or spotted gar (bottom) as the base sequence. Sequences were aligned using SLAGAN and CNEs were predicted using VISTA. The CNE definition used was: >65% identity and ≥50 bp window size. Blue peaks represent conserved exons whereas the pink ones denote CNEs. In elephant shark, the canonical form of *Pax6* is known as *Pax6.1*. Amongst teleost fishes, zebrafish retains duplicate copies of the *pax6* locus (*pax6.1a* and *pax6.1b*) whereas stickleback possesses a single *pax6* (*pax6.1a*) locus.



Supplementary Figure 36. phyloFit trees of 13-way whole genome alignments based on 4d sites. Left to right: gar-, zebrafish-, human-centric alignments. Numbers on branches indicate their lengths and numbers in parentheses for each taxon indicate total branch lengths to the root of the tree.



Supplementary Figure 37. Evolution of human limb enhancers. Limb enhancer complements for various phylogenetic nodes are given in cornered boxes. Phylogenetic gains (green) of limb enhancers since the gnathostome ancestor of human and elephant shark were determined from the human-centric genome alignment. Losses in other lineages (red) were inferred from the human-centric and the gar-centric genome alignment. Elements whose evolution was specifically clarified by the inclusion of gar are indicated in bold, i.e., presence of 14 enhancers in teleosts established by connectivity through gar (asterisk) as well as elements lost specifically in the teleost lineage (underlined). See Supplementary Tab. 22 for further information on individual elements.

D. Supplementary Files

Supplementary File 1. Phylogenomic alignment file in phylip format.
[separate .txt file]

E. Source Data Files

Source Data Set 1. Source Data for Figure 1. The concatenated protein alignment from the coelacanth genome analysis by Amemiya et al. (2013) was expanded by the following sequences from gar, bowfin, and western painted turtle.
[separate .xls file]

Source Data Set 2. Source Data for Figure 2. Genomic locations of genes depicted in Figs. 2b,c,e,f and chromosome sizes shown in Fig. 2d.
[separate .xls file]

Source Data Set 3. Source Data for Figure 3. Genomic locations of genes depicted in Figs. 3a-c.
[separate .xls file]

Source Data Set 4. Source Data for Figure 6. DESeq gene expression values for Fig. 6c-h (normalized read counts).
[separate .xls file]

F. SUPPLEMENTARY REFERENCES

1 Hoegg, S., Brinkmann, H., Taylor, J. S. & Meyer, A. Phylogenetic timing of the fish-specific genome duplication correlates with the diversification of teleost fish. *J Mol Evol* **59**, 190-203 (2004).

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

3 Braasch, I. *et al.* A new model army: Emerging fish models to study the genomics of vertebrate Evo-Devo. *J Exp Zool B Mol Dev Evol* **324**, 316-341 (2015).

4 Wright, J. J., David, S. R. & Near, T. J. Gene trees, species trees, and morphology converge on a similar phylogeny of living gars (Actinopterygii: Holostei: Lepisosteidae), an ancient clade of ray-finned fishes. *Mol Phylogenet Evol* **63**, 848-856 (2012).

5 Alfaro, R. M., Gonzalez, C. A. & Ferrara, A. M. Gar biology and culture: status and prospects. *Aquaculture Research* **39**, 748-763 (2008).

6 Page, L. M. & Burr, B. M. *Peterson field guide to the freshwater fishes of North American north of Mexico*. 2 edn, (Houghton Mifflin Harcourt, 2011).

7 Miller, R. R., Minckley, W. L. & Norris, S. M. *Freshwater fishes of Mexico*. (University of Chicago Press, 2005).

8 David, S. R., Kik, R. S., Diana, J. S., Rutherford, E. S. & Wiley, M. J. Evidence of Countergradient Variation in Growth of Spotted Gars from Core and Peripheral Populations. *Transactions of the American Fisheries Society* **144**, 837-850 (2015).

9 Grande, L. An Empirical Synthetic Pattern Study of Gars (Lepisosteiformes) and Closely Related Species, Based Mostly on Skeletal Anatomy. The Resurrection of Holostei. *Copeia*, 1-863 (2010).

10 Darwin, C. *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*. (John Murray, 1859).

11 Rabosky, D. L. *et al.* Rates of speciation and morphological evolution are correlated across the largest vertebrate radiation. *Nat Commun* **4**, 1958 (2013).

12 Sallan, L. C. Major issues in the origins of ray-finned fish (Actinopterygii) biodiversity. *Biol Rev Camb Philos Soc* **89**, 950–971(2014).

13 Metscher, D. & Ahlberg, P. E. in *Major Events in Early Vertebrate Evolution The Systematics Association Special Volume Series* (ed P. E. Ahlberg) Ch. 19, 333-349 (Taylor & Francis, 2001).

14 Gehrke, A. R. *et al.* Deep conservation of wrist and digit enhancers in fish. *Proceedings of the National Academy of Sciences of the United States of America* **112**, 803-808 (2015).

15 Graham, J. B. *Air-Breathing Fishes*. (Academic Press, 1997).

16 Rahn, H., Rahn, K. B., Howell, B. J., Gans, C. & Tenney, S. M. Air breathing of the garfish (Lepisosteus osseus). *Respir Physiol* **11**, 285-307 (1971).

17 Sasagawa, I., Ishiyama, M., Yokosuka, H. & Mikami, M. Fine structure and development of the collar enamel in gars, Lepisosteus oculatus, Actinopterygii. *Frontiers of Materials Science in China* **2**, 134-142 (2008).

18 Sire, J. Y. Light and TEM study of nonregenerated and experimentally regenerated scales of Lepisosteus oculatus (Holostei) with particular attention to ganoine formation. *Anat Rec* **240**, 189-207 (1994).

19 Bowmaker, J. K. in *How Animals See the World: Comparative Behavior, Biology, and Evolution of Vision* (eds O. F. Lazareva, T. Shimizu, & E. A. Wasserman) (Oxford University Press, 2012).

20 Suttkus, R. D. in *Fishes of the western North Atlantic: Soft-rayed fishes. Memoirs of the Sears Foundation for Marine Research I, Part 3.* (eds H. B. Bigelow & W. C. Schroeder) 61-88 (Sears Foundation for Marine Research, 1963).

21 Long, W. L. & Ballard, W. W. Normal embryonic stages of the longnose gar, Lepisosteus osseus. *BMC Dev Biol* **1**, 6 (2001).

22 Eames, B. F., Amores, A., Yan, Y. L. & Postlethwait, J. H. Evolution of the osteoblast: skeletogenesis in gar and zebrafish. *BMC evolutionary biology* **12**, 27 (2012).

23 Braasch, I. *et al.* Connectivity of vertebrate genomes: Paired-related homeobox (Prrx) genes in spotted gar, basal teleosts, and tetrapods. *Comparative biochemistry and physiology. Toxicology & pharmacology : CBP* **163**, 24-36 (2014).

24 Braasch, I. & Postlethwait, J. H. in *Polyploidy and Genome Evolution* (eds P. S. Soltis & D. E. Soltis) Ch. 17, 341-383 (Springer, 2012).

25 Williams, L. J. *et al.* Paired-end sequencing of Fosmid libraries by Illumina. *Genome research* **22**, 2241-2249 (2012).

26 Gnerre, S. *et al.* High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 1513-1518 (2011).

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

28 Levin, J. Z. *et al.* Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nat Methods* **7**, 709-715 (2010).

29 Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* **29**, 644-652 (2011).

30 Schulz, M. H., Zerbino, D. R., Vingron, M. & Birney, E. Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics* **28**, 1086-1092 (2012).

31 Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658-1659 (2006).

32 Pertea, G. *et al.* TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. *Bioinformatics* **19**, 651-652 (2003).

33 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760 (2009).

34 Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* **16**, 276-277 (2000).

35 Holt, C. & Yandell, M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* **12**, 491 (2011).

36 Jurka, J. *et al.* Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res* **110**, 462-467 (2005).

37 Korf, I., Yandell, M. & Bedell, J. *BLAST*. (O'Reilly Media, 2003).

38 Cantarel, B. L. *et al.* MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome research* **18**, 188-196 (2008).

39 Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **5**, 59 (2004).

40 Stanke, M. & Waack, S. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* **19 Suppl 2**, ii215-225 (2003).

41 Stanke, M., Diekhans, M., Baertsch, R. & Haussler, D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* **24**, 637-644 (2008).

42 Ter-Hovhannisyan, V., Lomsadze, A., Chernoff, Y. O. & Borodovsky, M. Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training. *Genome research* **18**, 1979-1990 (2008).

43 Morgulis, A., Gertz, E. M., Schaffer, A. A. & Agarwala, R. A fast and symmetric DUST implementation to mask low-complexity DNA sequences. *J Comput Biol* **13**, 1028-1040 (2006).

44 Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **27**, 573-580 (1999).

45 Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. *Genome research* **14**, 988-995 (2004).

46 Collins, J. E., White, S., Searle, S. M. & Stemple, D. L. Incorporating RNA-seq data into the zebrafish Ensembl genebuild. *Genome research* **22**, 2067-2078 (2012).

47 Slater, G. S. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**, 31 (2005).

48 Burge, S. W. *et al.* Rfam 11.0: 10 years of RNA families. *Nucleic Acids Res* **41**, D226-232 (2013).

49 Kozomara, A. & Griffiths-Jones, S. miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res* **39**, D152-157 (2011).

50 Wicker, T. *et al.* A unified classification system for eukaryotic transposable elements. *Nat Rev Genet* **8**, 973-982 (2007).

51 Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921 (2001).
52 Howe, K. *et al.* The zebrafish reference genome sequence and its relationship to the human genome. *Nature* **496**, 498-503 (2013).
53 Kasahara, M. *et al.* The medaka draft genome and insights into vertebrate genome evolution. *Nature* **447**, 714-719 (2007).
54 Chalopin, D., Naville, M., Plard, F., Galiana, D. & Volff, J. N. Comparative analysis of transposable elements highlights mobilome diversity and evolution in vertebrates. *Genome Biol Evol* **7**, 567-580 (2015).
55 Bohne, A. *et al.* Zisupton--a novel superfamily of DNA transposable elements recently active in fish. *Molecular biology and evolution* **29**, 631-645 (2012).
56 Chalopin, D. *et al.* Evolutionary active transposable elements in the genome of the coelacanth. *J Exp Zool B Mol Dev Evol* **322**, 322-333 (2014).
57 Amemiya, C. T. *et al.* The African coelacanth genome provides insights into tetrapod evolution. *Nature* **496**, 311-316 (2013).
58 Shaffer, H. B. *et al.* The western painted turtle genome, a model for the evolution of extreme physiological adaptations in a slowly evolving lineage. *Genome biology* **14**, R28 (2013).
59 Ebersberger, I., Strauss, S. & von Haeseler, A. HaMStR: profile hidden markov model based search for orthologs in ESTs. *BMC evolutionary biology* **9**, 157 (2009).
60 Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution* **30**, 772-780 (2013).
61 Kuck, P. & Meusemann, K. FASconCAT: Convenient handling of data matrices. *Mol Phylogenet Evol* **56**, 1115-1118 (2010).
62 Castresana, J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular biology and evolution* **17**, 540-552 (2000).
63 Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312-1313 (2014).
64 Lartillot, N., Rodrigue, N., Stubbs, D. & Richer, J. PhyloBayes MPI: phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. *Syst Biol* **62**, 611-615 (2013).
65 Lartillot, N. & Philippe, H. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Molecular biology and evolution* **21**, 1095-1109 (2004).
66 Salichos, L. & Rokas, A. Inferring ancient divergences requires genes with strong phylogenetic signals. *Nature* **497**, 327-331 (2013).
67 Salichos, L., Stamatakis, A. & Rokas, A. Novel information theory-based measures for quantifying incongruence among phylogenetic trees. *Molecular biology and evolution* **31**, 1261-1271 (2014).
68 Near, T. J. *et al.* Resolution of ray-finned fish phylogeny and timing of diversification. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 13698-13703 (2012).
69 Betancur, R. R. *et al.* The tree of life and a new classification of bony fishes. *PLoS Curr* **5**, doi:10.1371/currents.tol.53ba26640df0ccaee75bb165c8c26288 (2013).
70 Broughton, R. E., Betancur, R. R., Li, C., Arratia, G. & Orti, G. Multi-locus phylogenetic analysis reveals the pattern and tempo of bony fish evolution. *PLoS Curr* **5**, doi:10.1371/currents.tol.2ca8041495ffafd0c92756e75247483e (2013).
71 Faircloth, B. C., Sorenson, L., Santini, F. & Alfaro, M. E. A Phylogenomic Perspective on the Radiation of Ray-Finned Fishes Based upon Targeted Sequencing of Ultraconserved Elements (UCEs). *PLoS One* **8**, e65923 (2013).
72 Morgan, C. C. *et al.* Heterogeneous models place the root of the placental mammal phylogeny. *Molecular biology and evolution* **30**, 2145-2156 (2013).
73 Romiguier, J., Ranwez, V., Delsuc, F., Galtier, N. & Douzery, E. J. Less is more in mammalian phylogenomics: AT-rich genes minimize tree conflicts and unravel the root of placental mammals. *Molecular biology and evolution* **30**, 2134-2144 (2013).
74 Tajima, F. Simple methods for testing the molecular evolutionary clock hypothesis. *Genetics* **135**, 599-607 (1993).
75 Takezaki, N., Rzhetsky, A. & Nei, M. Phylogenetic test of the molecular clock and linearized trees. *Molecular biology and evolution* **12**, 823-833 (1995).
76 Paradis, E., Claude, J. & Strimmer, K. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* **20**, 289-290 (2004).

77 Amores, A. & Postlethwait, J. H. Banded chromosomes and the zebrafish karyotype. *Methods in cell biology* **60**, 323-338 (1999).
78 Rab, P., Rabova, M., Reed, K. M. & Phillips, R. B. Chromosomal characteristics of ribosomal DNA in the primitive semionotiform fish, longnose gar Lepisosteus osseus. *Chromosome Res* **7**, 475-480 (1999).
79 Arias-Rodrigez L., Paramo-Delgadillo S., W.M., C.-S. & C.A., A.-G. Karyotype of the tropical gar Atractosteus tropicus (Lepisosteiformes: Lepisosteidae) and chromosomal variation in their larval and adults. . *Rev. Biol. Trop.* **57**, 529-539 (2009).
80 Ohno, S. *et al.* Microchromosomes in holocephalian, chondrostean and holostean fishes. *Chromosoma* **26**, 35-40 (1969).
81 Bogart, J. P., Balon, E. K. & Bruton, M. N. The chromosomes of the living coelacanth and their remarkable similarity to those of one of the most ancient frogs. *J Hered* **85**, 322-325 (1994).
82 Wickbom, T. The chromosomes of Ascaphus true and the evolution of the anuran karyotypes. *Hereditas* **36**, 406-418 (1950).
83 Pokorna, M. *et al.* Differentiation of sex chromosomes and karyotypic evolution in the eye-lid geckos (Squamata: Gekkota: Eublepharidae), a group with different modes of sex determination. *Chromosome Res* **18**, 809-820 (2010).
84 Krzywinski, M. *et al.* Circos: an information aesthetic for comparative genomics. *Genome research* **19**, 1639-1645 (2009).
85 Kinsella, R. J. *et al.* Ensembl BioMarts: a hub for data retrieval across taxonomic space. *Database (Oxford)* **2011**, bar030 (2011).
86 Catchen, J. M., Conery, J. S. & Postlethwait, J. H. Automated identification of conserved synteny after whole-genome duplication. *Genome research* **19**, 1497-1505 (2009).
87 Smith, J. J. *et al.* Sequencing of the sea lamprey (Petromyzon marinus) genome provides insights into vertebrate evolution. *Nat Genet* **45**, 415-421 (2013).
88 Putta, S., Smith, J. J., Staben, C. & Voss, S. R. MapToGenome: a comparative genomic tool that aligns transcript maps to sequenced genomes. *Evol Bioinform Online* **3**, 15-25 (2007).
89 Voss, S. R. *et al.* Origin of amphibian and avian chromosomes by fission, fusion, and retention of ancestral chromosomes. *Genome research* **21**, 1306-1312 (2011).
90 Uno, Y. *et al.* Inference of the protokaryotypes of amniotes and tetrapods and the evolutionary processes of microchromosomes from comparative gene mapping. *PLoS One* **7**, e53027 (2012).
91 Martin, K. J. & Holland, P. W. Enigmatic orthology relationships between Hox clusters of the african butterfly fish and other teleosts following ancient whole-genome duplication. *Molecular biology and evolution* **31**, 2592-2611 (2014).
92 Crow, K. D., Smith, C. D., Cheng, J. F., Wagner, G. P. & Amemiya, C. T. An independent genome duplication inferred from Hox paralogs in the American paddlefish--a representative basal ray-finned fish and important comparative reference. *Genome Biol Evol* **4**, 937-953 (2012).
93 Mulley, J. F. & Holland, P. W. Genomic organisation of the seven ParaHox genes of coelacanths. *J Exp Zool B Mol Dev Evol* **322**, 352-358 (2014).
94 Mulley, J. F., Chiu, C. H. & Holland, P. W. Breakup of a homeobox cluster after genome duplication in teleosts. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 10369-10372 (2006).
95 Mulley, J. F. & Holland, P. W. Parallel retention of Pdx2 genes in cartilaginous fish and coelacanths. *Molecular biology and evolution* **27**, 2386-2391 (2010).
96 Prohaska, S. J. & Stadler, P. F. Evolution of the vertebrate ParaHox clusters. *J Exp Zool B Mol Dev Evol* **306**, 481-487 (2006).
97 Siegel, N., Hoegg, S., Salzburger, W., Braasch, I. & Meyer, A. Comparative genomics of ParaHox clusters of teleost fishes: gene cluster breakup and the retention of gene sets following whole genome duplications. *BMC genomics* **8**, 312 (2007).
98 Guindon, S. *et al.* New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* **59**, 307-321 (2010).
99 Larsson, A. AliView: a fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics* **30**, 3276-3278 (2014).
100 Canestro, C., Catchen, J. M., Rodriguez-Mari, A., Yokoi, H. & Postlethwait, J. H. Consequences of lineage-specific gene loss on functional evolution of surviving paralogs: ALDH1A and retinoic acid signaling in vertebrate genomes. *PLoS Genet* **5**, e1000496 (2009).

101 Canestro, C., Postlethwait, J. H., Gonzalez-Duarte, R. & Albalat, R. Is retinoic acid genetic machinery a chordate innovation? *Evol Dev* **8**, 394-406 (2006).
102 Panda, S., Hogenesch, J. B. & Kay, S. A. Circadian rhythms from flies to human. *Nature* **417**, 329-335 (2002).
103 Dunlap, J. C. Molecular bases for circadian clocks. *Cell* **96**, 271-290 (1999).
104 Wang, H. Comparative analysis of period genes in teleost fish genomes. *J Mol Evol* **67**, 29-40 (2008).
105 Liu, C. *et al.* Molecular evolution and functional divergence of zebrafish (Danio rerio) cryptochrome genes. *Sci Rep* **5**, 8113 (2015).
106 Tamura, K., Stecher, G., Peterson, D., Filipski, A. & Kumar, S. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Molecular biology and evolution* **30**, 2725-2729 (2013).
107 Jones, D. T., Taylor, W. R. & Thornton, J. M. The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci* **8**, 275-282 (1992).
108 Rennison, D. J., Owens, G. L. & Taylor, J. S. Opsin gene duplication and divergence in ray-finned fish. *Mol Phylogenet Evol* **62**, 986-1008 (2012).
109 Lagman, D. *et al.* 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 evolutionary biology* **13**, 238 (2013).
110 Venkatesh, B., Ning, Y. & Brenner, S. Late changes in spliceosomal introns define clades in vertebrate evolution. *Proceedings of the National Academy of Sciences of the United States of America* **96**, 10267-10271 (1999).
111 Bellingham, J., Tarttelin, E. E., Foster, R. G. & Wells, D. J. Structure and evolution of the teleost extraretinal rod-like opsin (errlo) and ocular rod opsin (rho) genes: is teleost rho a retrogene? *J Exp Zool B Mol Dev Evol* **297**, 1-10 (2003).
112 Minamoto, T. & Shimizu, I. Molecular cloning of cone opsin genes and their expression in the retina of a smelt, Ayu (Plecoglossus altivelis, Teleostei). *Comp Biochem Physiol B Biochem Mol Biol* **140**, 197-205 (2005).
113 Yokoyama, S. Evolution of dim-light and color vision pigments. *Annu Rev Genomics Hum Genet* **9**, 259-282 (2008).
114 Hunt, D. M. *et al.* Spectral tuning of shortwave-sensitive visual pigments in vertebrates. *Photochem Photobiol* **83**, 303-310 (2007).
115 Peirson, S. N., Halford, S. & Foster, R. G. The evolution of irradiance detection: melanopsin and the non-visual opsins. *Philos Trans R Soc Lond B Biol Sci* **364**, 2849-2865 (2009).
116 Mano, H., Kojima, D. & Fukada, Y. Exo-rhodopsin: a novel rhodopsin expressed in the zebrafish pineal gland. *Brain Res Mol Brain Res* **73**, 110-118 (1999).
117 Bingulac-Popovic, J. *et al.* Mapping of Mhc class I and class II regions to different linkage groups in the zebrafish, *Danio rerio*. *Immunogenetics* **46**, 129-134 (1997).
118 Dijkstra, J. M., Grimholt, U., Leong, J., Koop, B. F. & Hashimoto, K. Comprehensive analysis of MHC class II genes in teleost fish genomes reveals dispensability of the peptide-loading DM system in a large part of vertebrates. *BMC evolutionary biology* **13**, 260 (2013).
119 Grimholt, U. *et al.* A comprehensive analysis of teleost MHC class I sequences. *BMC evolutionary biology* **15**, 32 (2015).
120 Basler, M., Kirk, C. J. & Groettrup, M. The immunoproteasome in antigen processing and other immunological functions. *Curr Opin Immunol* **25**, 74-80 (2013).
121 McConnell, S. C., Restaino, A. C. & de Jong, J. L. Multiple divergent haplotypes express completely distinct sets of class I MHC genes in zebrafish. *Immunogenetics* **66**, 199-213 (2014).
122 Dirscherl, H., McConnell, S. C., Yoder, J. A. & de Jong, J. L. The MHC class I genes of zebrafish. *Developmental and comparative immunology* **46**, 11-23 (2014).
123 Flajnik, M. F. & Kasahara, M. Origin and evolution of the adaptive immune system: genetic events and selective pressures. *Nat Rev Genet* **11**, 47-59 (2010).
124 Dirscherl, H. & Yoder, J. A. Characterization of the Z lineage Major histocompatability complex class I genes in zebrafish. *Immunogenetics* **66**, 185-198 (2014).
125 Milner, C. M. & Campbell, R. D. Genetic organization of the human MHC class III region. *Front Biosci* **6**, D914-926 (2001).

126 Danilova, N., Bussmann, J., Jekosch, K. & Steiner, L. A. The immunoglobulin heavy-chain locus in zebrafish: identification and expression of a previously unknown isotype, immunoglobulin Z. *Nat Immunol* **6**, 295-302 (2005).
127 Hansen, J. D., Landis, E. D. & Phillips, R. B. Discovery of a unique Ig heavy-chain isotype (IgT) in rainbow trout: Implications for a distinctive B cell developmental pathway in teleost fish. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 6919-6924 (2005).
128 Parra, Z. E., Ohta, Y., Criscitiello, M. F., Flajnik, M. F. & Miller, R. D. The dynamic TCRdelta: TCRdelta chains in the amphibian Xenopus tropicalis utilize antibody-like V genes. *Eur J Immunol* **40**, 2319-2329 (2010).
129 Aderem, A. & Ulevitch, R. J. Toll-like receptors in the induction of the innate immune response. *Nature* **406**, 782-787 (2000).
130 Roach, J. C. *et al.* The evolution of vertebrate Toll-like receptors. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 9577-9582 (2005).
131 Quiniou, S. M., Boudinot, P. & Bengten, E. Comprehensive survey and genomic characterization of Toll-like receptors (TLRs) in channel catfish, Ictalurus punctatus: identification of novel fish TLRs. *Immunogenetics* **65**, 511-530 (2013).
132 Jault, C., Pichon, L. & Chluba, J. Toll-like receptor gene family and TIR-domain adapters in Danio rerio. *Mol Immunol* **40**, 759-771 (2004).
133 Kanwal, Z., Wiegertjes, G. F., Veneman, W. J., Meijer, A. H. & Spaink, H. P. Comparative studies of Toll-like receptor signalling using zebrafish. *Developmental and comparative immunology* **46** (2014).
134 Palti, Y. Toll-like receptors in bony fish: from genomics to function. *Developmental and comparative immunology* **35**, 1263-1272 (2011).
135 Cannon, J. P. *et al.* A bony fish immunological receptor of the NITR multigene family mediates allogeneic recognition. *Immunity* **29**, 228-237 (2008).
136 Yoder, J. A. Form, function and phylogenetics of NITRs in bony fish. *Developmental and comparative immunology* **33**, 135-144 (2009).
137 Desai, S., Heffelfinger, A. K., Orcutt, T. M., Litman, G. W. & Yoder, J. A. The medaka novel immune-type receptor (NITR) gene clusters reveal an extraordinary degree of divergence in variable domains. *BMC evolutionary biology* **8**, 177 (2008).
138 Ferraresso, S. *et al.* Identification and characterisation of a novel immune-type receptor (NITR) gene cluster in the European sea bass, Dicentrarchus labrax, reveals recurrent gene expansion and diversification by positive selection. *Immunogenetics* **61**, 773-788 (2009).
139 Strong, S. J. *et al.* A novel multigene family encodes diversified variable regions. *Proceedings of the National Academy of Sciences of the United States of America* **96**, 15080-15085 (1999).
140 Yoder, J. A. *et al.* Evidence for a transposition event in a second NITR gene cluster in zebrafish. *Immunogenetics* **60**, 257-265 (2008).
141 Yoder, J. A. *et al.* Resolution of the novel immune-type receptor gene cluster in zebrafish. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 15706-15711 (2004).
142 Kawasaki, K. Odontogenic ameloblast-associated protein (ODAM) and amelotin: Major players in hypermineralization of enamel and enameloid. *Journal of Oral Biosciences* **55**, 85-90 (2013).
143 Sire, J. Y., Donoghue, P. C. J. & Vickaryous, M. K. Origin and evolution of the integumentary skeleton in non-tetrapod vertebrates. *Journal of Anatomy* **214**, 409-440 (2009).
144 Prostak, K., Seifert, P. & Skobe, Z. in *Tooth enamel V* (ed R. Fearnhead) (Florence, 1989).
145 Kawasaki, K., Buchanan, A. V. & Weiss, K. M. Biomineralization in humans: making the hard choices in life. *Annu Rev Genet* **43**, 119-142 (2009).
146 Sire, J. Y., Geraudie, J., Meunier, F. J. & Zylberberg, L. On the origin of ganoine: histological and ultrastructural data on the experimental regeneration of the scales of Calamoichthys calabaricus (Osteichthyes, Brachyopterygii, Polypteridae). *Am J Anat* **180**, 391-402 (1987).
147 Kawasaki, K., Suzuki, T. & Weiss, K. M. Phenogenetic drift in evolution: the changing genetic basis of vertebrate teeth. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 18063-18068 (2005).
148 Kawasaki, K. The SCPP gene family and the complexity of hard tissues in vertebrates. *Cells Tissues Organs* **194**, 108-112 (2011).

149 Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J Mol Biol* **268**, 78-94 (1997).
150 Kawasaki, K. & Weiss, K. M. Mineralized tissue and vertebrate evolution: the secretory calcium-binding phosphoprotein gene cluster. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 4060-4065 (2003).
151 Wheelan, S. J., Church, D. M. & Ostell, J. M. Spidey: a tool for mRNA-to-genomic alignments. *Genome research* **11**, 1952-1957 (2001).
152 Quinlan, A. R. BEDTools: The Swiss-Army Tool for Genome Feature Analysis. *Curr Protoc Bioinformatics* **47**, 11 12 11-11 12 34 (2014).
153 Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17** (2011).
154 Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120 (2014).
155 Wu, T. D. & Nacu, S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* **26**, 873-881 (2010).
156 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).
157 Anders, S., Pyl, P. T. & Huber, W. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166-169 (2015).
158 Kawasaki, K., Suzuki, T. & Weiss, K. M. Genetic basis for the evolution of vertebrate mineralized tissue. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 11356-11361 (2004).
159 Kawasaki, K. & Amemiya, C. T. SCPP genes in the coelacanth: tissue mineralization genes shared by sarcopterygians. *J Exp Zool B Mol Dev Evol* **322**, 390-402 (2014).
160 Griffiths-Jones, S., Saini, H. K., van Dongen, S. & Enright, A. J. miRBase: tools for microRNA genomics. *Nucleic Acids Res* **36**, D154-158 (2008).
161 Griffiths-Jones, S., Grocock, R. J., van Dongen, S., Bateman, A. & Enright, A. J. miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res* **34**, D140-144 (2006).
162 Griffiths-Jones, S. The microRNA Registry. *Nucleic Acids Res* **32**, D109-111 (2004).
163 Johnson, M. *et al.* NCBI BLAST: a better web interface. *Nucleic Acids Res* **36**, W5-9 (2008).
164 Lorenz, R. *et al.* ViennaRNA Package 2.0. *Algorithms Mol Biol* **6**, 26 (2011).
165 Desvignes, T., Beam, M. J., Batzel, P., Sydes, J. & Postlethwait, J. H. Expanding the annotation of zebrafish microRNAs based on small RNA sequencing. *Gene* **546**, 386-389 (2014).
166 Batzel, P., Desvignes, T., Sydes, J., Eames, B. F. & Postlethwait, J. H. Prost!, a tool for miRNA annotation and next generation smallRNA sequencing experiment analysis. doi: http://dx.doi.org/10.5281/zenodo.35461 (2015).
167 Flicek, P. *et al.* Ensembl 2013. *Nucleic Acids Res* **41**, D48-55 (2013).
168 Zuker, M. & Stiegler, P. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res* **9**, 133-148 (1981).
169 Westerfield, M. *The Zebrafish Book: A guide for the laboratory use of zebrafish*. 5th edn, (University of Oregon Press, 2007).
170 Desvignes, T. *et al.* miRNA Nomenclature: A View Incorporating Genetic Origins, Biosynthetic Pathways, and Sequence Variants. *Trends Genet* **31**, 613-626 (2015).
171 Muffato, M., Louis, A., Poisnel, C. E. & Roest Crollius, H. Genomicus: a database and a browser to study gene synteny in modern and ancestral genomes. *Bioinformatics* **26**, 1119-1121 (2010).
172 Louis, A., Muffato, M. & Roest Crollius, H. Genomicus: five genome browsers for comparative genomics in eukaryota. *Nucleic Acids Res* **41**, D700-705 (2013).
173 Olena, A. F. & Patton, J. G. Genomic organization of microRNAs. *J Cell Physiol* **222**, 540-545 (2010).
174 Thatcher, E. J., Bond, J., Paydar, I. & Patton, J. G. Genomic organization of zebrafish microRNAs. *BMC genomics* **9**, 253 (2008).
175 Yuan, X. *et al.* Clustered microRNAs' coordination in regulating protein-protein interaction network. *BMC Syst Biol* **3**, 65 (2009).
176 Montavon, T. *et al.* A regulatory archipelago controls Hox genes transcription in digits. *Cell* **147**, 1132-1145 (2011).

177 Maeso, I. *et al.* An ancient genomic regulatory block conserved across bilaterians and its dismantling in tetrapods by retrogene replacement. *Genome research* **22**, 642-655 (2012).
178 Ravi, V. *et al.* Sequencing of Pax6 loci from the elephant shark reveals a family of Pax6 genes in vertebrate genomes, forged by ancient duplications and divergences. *PLoS Genet* **9**, e1003177 (2013).
179 Brudno, M. *et al.* Glocal alignment: finding rearrangements during alignment. *Bioinformatics* **19 Suppl 1**, i54-62 (2003).
180 Frazer, K. A., Pachter, L., Poliakov, A., Rubin, E. M. & Dubchak, I. VISTA: computational tools for comparative genomics. *Nucleic Acids Res* **32**, W273-279 (2004).
181 Woolfe, A. *et al.* Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol* **3**, e7 (2005).
182 Schartl, M. *et al.* The genome of the platyfish, Xiphophorus maculatus, provides insights into evolutionary adaptation and several complex traits. *Nat Genet* **45**, 567-572 (2013).
183 Amores, A. *et al.* A RAD-tag Genetic Map for the Platyfish (Xiphophorus maculatus) Reveals Mechanisms of Karyotype Evolution Among Teleost Fish. *Genetics* **197**, 625-641 (2014).
184 Harris, R. S. *Improved pairwise alignment of genomic DNA* PhD thesis, Pennsylvania State University (2007).
185 Blanchette, M. *et al.* Aligning multiple genomic sequences with the threaded blockset aligner. *Genome research* **14**, 708-715 (2004).
186 Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome research* **15**, 1034-1050 (2005).
187 Anderson, J. L. *et al.* Multiple sex-associated regions and a putative sex chromosome in zebrafish revealed by RAD mapping and population genomics. *PLoS One* **7**, e40701 (2012).
188 Wilson, C. A. *et al.* Wild Sex in Zebrafish: Loss of the Natural Sex Determinant in Domesticated Strains. *Genetics* **198**, 1291-1308 (2014).
189 Hindorff, L. A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 9362-9367 (2009).
190 Maurano, M. T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190-1195 (2012).
191 Fuxman Bass, J. I. *et al.* Human gene-centered transcription factor networks for enhancers and disease variants. *Cell* **161**, 661-673 (2015).
192 Hinrichs, A. S. *et al.* The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res* **34**, D590-598 (2006).
193 Visel, A., Minovitsky, S., Dubchak, I. & Pennacchio, L. A. VISTA Enhancer Browser--a database of tissue-specific human enhancers. *Nucleic Acids Res* **35**, D88-92 (2007).
194 Nikaido, M. *et al.* Coelacanth genomes reveal signatures for evolutionary transition from water to land. *Genome research* **23**, 1740-1748 (2013).
195 Berlivet, S. *et al.* Clustering of tissue-specific sub-TADs accompanies the regulation of HoxA genes in developing limbs. *PLoS Genet* **9**, e1004018 (2013).
196 Brudno, M. *et al.* LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome research* **13**, 721-731 (2003).
197 Fisher, S. *et al.* Evaluating the biological relevance of putative enhancers using Tol2 transposon-mediated transgenesis in zebrafish. *Nat Protoc* **1**, 1297-1305 (2006).
198 Kimmel, C. B., Ballard, W. W., Kimmel, S. R., Ullmann, B. & Schilling, T. F. Stages of embryonic development of the zebrafish. *Dev Dyn* **203**, 253-310 (1995).
199 Suster, M. L., Abe, G., Schouw, A. & Kawakami, K. Transposon-mediated BAC transgenesis in zebrafish. *Nat Protoc* **6**, 1998-2021 (2011).
200 Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology* **10**, R25 (2009).
201 Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome biology* **11**, R106 (2010).
202 R: A Language and Environment for Statistical Computing (R Foundation for Statistical Computing, Vienna, Austria, 2015).
203 Star, B. *et al.* The genome sequence of Atlantic cod reveals a unique immune system. *Nature* **477**, 207-210 (2011).

204 Hellsten, U. *et al.* The genome of the Western clawed frog Xenopus tropicalis. *Science* **328**, 633-636 (2010).
205 Alfoldi, J. *et al.* The genome of the green anole lizard and a comparative analysis with birds and mammals. *Nature* **477**, 587-591 (2011).
206 Wan, Q. H. *et al.* Genome analysis and signature discovery for diving and sensory properties of the endangered Chinese alligator. *Cell Res* **23**, 1091-1105 (2013).
207 International Chicken Genome Sequencing, C. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* **432**, 695-716 (2004).
208 Waterston, R. H. *et al.* Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520-562 (2002).
209 Aparicio, S. *et al.* Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. *Science* **297**, 1301-1310 (2002).
210 Nonaka, M. I., Aizawa, K., Mitani, H., Bannai, H. P. & Nonaka, M. Retained orthologous relationships of the MHC Class I genes during euteleost evolution. *Molecular biology and evolution* **28**, 3099-3112 (2011).
211 Lukacs, M. F. *et al.* Comprehensive analysis of MHC class I genes from the U-, S-, and Z-lineages in Atlantic salmon. *BMC genomics* **11**, 154 (2010).
212 Saha, N. R. *et al.* Genome complexity in the coelacanth is reflected in its adaptive immune system. *J Exp Zool B Mol Dev Evol* **322**, 438-463 (2014).
213 Letunic, I., Doerks, T. & Bork, P. SMART: recent updates, new developments and status in 2015. *Nucleic Acids Res* **43**, D257-260 (2015).
214 Saitou, N. & Nei, M. The Neighbor-Joining Method - a New Method for Reconstructing Phylogenetic Trees. *Molecular biology and evolution* **4**, 406-425 (1987).
215 Felsenstein, J. Confidence-Limits on Phylogenies - an Approach Using the Bootstrap. *Evolution* **39**, 783-791 (1985).
216 Zuckerkandl, E. & Pauling, L. in *Evolving Genes and Proteins* (eds V. Bryson & H. J. Vogel) 97-166 (Academic Press, 1965).