SUPPLEMENTARY NOTE

The genome of the bowfin (*Amia calva*) illuminates the developmental evolution of ray-finned fishes

Andrew W. Thompson, M. Brent Hawkins, Elise Parey, Dustin J. Wcisel, Tatsuya Ota, Kazuhiko Kawasaki, Emily Funk, Mauricio Losilla, Olivia E. Fitch, Qiaowei Pan, Romain Feron, Marine Milhes, Brett L. Racicot, Kevin L. Childs, Quenton Fontenot, Allyse Ferrara, Solomon R. David, Amy R. McCune, Alex Dornburg, Jeffrey A. Yoder, Yann Guiguen, Hugues Roest Crollius, Camille Berthelot, Matthew P. Harris, and Ingo Braasch*

* Corresponding author: Ingo Braasch

Email: braasch@msu.edu

Supplementary Notes 1-9

Supplementary Figures 1-17

- Suppl. Fig. 1. Size distribution of the 30 largest scaffolds in the AmiCal1 assembly
- Suppl. Fig. 2. OrthoFinder species tree
- Suppl. Fig. 3. Bowfin pool-sex analysis
- Suppl. Fig. 4. Orthology of holostean and chicken karyotypes
- Suppl. Fig. 5. MHC cladograms
- Suppl. Fig. 6. Genomic organization of bowfin immunoglobulin and T cell receptor genes
- Suppl. Fig. 7. IgH cladograms
- Suppl. Fig. 8. IGL cladograms
- Suppl. Fig. 9. T cell receptor (TCR) cladograms
- Suppl. Fig. 10. Toll-like receptor phylogeny
- Suppl. Fig. 11. Teleost *scpp5* and *scpp1* genes
- Suppl. Fig. 12. Nucleosome periodicity in ATAC-Seq data
- Suppl. Fig. 13. Heatmap of pairwise Jaccard distances among bowfin OCR profiles
- Suppl. Fig. 14. The holostean bridge connects vertebrate *Tbx4* gene enhancers
- Suppl. Fig. 15. Bowfin *hoxd14* pseudogene
- Suppl. Fig. 16. Bowfin pectoral fin Evo-Devo
- Suppl. Fig. 17. Genomic alignment of bony vertebrate *fgf8* gene regions

Supplementary Tables 1-21

- Suppl. Tab. 1. Assembly statistics for bowfin genome AmiCal1
- Suppl. Tab. 2. Repeat content in bowfin in comparison to spotted gar
- Suppl. Tab. 3. OrthoFinder summary
- Suppl. Tab. 4. Orthogroup statistics for bowfin and 11 other vertebrates
- Suppl. Tab. 5. Reference-free k-mer analysis of bowfin sex
- Suppl. Tab. 6. Bowfin MHC genes*
- Suppl. Tab. 7. Gar MHC genes*
- Suppl. Tabs. 8-12. Immune gene accessions: MHC, IgH, IgL, TCR, TLR*
- Suppl. Tab. 13. SCPP gene expression in zebrafish skin at three different ages
- Suppl. Tab. 14. Open Chromatin Regions found for each bowfin developmental stage
- Suppl. Tab. 15. Number of bowfin OCRs found in N developmental stages
- Suppl. Tab. 16. HOMER annotation of OCRs in each developmental stage
- Suppl. Tab. 17. Overlap of bowfin OCRs, bowfin UCEs and gar-centric CNEs
- Suppl. Tab. 18. Number of VISTA enhancers detected in bowfin and zebrafish genomes
- Suppl. Tab. 19. Human VISTA enhancers in bowfin OCRs*
- Suppl. Tab. 20. Mouse OCRs found in bowfin
- Suppl. Tab. 21. Number of tissue specific mouse OCRs in bowfin
- Suppl. Tab. 22. Putative *fgf8* regulatory regions in bowfin and their ATAC-Seq profile

Supplementary Files 1-2

- Suppl. File 1. Bowfin SCPP gene predictions*
- Suppl. File 2. Bowfin hox gene transcripts*

* separate files

Supplementary Note 1. Genome sequencing and assembly

1.1 Genome specimen. The bowfin genome assembly is based on DNA extracted from blood drawn of a single wild male individual ("Calvin", sample ID Aca15.3), collected from the Atchafalaya Basin population near Stephensville, Louisiana, USA (coordinates 29.812569 N 91.220803 W), and sacrificed in the laboratory of Allyse Ferrara (Nicholls State University, Thibodaux, LA). Information on the genome specimen is filed under NCBI BioProject PRJNA417081, BioSample SAMN07977036.

1.2 Genome assembly. The bowfin genome was sequenced and assembled in partnership with Dovetail Genomics [https://dovetailgenomics.com/]. The Meraculous[2] *de novo* genome assembly consisted of 433.5 million read pairs assembled into 64,769 1Kb+ scaffolds with a total length of 768.2 Mb. The Chicago[3] assembly consisted of 177M read pairs and resulted in 62,981 joins reducing the number of 1Kb+ scaffolds to 1,846 and increasing total assembly length to 775.8 Mb. The Hi-C library consisted of 527M read pairs and the final assembly (Chicago+HiC scaffolded with HiRise[3] resulted in 1,958 1Kb+ scaffolds with a total length of 831 Mb after additional 251 additional joins. The kmer-based estimated genome size of bowfin is 0.91 Gb. The bowfin genome size was previously estimated to be ~1.1 Gb based on Feulgen staining[4]. Scaffolds of the genome were ordered from longest to shortest and labeled as Aca_scaf_1 to Aca_scaf_1957. The bowfin karyotype is n=23[5,6]. The 23 largest assembly scaffolds range from 57.1 Mb and 20.5 Mb in size, with a clear drop in length to 319 kb for the 24th largest scaffold (Supplementary Fig. 1). The assembly is thus at the chromosome-level with 99% of the assembly being represented by the 23 largest superscaffolds. The mitochondrial genome was identified as a single scaffold (Aca_scaf_1958) based on sequence comparison to the published *A. calva* mitogenome (NCBI RefSeq: NC_004742.1). BUSCO[7] and CEGMA[8] scores of 100% (303/303 eukaryotic BUSCOs) and 98.3% (244/248 eukaryotic CEGs), respectively, further support the high quality and completeness of the final assembly AmiCal1 (NCBI accession: PESF00000000). See Supplementary Table 1 for assembly statistics.

Supplementary Note 2. Genome annotation

2.1 Repeat analysis. 22.1% of the bowfin genome is repetitive, very similar to the 20.6% repeat content of spotted gar[9], although there are clear differences in the distribution among repeat element types between the two species (Supplementary Table 2). Bowfin has a higher content of DNA transposons than gar (7.1% vs. 3.5% of the genome assembly), a lower LINE content (3.9% vs. 5.8%), a higher level of satellite/simple repeat elements (2.1% vs. 0.3%), and a lower SINE content (0.3% vs. 2.8%). The DNA transposon Zisupton, absent from spotted gar[9] and considered specific to teleosts[10], is found in 6,000+ copies in the bowfin genome, showing that this superfamily is present outside of teleosts and that it has been secondarily lost in the spotted gar lineage after divergence from bowfin.

2.2 Gene annotation and orthology prediction. MAKER[11] annotated 49,283 putative gene models based on evidence and model predictions (MAKER-Max). After screening for homology with Pfam protein domains[12], we retained 21,948 genes with transcriptional and/or Pfam domain evidence as the final MAKER-Standard gene set. This number is a realistic expectation when compared to spotted gar, which has 21,443 MAKER annotated genes[9]. OrthoFinder[13] placed 86.8% of bowfin genes into orthogroups and identified 7,532 orthogroups present in all included species (gar, coelacanth, mouse, chicken, human, western clawed frog, anole lizard, zebrafish, medaka, arowana, and elephant shark). OrthoFinder output is summarized in Supplementary Tables 3-4. The species tree generated by OrthoFinder via STAG[14] supports holostean monophyly (Supplementary Fig. 2).

Supplementary Note 3. Sex determination analysis

Results and Discussion. Pool-Seq analyses have been shown to be effective in revealing sex determination systems in many species (e.g.[15-18]). Here, we search for regions showing male versus female differences on the bowfin genome assembly using sex-specific Pool-Seq reads from 30 males and 30 females. We computed FST, number of sex-specific SNPs, and ratio of read depth between males and females in a sliding window along the genome and found no large region showing a clear difference between sexes (Supplementary Fig. 3). It is possible that the male reference genome represents the homogametic sex, and thus would not allow the detection of sex-specific regions using a reference-based approach. Consequently, we complemented our reference-based analyses with a reference-free approach searching for enrichment of sex-specific k-mers, which has been used as an indication for sex determination systems in other species[19-21]. These k-mer analyses consistently revealed a high number of both male- and female-biased k-mers with only a slight excess of male-biased k-mers compared to female-biased k-mers (Supplementary Table 5).

Altogether, these results suggest that if bowfin has a genetic sex determination system, it is either more complex than a simple monofactorial heterogametic system (XX/XY or ZZ/ZW) or its sex locus is too small to be detected with confidence using a Pool-Seq approach. Complex polyfactorial systems have been described in vertebrates[22], and pure environmental sex determination or combinations of environmental effects and genetic sex determination are common among bony fishes[23]. But these complex cases cannot be easily identified by contrasting pools of male and female reads. To shed light on bowfin sex determination, other approaches would be needed. For example, whole genome resequencing of multiple males and females at single base pair resolution for individual fish. This could enable the identification of complex sex determination systems and/or very small sex loci that are found in some fish species (e.g.[24]).

Supplementary Note 4. The bowfin immune gene repertoire

4.1 Immune tissue transcriptome. A NovaSeq6000 sequencing run from immune tissues (gill, spleen, and intestine) generated 7.6 Gb of reads that were mapped to the reference genome and assembled into 91,498 transcripts with Trinity[49,50] (N50 = 3,054 bp). More than 93% of cleaned reads mapped to the genome-guided transcriptome assembly, suggesting a high-quality assembly. BUSCO[7] analysis confirmed 72.9% (3,630/4,584) complete Actinopterygii (orthoDB 9) ultraconserved orthologs are present in the transcriptome. NCBI accession numbers: raw sequencing reads SRR11303972 (SRA); assembled transcripts GIOP00000000 (TSA).

4.2 Major Histocompatibility Complex (MHC) sequences

4.2.1 MHC Background. MHC molecules present antigenic (mostly foreign) peptides processed in host cells to T-cells, thereby enabling immunological cells to recognize foreign antigens and eliminate infected host cells upon binding of cell surface T-Cell Receptors (TCRs) with MHC-antigen complexes. Two different types of MHC molecules are present in vertebrates: class I and class II. Classical class I molecules consist of MHC class I alpha and beta-2-microglobulin (B2M) heterodimers and classical class II molecules consist of MHC class II alpha and beta heterodimers. Class I alpha proteins consist of three extracellular domains: alpha1, alpha2, and alpha3 domains, that play roles in displaying peptide antigens (alpha1 and alpha2) and dimerize with B2M (alpha3). In humans, MHC class III genes are present between classical class I and class II genes and, although not directly involved in antigen presentation, many class III genes play important roles in inflammation and cell signaling[25,26]. In the genomes of tetrapods (including humans) and cartilaginous fish, the MHC class I and class II genes are linked on a

single chromosome[27]. However, in teleosts, class I and class II genes are typically present on different chromosomes[28,29]. Earlier efforts to determine the genomic organization of MHC genes in the non-teleost spotted gar was hampered by its MHC genes being located on a large number of small scaffolds[9].

4.2.2 MHC Approach. BLAST searches[30] of the bowfin reference genome using spotted gar MHC sequences as queries identified bowfin scaffolds encoding MHC class I and class II loci. Portions of these scaffolds (encoding MHC-related sequences) were used as queries in subsequent BLASTN and TBLASTN searches of the bowfin genome and identified additional genes encoding MHC class I and MHC class II loci. BLASTX analyses were used to confirm the identity of MHC sequences.

The majority of MHC class I and all class II genes were identified in a single scaffold, Aca_scaf_14. This scaffold was examined closely to identify MHC-related as well as non-MHC genes. Repeat masked scaffold sequences were used as queries for BLASTX searches to identify coding segments within the scaffold. Coding sequences were then used as queries in BLASTN searches of bowfin RNA-seq transcriptomes to identify expressed genes. If alternatively spliced transcripts were identified, the longest isoforms were selected and used to identify exons within Aca_scaf_14. Exon/intron boundaries for each gene were predicted manually. If only partial bowfin transcripts were identified, homologous genes and transcripts from other vertebrates (mostly fishes) were collected and used to predict missing peptide sequences. When conserved peptide sequences for a specific gene were absent from bowfin transcriptomes, TBLASTN searches were conducted to identify corresponding fragments in the genomic region. In general, BLASTN searches were used with identified nucleotide sequences to obtain exact genomic positions of exons as well as any homologous genes/exons present in the bowfin genome.

The immune transcriptome and exon sequences on Aca_scaf_14 were used to infer the protein sequences encoded by these genes. Using the identified nucleotide and/or protein sequences encoded on Aca_scaf_14 as queries for BLASTN or BLASTP analyses, respectively, homologs were identified in human, zebrafish and spotted gar. Human and zebrafish genes with the lowest and/or near the lowest e-values from BLAST searches were located on human chromosome 6 and zebrafish chromosomes 15, 16, and 19, and, in many cases and they were considered as orthologs. When some ambiguity remained (e.g. human genes identified with the lowest or near the lowest e-values were located on Hsa1, Hsa9, or Hsa19 which contain the MHC paralogous groups), sequences were analyzed phylogenetically to determine if the identified bowfin gene was a paralog or ortholog of human/zebrafish genes. MHC class I and class II sequences were aligned with ClustalW[31] and their evolutionary relationships were inferred using maximum likelihood in MEGAX[32] with the complete deletion option, 200 bootstrap replicates, and the best fit substitution model.

4.2.3 MHC Results and Discussion. In bowfin, we identified numerous linked MHC class I, class II, and class III genes on a single superscaffold, Aca_scaf_14, that shares conserved synteny with the human MHC and multiple zebrafish loci (main text Fig. 2; Supplementary Fig. 5, Supplementary Table 6). This observation suggests that the class I and class II genes of other non-teleost ray finned fishes are linked, and that the “unlinking” of class I and class II genes in teleosts may have resulted from the teleost genome duplication and subsequent events (e.g. recombination, gene gain/loss, pseudogenization, etc.).

Although all identified bowfin MHC class II (alpha and beta) genes are present in a single cluster on Aca_scaf_14 (main text Fig. 2; Supplementary Table 6), several class I (alpha) genes are present on other scaffolds (Aca_scaf_2: 38,351,553-38,358,157; Aca_scaf_5: 32,640,311-32,642,730; Aca_scaf_6: 26,898,782-26,902,075; Aca_scaf_8: 39,209,049-39,217,024; Aca_scaf_12: 4,105,156-4,123,775, 9,159,909-9,785,271; Aca_scaf_18: 1,532,678-1,533,916, 13,231,672-13,233,334; Aca_scaf_49: 30,866-36,567), and B2M is encoded on scaffold Aca_scaf_18 (Supplementary Fig. 6A). These observations illustrate that duplicated MHC class I genes have been translocated to other genomic regions in bowfin and suggest that similar variations in MHC organization may be present across holosteans. Tight linkage of MHC class I and class II genes is characteristic of tetrapod MHC genes, while dispersed genomic

localization of MHC class I genes is typical for teleosts. As such, the bowfin reference genome provides important clues on the evolutionary origin and transition to the teleost type of MHC gene organization.

Teleosts encode up to six different lineages of MHC class I sequences (U, Z, S, L, P and H). All have been described in spotted gar, though some ambiguity remains for the S lineage[9,33,34] (Supplementary Table 7). The U lineage has been identified in spotted gar and all studied teleosts (except a few sexually parasitic anglerfishes[35]) and is the only one with genes that display classical type polymorphisms, although it also includes monomorphic non-classical class I genes. The Z lineage also has been described in all studied teleosts and spotted gar. Z lineage proteins share some characteristics with classical MHC class I genes (e.g. may bind peptide antigens) but display more conservation in their alpha1 and alpha2 domains suggesting a limited capacity for antigen display. In contrast, the nonclassical lineages, S, L, P and H have been identified in some, but not all teleost lineages[33,34].

Here we identify MHC class I sequences of the U, Z, S, P, L and H lineages encoded within the bowfin genome (Supplementary Fig. 5, Supplementary Table 8). Phylogenetic analyses of MHC class I genes indicate that MHC class I genes of the U lineage as well as the P, Z and L lineages are present on Aca_scaf_14, even though many of them are likely pseudogenes because of in-frame stop codons, frameshift mutations, deletions or insertions, and some missing exons. MHC genes/pseudogenes of the P lineage are also located on Aca_scaf_12, Aca_scaf_2 and Aca_scaf_18, while an MHC pseudogene of the L lineage is located on Aca_scaf_18. MHC genes/pseudogenes of the S lineage are found on Aca_scaf_5, Aca_scaf_8, Aca_scaf_12, and Aca_scaf_49. Additionally, the recently annotated MHC gene of the H lineage[34] is detected on Aca_scaf_6. Among MHC class I molecules identified, most are composed of alpha1, alpha2, and alpha3 domains with transmembrane domains and cytoplasmic tails. A few exceptional cases nonetheless exist: *mhc1uaa*, *mhc1uea*, *mhc1pda*, and *mhc1pfa* seem to contain multiple exons coding alpha1, alpha2, and/or alpha3 domains, but it remains to be tested if these genes encode functional proteins.

In contrast to MHC class I genes, diversified lineages of both alpha and beta MHC II genes are located in a cluster on Aca_scaf_14 and are not identifiable in other scaffolds/regions (main text Fig. 2; Supplementary Table 6). A pair of alpha and beta loci are juxtaposed with the same transcriptional orientation for most of the bowfin MHC class II genes, although this organization needs to be confirmed. Until now, MHC class II genes of teleost fishes have been classified into three lineages (DA, DB, and DE) and MHC class II genes of tetrapods have been classified into two lineages (DM and tetrapod classical)[36]. Most of bowfin and spotted gar MHC class II genes identified are distinct from teleost MHC class II genes except those of the DE lineage. Additional bowfin and spotted gar lineages identified here were named as DG, DH, DI, DJ, DK, DL, and DN. Although the DG lineage sequences may be orthologous to teleost DA/DB lineages, it was named here as DG, because they are somewhat different from those of teleost DA/DB lineages (see also[36]). It is nonetheless apparent that holostean fishes possess distinct and diverse lineages of MHC class II genes that are not present in teleosts.

The analyses of the genes identified from one end of Aca_scaf_14 to the MHC class I genes of the L lineage (main text Fig. 2; Supplementary Table 6) revealed that synteny is well conserved between bowfin and humans, emphasizing the utility of the bowfin genome to infer the ancestral states of early bony vertebrate genomes. Synteny with zebrafish was observed to lesser extent and restricted to regions of chromosomes where one-to-two orthologs corresponded to some genes though one-to-one orthologous correspondences were most commonly observed. Due to the teleost genome duplication, it is possible that two co-orthologs exist in teleosts for some bowfin and human genes, but a more common one-to-one mapping in zebrafish suggests that frequently only one of two TGD duplicates has been retained. Comparison of the bowfin and human genomes supports the hypothesis that the ancestral genome of vertebrates exhibited tight linkage among MHC class I, II and III genes. Such a condition was dissolved in teleosts by reciprocal loss of TGD paralogs from the initial two teleost MHC I/II/III paralogons and further chromosomal rearrangements, leading to the apparent dissolution of MHC class I, II and III linkage in teleosts. For example, zebrafish classical MHC class I U lineage genes are encoded on

chromosomes 19 and 22, MHC class I Z lineage genes are present on zebrafish chromosomes 1 and 3, and classical MHC class II genes are encoded on zebrafish chromosomes 4, 8, and 18[37,38]. It is also worth noting that numerous orthologs from the human and bowfin MHC regions are present on Dre19 and Dre16 (main text Fig. 2) which are TGD-paralogous chromosomes carrying for example the *hoxAa* and *hoxAb* gene clusters, respectively.

4.3 Immunoglobulin (Ig) and T Cell Receptor (TCR) genes

4.3.1 Ig and TCR Background. Antigen recognition receptors of the adaptive immune system, i.e. immunoglobulin (Ig) genes and T-cell receptor (TCR) genes that undergo V(D)J recombination, have been identified in all lineages of jawed vertebrates. However, their structure and genomic organization can differ between major lineages[39,40]. Most teleosts encode three classes of antibodies [IgM, IgD and IgT (a.k.a. IgZ in zebrafish)] and canonical TCR loci (TCRα, TCRβ, TCRγ, and TCRδ). We previously reported a single, tightly linked TCRα/δ locus in spotted gar[9].

4.3.2 Ig and TCR Approach. BLAST searches[30] using immunoglobulin (Ig) and T cell receptor (TCR) constant gene segments of spotted gar and other vertebrates as queries identified bowfin scaffolds encoding various types of antigen recognition molecules of the adaptive immune system using the strategy described in 4.2.2. The immune transcriptome was utilized to clarify gene structures. In addition, a custom program was used to locate potential recombination-signal sequence (RSS) neighboring V, D, and J gene segments of immunoglobulin (Ig) and T-cell receptor (TCR) genes by identifying heptamer/nanomer motifs with 12/13 or 22/23 bp spacers. Deduced amino acid sequences of Ig and TCR were aligned using ClustalW[31] and analyzed to infer their phylogenetic relationships in MEGAX[32]. The best fitting amino acid substitution models (see figure legends) were used to infer phylogenetic relationships in a maximum likelihood framework with complete deletion options and 200 bootstrap replicates.

4.3.3 Ig and TCR Results and Discussion. Our investigation of the bowfin genome revealed a single TCRα/δ locus, a single TCRβ locus, and a single TCRγ locus (Supplementary Fig. 6A-C). We identified a single Ig heavy (IgH) chain locus (Supplementary Fig. 6A), and a few Ig light (IgL) chain loci including an Igκ locus (Supplementary Fig. 6D) and Igσ loci (Supplementary Fig. 6E). The IgH chain locus and the TCRα/δ locus are linked on a single superscaffold (Aca_scaf_18), separated by approximately 12 Mbp (Supplementary Fig. 6A). With the exception of a few IgL chain loci, the Ig loci (including the Igκ locus), and TCR loci exhibit translocon type gene organization, i.e., tandem duplications of V, D, J gene segments and C exons represented by Vn-(Dn-)-Jn-C (Supplementary Fig. 6).

Immunoglobulin heavy chain (IgH) genes: One notable feature of the IgH locus on Aca_scaf_18 is the presence of three different types of constant gene segments. The two constant gene segments located closer to the TCRα/δ locus are of a Cμ and Cδ type, while another constant gene segment is present at the other end (Supplementary Fig. 6A). This other constant region consists of four Ig domains (C1 - C4). Maximum likelihood inference of the phylogenetic relationships of this sequence with other vertebrate Ig constant regions with four Ig domains (Supplementary Fig. 7A**,** Supplementary Table 9) indicated a closer affinity of this novel bowfin C region to the Cτ (Cζ in zebrafish) constant region of teleost fish. Likewise, separate phylogenetic analyses of the C1 and C4 domains with C1 domains and with the carboxyl-terminal constant Ig domains (respectively) from a range of immunoglobulins and species suggest an affinity of the bowfin C4 gene segment to the Cτ/Cζ of teleosts at the carboxyl-terminal domain (Supplementary Fig. 7C, Supplementary Table 9). On the other hand, they suggest an affinity to Cω (of the chondrichthyan sandbar shark) at the C1 domain (Supplementary Fig. 7B, Supplementary Table 9), though with weak support (BSS =33%). Our results also indicate that teleost fish Cτ have an affinity with teleost C1, suggesting past genetic exchange (i.e. exon shuffling, gene conversion, etc.) between these domains in teleosts. We further determined that an ortholog of this Cτ constant gene segment found in teleosts and bowfin is in fact present in the spotted gar genome on an

unplaced scaffold JH591552.1 (Supplementary Fig. 7), despite being considered absent in earlier analyses[9]. As this C gene segment found in both bowfin and spotted gar likely represents an ancestral homolog of the teleost Cτ, we consider it a Cτ constant gene segment that dates back at least to the common ancestor of crown neopterygians. Our conclusion based on bowfin genomic analyses is consistent with a recent investigation of bowfin transcriptome data[41]. These observations thus indicate that IgT-like antibodies arose prior to the origin of teleosts and raise the question about the function of IgT. Further studies in bowfin, gars, and other non-teleosts will be necessary to clarify if IgT mucosal host defense is a derived feature of teleosts or a more universal role.

Our results raise the question of whether the mu, delta, or tau type of IgH utilize the entire repertoire of V gene segment of the IgH locus, since the presence of non-IgH genes (e.g. Ig superfamily proteins, C-type lectin domain family 4, phosphatidylinositol-4,5-bisphosphate 4-phosphatase 1, and immunoglobulin light chain), in the middle of V gene segment cluster of IgH locus, may interfere with the utilization of distantly located V gene segments and allow only the utilization of proximal V gene segments for each isotype.

Immunoglobulin light chain (IgL) genes: Using sequences from multiple vertebrate lineages, including the genome of the spotted gar and transcriptomes from saddled bichir (*Polypterus endlicheri*) and ropefish (*Erpetoichthys calabaricus*), five ancient vertebrate IgL isotypes have been described: kappa (κ), lambda (λ), lambda-2 (λ-2), sigma (σ), and sigma-2 (σ-2; previously sigma-cart) [42]. Spotted gar was reported as the only known species to possess all five isotypes (although the spotted gar IgL lambda sequence was reported as a pseudogene). Accordingly, we identified evidence for all five isotypes of bowfin immunoglobulin light chain loci on different scaffolds (Supplementary Fig. 8, Supplementary Table 10). Among them, an IgL kappa locus (Aca_scaf_11: 2,079,706 -2,103,176, 5,091,825-7,188,393) contains large and extensive numbers of V gene segments under a translocon type gene organization, likely representing the primary immunoglobulin light chain of the species, together with IgL sigma loci (Aca_scaf_22: 14,393,451-14,401,326, 16,948,268-16,963,957) (Supplementary Fig. 6D,E). Bowfin IgL sigma-2 sequences can be identified in the middle of the IgH locus (Aca_scaf_18) and shares homology with sigma-2 sequences in shark and spotted gar. Although bowfin sequences similar to IgL lambda (Aca_scaf_15: 25,075,577-25,090,458) and lambda-2 (Aca_scaf_1: 30,303,148-30,306,638) can be identified, the presence of in-frame stop codons (in both genome and transcriptome sequences) in the single J gene segment in the lambda locus and in the single V gene segment in the lambda-2 locus indicates that, unless these sequences are properly mutated at the somatic level, functional protein products would not be generated. Further investigation is required to determine the functionality of all five IgL lineages in bowfin and gar.

T cell receptor (TCR) genes: Three bowfin loci encoding four different types of TCRs were identified: TCRα/δ (Aca_scaf_18: 22,416,394-23,841,556), TCRβ (Aca_scaf_21: 27,399-316,845), and TCRγ (Aca_scaf_4: 38,763,058-38,840,380). Schematic representations of TCR loci with translocon type gene organization are shown in Supplementary Fig. 6. This mirrors TCR lineages in other vertebrates (e.g. see human, frog, and catfish TCR lineages in Supplementary Fig. 9, Supplementary Table 11). In bowfin, extensive diversity is observed for the V and J gene segments of the TCR Vα/δ gene segments and the Jα gene segments, as well as moderate diversity for Vβ and Vγ gene segments. Our finding of four different types of TCRs are consistent with expectations from the hypothesized early evolutionary origin of distinct sets of TCR subunits and their functional importance among vertebrate species[43].

4.4. Toll-like receptors (TLRs)

4.4.1 TLR Background. The family of Toll-like Receptor (TLRs) are present in both vertebrates and outside vertebrates and provide one of the initial immune responses to infection through recognition of a variety of pathogen-associated molecular patterns (PAMPs)[44,45]. Six major families of TLRs have been

described (TLR1, TLR3, TLR4, TLR5, TLR7, and TLR 11) with different species possessing varying numbers of genes within each family. Human and mouse encode ten and twelve TLR genes respectively. However, teleost genomes can encode more than 20 TLR genes[38,44,46]. Although the larger number of teleost TLRs could be the result of the teleost genome duplication, a total of 16 TLR genes plus one pseudogene were reported from spotted gar representing all six TLR families[9,47]. Therefore, determining the extent of TLR diversity in bowfin is critical to further polarizing the changes in TLR diversity following the teleost genome duplication.

4.4.2 TLR Approach. Bowfin Toll-like receptors (TLRs) were identified via TBLASTN searches using the amino acid sequences of well-established TIR domains[47] and annotated based on clustering within the phylogenetic analyses. TIR domains were aligned via Clustal Omega[48] with priority given to TIR domains identified from the MAKER genome annotation, followed by sequences from the PhyloFish database[49], the immune tissue transcriptome, and significant hits in the genome. The evolutionary relationships among TLR receptors were analyzed using the maximum likelihood approach available in IQ-TREE 2[50]. Analyses were conditioned on the best-fit model of protein evolution identified by IQ-TREE from a candidate pool of models that spanned all common amino-acid exchange models to protein mixture models. We conducted 1,000 ultrafast phylogenetic bootstraps [51,52] to assess support for evolutionary relationships.

For genomic loci in which BLAST identified a TIR domain but lacked a gene model, we scanned upstream sequences for the presence of leucine-rich-repeats (LRR). 50 kbp upstream of each TIR domain was translated (3-frame) into amino acid sequences and searched for leucine-rich repeats (pfam clan CL0022) using HMMER3[53]. In all cases, at least one LRR was found, supporting the notion that these “orphan” TIR domains belong to unannotated TLR genes, and not to other genes that utilize TIRs (e.g., *myd88*).

4.4.3 TLR Results and Discussion. Here we report 20 bowfin TLRs Supplementary Table 12) annotated based on phylogenetic clustering with defined TLRs from other species (Supplementary Fig. 10). TLR4, TLR9, TLR20, TLR21 and TLR25 appear to be duplicated in bowfin. In four of these instances (TLR4, TLR9, TLR20, and TLR21), the duplications are found in close proximity on the same genomic scaffold suggesting tandem gene duplications (Supplementary Table 12). TLR25a and TLR25b are almost 10 Mbp apart on Aca_scaf_14 suggesting a different mechanism of duplication.

In spotted gar, TLR1, TLR2, and TLR3 are found on the same chromosome (LG4). Bowfin TLR2 and TLR3 are found on the same scaffold Aca_scaf_12, and bowfin TLR1 is isolated on scaffold Aca_scaf_11. The duplication of TLR8 observed in spotted gar was not detected in bowfin, but bowfin TLR7 and TLR8 are found in close genomic proximity as observed in spotted gar and other vertebrate species[47]. Bowfin appear to encode six members of the TLR11 subfamily, while only one was previously reported in spotted gar, further showing the diversity of TLR repertoires even among holosteans (Supplementary Fig. 10, Supplementary Table 12). Bowfin has an ortholog to a previously identified and highly diverged gar TLR which was originally named TLR19 or TLR19-like[9,47]. Due to the identification of an authentic TLR19 in bowfin (Supplementary Fig. 10), the uncertainty as to which TLR subfamily this previously identified sequence belongs, and the apparent restriction of this sequence to holosteans, we refer to this gene as TLR-HS (holostean-specific), providing another line of evidence in support of holostean monophyly.

Supplementary Note 5. SCPP genes and scale formation

5.1 Approach. We investigated differences in the repertoire of SCPP genes in bowfin, gar, and selected teleosts to reveal SCPP genes that are important for scale formation in Holostei and other actinopterygians. We searched for the bowfin genomic regions syntenic to the two spotted gar SCPP

gene clusters and gar *scpplpq20* using TBLASTN[30] with gar SCPP amino acid sequences[9,54] as queries. All intron-spanning reads in these syntenic regions and adjacent regions identified using RNA-seq data were investigated to identify SCPP genes that possess a characteristic exon-intron structure, including the presence of entirely untranslated exon 1, all phase-0 introns, and the entire signal-peptide encoding region in exon 2[55].

To analyze SCPP gene expression during zebrafish skin development, relative expression levels were estimated using Galaxy[56] for published skin RNA-Seq datasets[57], trimmed using Trimmomatic[58], and aligned to the zebrafish genome sequence (danRer11) using STAR[59]. Three different sets for each age from NCBI SRA (5 months: SRR850591, SRR850594, SRR850597; 24 months: SRR850598, SRR850601, SRR850603; 42 months: SRR850604, SRR850607, SRR850608) were merged. Genomic coordinates of SCPP genes were determined using Splign[60]. FPKM values were calculated using Cufflinks[61] based on BAM files obtained from STAR. Galaxy analyses used default conditions except additional options in Trimmomatic (ILLUMINACLIP and MINLEN=50) and Cufflinks (multi-read correction).

5.2 Results and Discussion. In the bowfin genome, we identified 22 SCPP genes Supplementary File 1, of which 21 genes form two large clusters. The arrangement of these bowfin SCPP genes is similar to that of gar SCPP genes (main text Figure 3). One cluster on bowfin 12/gar LG4 consists of two subclusters of SCPP genes; genes in one subcluster encode Pro and/or Gln (P/Q)-rich SCPPs, and genes in the other subcluster encode acidic SCPPs. These two subclusters are separated by *sparcl1*, which is evolutionarily related to SCPP genes[54]. Genes in the other cluster on bowfin 9/gar LG2 encode one acidic SCPP (Spp1) and three or more P/Q-rich SCPPs (main text Figure 3). The only exception is *scpplpq20* that is isolated from these two SCPP gene clusters on bowfin 22/gar LG22.

The total number of SCPP genes identified in the bowfin genome is considerably smaller than that found in the gar genome, in which we detected 38 SCPP genes, the largest number of SCPP genes found to date[9,54]. All 22 bowfin SCPP genes have orthologs in the gar genome, but the orthologs of 16 gar SCPP genes were not detected in the bowfin genome (main text Figure 3). This result implies that various SCPP genes were secondarily lost in the bowfin lineage and/or newly duplicated in the gar lineage. Among the 16 gar SCPP genes for which no bowfin ortholog was identified, nine gar genes (*scpplpq9*, *scpplpq10*, *scpplpq18*, *scpplpq11*, *scpplpq19*, *scpplpq12*, *scpplpq13*, *scpplpq15*, and *scpplpq17*) were clustered on LG2 (syntenic to the SCPP gene cluster on bowfin 9; main text Figure 3). Notably, our previous RT-PCR study showed that expression of all these *Lepisosteus* SCPP genes with the exception of *scpplpq14* and *scpplpq16a* was weak or undetectable in tooth germs but strong in the skin that includes cells forming scales[9]. This result suggests that many of these SCPP genes clustering on gar LG2 are involved in scale formation, but not in tooth formation. In order to test this hypothesis, we examined P/Q-rich SCPP genes of three phylogenetically close teleost species, zebrafish, Mexican tetra (cavefish), and channel catfish, of which both zebrafish and cavefish have scales but channel catfish secondarily lost scales[62].

It was previously reported that *scpp5*, among P/Q-rich SCPP genes, is involved in scale formation, because *scpp5* is absent in the scale-less three-spine stickleback and became a pseudogene in the scale-less channel catfish, but is present in two scaled catfish species[62]. However, we identified apparently functional *scpp5* genes in both three-spine stickleback and channel catfish (Supplementary Fig. 11A). Moreover, our transcriptome analysis revealed no significant expression of *scpp5* in the skin of various ages of zebrafish (Supplementary Table 13). For these reasons, we concluded that *scpp5* is not essential for scale formation. Liu *et al*. (2016) further argue that *scpp1* is also critical to scale formation, because of its presence in the common pleco genome, a scaled catfish, and its absence in the scale-less channel catfish[62]. However, we also identified *scpp1* in the channel catfish genome (Supplementary Fig. 11B).

Instead, the presence and absence of SCPP genes in scaled zebrafish and cavefish and in scale-less channel catfish suggest that *scpp11a*, *scpp11b*, *scpp13*, *scpp14*, *gsp37*, and *scpp12* clustered on

zebrafish chromosomes 5 and 10, previously shown to be TGD duplicate clusters of the SCPP gene cluster on gar LG2[9], are potentially important for scale formation. This suggestion is consistent with our transcriptome analysis that confirmed expression of all these genes in zebrafish skin (Supplementary Table 13). Furthermore, *gsp37* is known to encode a scale matrix protein in goldfish[63]. The evolutionary rate of all these genes is extremely high[54], and it is thus difficult to determine orthologies even between the relatively closely related zebrafish and cavefish (*Astyanax*) and not possible to clearly identify orthologies of these SCPP genes between Holostei and teleosts. However, the presence of twelve gar genes encoding P/Q-rich SCPPs, clustered on LG2, and nine missing orthologs of these twelve gar SCPP genes on bowfin 9 agree with the hypothesis that genes encoding P/Q-rich SCPPs involved in scale formation are clustered on LG2, and that these genes were secondarily lost from the bowfin genome, along with modifications of scale formation in bowfin.

Supplementary Note 6. ATAC-Seq chromatin profiling of bowfin development

Results and Discussion. ATAC-Seq data are available in NCBI under SRA accession SRP281665. The inferred fragment size distribution (Supplementary Fig. 12) confirm high library quality and visualizes nucleosome periodicity with peaks occurring around 200, 400, and 500 bp representing mononucleosome, dinucleosome, and trinucleosome sizes, respectively[64]. We find a total of 163,771 open chromatin regions (OCRs) when merging bed files of all stages with 1bp+ overlap, 132,119 of which (81%) were retained as non-coding OCRs (ncOCRs) that are candidates for gene regulatory regions through development. OCR similarity between stages quantified with Jaccard indices show that OCR profiles for each stage are most similar to those in the adjacent developmental stage (Supplementary Fig. 13), illustrating gradual changes in chromatin dynamics and gene regulation through development. OCR and ncOCR results are summarized in Supplementary Tables 14-15. See Supplementary Table 16 for HOMER[65] annotation results of bowfin OCRs.

To explore overlap of bowfin OCRs with genetic annotations (OCRs; Ultra Conserved Elements, UCEs; enhancers; Conserved Non-Coding Elements, CNEs) from other vertebrate species' studies and databases, these elements were blasted (BLASTN e< 10^{-5}) against the hard repeat-masked bowfin genome. Specifically, we used CNEs in the gar genome[9], bowfin UCEs[66], experimentally confirmed enhancer elements from human and mouse available from the Vista Enhancer Browser[67], and mouse embryonic single nucleus OCRs[1] as our queries. Overlap between bowfin OCRs and blast hits for these queries extracted with BedTools[68] is summarized in Supplementary Tables 17-21. Results described in the main text and here are reported for overlaps of both 1bp+ overlap between elements and 33%+ length overlap for either element.

For both mouse VISTA enhancers (Supplementary Table 18) and embryonic single nucleus OCRs (Supplementary Tables 20-21), we find a low level of sequence connectivity to the bowfin genome. This is likely due to the particularly high rate of molecular sequence evolution in the mouse lineage[69], which would also explain the difference in BLAST mapping to bowfin of VISTA enhancers of mouse (16%) when compared to human (61%) (Supplementary Table 18). Furthermore, since the genome-wide mouse OCRs were defined by ATAC-Seq without taking sequence conservation into account[1], we expect a large fraction of them to be mouse-lineage specific. Nevertheless, we located the orthologous position for a total of 8,594 murine ncOCR elements in the bowfin genome, of which 2,137 (25%) reside in bowfin ncOCRs (Supplementary Tables 20-21). These bowfin ncOCRs are thus putatively cell/tissue type-specific, attributable to the development of e.g. the forebrain (153), neural crest (46), cardiomyocytes (21), etc. (Supplementary Table 21), and prime candidates for experimental validation with reporter assays in fish and tetrapods.

Supplementary Note 7. Bowfin informs the evolution of the *tbx4* 'lung' enhancer

7.1 Background. The evolutionary origin and homology of bony vertebrate air-filled organs (AOs) in form of e.g. tetrapod lungs, respiratory gas bladders as in bowfin and gar, and teleost swim bladders has been debated since the days of Owen and Darwin[70,71]. For example, it was proposed that an ancestral lung with a dual function as a respiratory as well as a hydrostatic organ evolved before the rise of bony vertebrates (but AOs are not found in cartilaginous fishes or cyclostomes). While the respiratory function was intensified in tetrapods in form of the lung, the hydrostatic function became more elaborate in teleosts in form of the swim bladder for buoyancy control[72]. Lungs in tetrapods and lungfishes as well as in the basal ray-finned lineage of polypterids (bichirs, reedfish) develop on the ventral side of the gut, while teleost swim bladders as well as the gas bladders of gars and bowfin develop on the gut's dorsal side. Therefore, it has also been suggested that these organs could have originated independently from a respiratory posterior pharynx in the bony vertebrate ancestor that then evolved into a dorsal respiratory gas bladder in ray-finned fishes and into ventral lungs in tetrapods. The respiratory function of the dorsal gas bladder would later be lost in many teleosts after divergence from holosteans[73].

Support for the homology of tetrapod lungs and teleost swim bladders comes from molecular and developmental studies that recognized similarities between transcriptional profiles of the adult zebrafish swim bladders and adult mammalian lungs[74] and the co-expression of genes during early zebrafish swim bladder development that also interact during tetrapod lung formation[71]. Most recently, some of us (EF and ARM) have shown that key genes of the Fgf, Bmp, and Tbx signaling pathways known to be expressed during ventral lung development in tetrapods are also expressed during dorsal gas bladder development in bowfin[75,76]. This suggests a ventral-to-dorsal inversion of the AO gene regulatory network in an ancestor of neopterygians fishes after divergence from the more basally branching polypterid lineage that express these genes during the formation of their ventral lungs[75-77].

One of the genes with a ventral-to-dorsal expression inversion in neopterygians is *tbx4*. Previously, it was found that an enhancer region located in the third intron of the *Tbx4* transcription factor gene drives *Tbx4* expression in the early lung bud in mouse, where it controls lung bud formation[78]. This 'lung mesenchyme enhancer' (LME) region shows conservation among tetrapods, but is also conserved in coelacanth[79,80], which possesses a fat-filled lung-like organ[81], as well as in bichir, where LME likely functions as a lung enhancer[77]. The LME is also conserved in gar[9,77], but no conservation has thus far been demonstrated with teleosts that possess swim bladders[77,79,80,82,83]. Other non-coding elements in the *tbx4* gene region conserved across teleosts but not with lobe-finned vertebrates, in contrast, were hypothesized to represent teleost-specific 'swim bladder enhancers'[79]. Here, we have analyzed the conservation of *Tbx4* non-coding regions across bony vertebrates including bowfin genome information.

7.2 Approach. The *tbx4* gene regions of bowfin (AmiCal1; Aca_scaf_22), gar (LepOcu1; LG22), stickleback (BROADS1; groupI), Japanese puffer (GCA_901000725.2; chr1), medaka (HdrR ASM223467v1; chr13), pike (Eluc_v4; LG01), zebrafish (GRCz11; chr15), coelacanth (LatCha1; JH126567), western clawed frog (Xenopus_tropicalis_v9.1; chr2), chicken (GRCg6a; chr19) and human (GRCh38; chr17) were aligned with mVISTA[84] [http://genome.lbl.gov/vista/mvista/submit.shtml] using Shuffle-Lagan[85] (calculation window: 75bp, minimum conservation width: 75bp; conservation identity: 65%). Repeat-masked sequences and gene annotations were downloaded from Ensembl [http://ensembl.org; Release 100].

7.3 Results and Discussion. In a human-centric VISTA plot (Supplementary Fig. 14, top), we find that conservation of the 'lung enhancer' extends to holosteans in general as a conservation peak is found within the third intron of bowfin and gar *tbx4*, providing evidence for the presence of the LME in the common ancestor of bony vertebrates.

Based on a stickleback-centric VISTA plot (Supplementary Fig. 14, bottom), we further find evidence for the presence of one of the three putative 'swim bladder' enhancers[79] (SBE2) in bowfin. This region is found in stickleback *tbx4* intron 3 in addition to two other regions conserved among neopterygians, but that appear to be not conserved with tetrapods. We name these two other apparently ray-finned-specific regions 'R1' and 'R2'.

In the bowfin-centric VISTA plot (Supplementary Fig. 14, middle; main text Fig. 4F), however, it becomes apparent that in intron 3 of bowfin *tbx4* the regions defined as LME lung enhancer by the human-centric alignment and region R1 defined by the stickleback-centric alignment in fact overlap in the bowfin genome. For the first time, this serial analysis of sequence conservation in intron3 of the *Tbx4* gene shows that at least the overlap of the LME and R1 elements has to be considered orthologous between tetrapods and teleosts, and this hidden orthology only becomes apparent by the inclusion of holostean genomes.

Bowfin *tbx4* is expressed in the developing gas bladder, starting with strong expression in the dorsal and ventral mesoderm during the budding phase (stage 25) and then increasing dorsal expression during the outgrowth phase (stage 27), generating a dorso-ventral gradient[75]. Importantly, our ATAC-Seq profile of *tbx4* shows that the LME region is indeed accessible during that time period (main text Fig. 4E), suggesting regulation of *tbx4* expression, and SBE2 is characterized by open chromatin as well (main text Fig. 4E).

This example highlights the power of the holostean genome bridge to connect tetrapod and teleosts non-coding regions (Supplementary Fig. 14, right). It further supports deep homology of *tbx4* gene regulation among bony vertebrates leading us to postulate that *cis*-regulatory elements embedded in the third intron of *tbx4* provide further molecular evidence for the homology of tetrapod lungs, holostean respiratory gas bladders, and teleost swim bladders. This hypothesis awaits functional examination of neopterygian intron 3 sequences. Intriguingly, LME is also conserved in some shark species[82,83] and cartilaginous fish may have vestigial AOs[73], which could suggest that the LME function even dates back to the gnathostome ancestor[83].

Supplementary Note 8. Hox gene cluster analyses

8.1 Approach. Bowfin Hox cluster genes were manually annotated using spotted gar *hox* gene orthologs, including the *hoxd14* pseudogene[9]. Manual curations were performed in Geneious 9.1.8 supported by AUGUSTUS[86] *ab initio* gene predictions on the Hox cluster regions and following annotation strategies used for other fish Hox gene clusters[87,88]. For the annotation of the *hoxd14* gene structure, we also leveraged information from aligned fin transcriptome RNA-seq reads and cloned and sequenced *hoxd14* transcripts as described below in Supplementary Note 9.

8.2 Results. Predicted bowfin *hox* gene transcripts are provided in fasta format in Supplementary File 2. The presence of the bowfin hoxd14 pseudogene is supported by fin bud RNA-seq and whole embryo ATAC-Seq data (Supplementary Fig. 15). We cloned two different partial splice variants of bowfin *hoxd14* as shown below (alternating color boxes refer to exons in main text Fig. 5B; grey shaded sequence indicates difference between the long and short variant).

>Aca_HoxD14_long

CCTTCGTCTCTGACCTCTGGATGTCTGAAAACCGCTGCGGACAAGAGGTTTGGGGATTTTCGTGATGTTTTA
GCGTATATGGTTACTCGAAAAAGGAAGTGGATGCATCTCTGCCCCAGGCTCCTACAAACTTTAAAACATATG
CGTGCTGCAGACAGTACTATATATCTTACAGTAATATATCTTCTTTATCTAGACCAAACACCTATAATCGGAG
CGAATACCCTCCACGTCCTTCCATATATCATCAGTGGGCGTTTGAACACCCTCGGGGGTCTTCTGCCGTCTC
AGACCAG CTTTAGGATCCCCTCGATATGGGTCTCAAAGTACACCTGCGAAGTGAGAAGAAAAGGACCCATT

AATAGCAAATAGCAGATCGCTGGATTGTAACTGGAATTTGAGAGGAATAATTTTCTCACTCCCAAAAATTTCA
TTAAAACAGAAAGACAG GTTAAACTGTGGTTTCAGAACCAGAGACTGAAACAGAAAAAGTTTCTTCGAGTTC
CACTCACAAAACCCAGTTTAATTTGAGCTGCACTTATGTTTTTCAAGAAATGCTTTTAGGCGGTACTATTTTCC
CAGTGTATAAATGCGTCTTCGAAATTGATAAAGTTCGTATGTGTCGCGATTAACATGTTTTAAATGAGAAACA
TCTGCATGTAAGCGTACATTTGAATGTGCCAATTGCCTCTT

>Aca_HoxD14_short
CCTTCGTCTCTGACCTCTGGATGTCTGAAAACCGCTGCGGACAAGAG ACCAAACACCTATAATCGGAGCG
AATACCCTCCACGTCCTTCCATATATCATCAGTGGGCGTTTGAACACCCTCGGGGGTCTTCTGCCGTCTCAG
ACCAG CTTTAGGATCCCCTCGATATGGGTCTCAAAGTACACCTGCGAAGTGAGAAGAAAAGGACCCATTAA
TAGCAAATAGCAGATCGCTGGATTGTAACTGGAATTTGAGAGGAATAATTTTCTCACTCCCAAAAATTTCATT
AAAACAGAAAGACAG GTTAAACTGTGGTTTCAGAACCAGAGACTGAAACAGAAAAAGTTTCTTCGAGTTCC
ACTCACAAAACCCAGTTTAATTTGAGCTGCACTTATGTTTTTCAAGAAATGCTTTTAGGCGGTACTATTTTCCC
AGTGTATAAATGCGTCTTCGAAATTGATAAAGTTCGTATGTGTCGCGATTAACATGTTTTAAATGAGAAACAT
CTGCATGTAAGCGTACATTTGAATGTGCCAATTGCCTCTT

Supplementary Note 9. Fin development analyses

9.1 Collection of bowfin embryos. Bowfin embryos were collected from nests in Oneida Lake, New York. Eggs attached to nest material were collected in lake water and methylene blue was added to abate fungal growth. Once in the lab, eggs were separated from nesting material by hand and placed in fresh lake water. Eggs and embryos were raised in static containers of lake water and moved to fresh water every other day. Embryos and larvae were sampled at the relevant stages following Ballard's bowfin staging series[89] (Supplementary Fig. 16).

9.2 cDNA cloning. Nested PCR was used to amplify cDNAs of target genes. *fgf8* and *sp8* were amplified from a cDNA library pooled from whole-embryo stage 23 and stage 25 libraries. *hoxD14* was amplified from stage 26 fin bud RNA. Primers and target sizes are listed below (Ac: *Amia calva*; Lo: *Lepisosteus oculatus*).

AcFgf8_F1: 5'-ACCATTCAGTCCTCGCCTAA-3'
AcFgf8_R1: 5'-GGGCAGACGCTTCATAAAAT-3'

➔ AcFgf8_F1 + AcFgf8_R1 = 465 bp product (for ISH)

LoFgf8_F1: 5'-ACCATTCAGTCCTCGCCTAA-3'
LoFgf8_R1: 5'-TTGTTCAAAAGGTGAAATTCCT-3'
LoFgf8_R2: 5'-GGCAATTCTCATGGACGTTT-3'

➔ LoFgf8_F1 + LoFgf8_R1 = 414 bp product
➔ LoFgf8_F1 + LoFgf8_R2 = 350 bp product (for ISH)

AcHoxD14_F1: 5'-CCTTCGTCTCTGACCTCTGG-3'
AcHoxD14_R1: 5'-AAGAGGCAATTGGCACATTC-3"

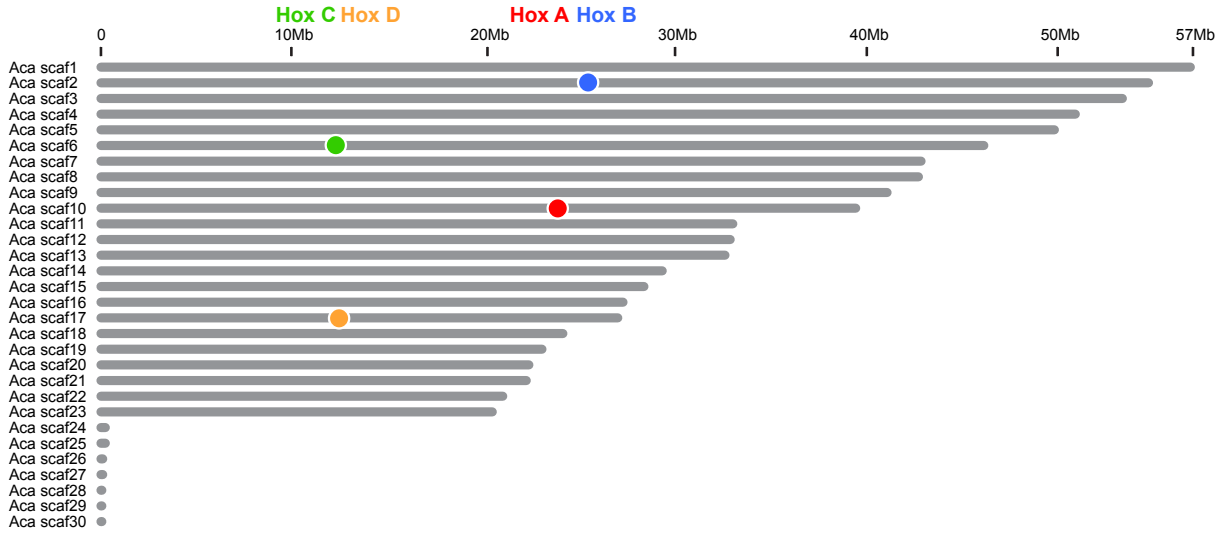
➔ AcHoxD14_F1 + AcHoxD14_R1 = 545 bp product (for ISH)

AcSp8_F1: 5'-GGAAGAGCCGAGGTTAGGAT-3'
AcSp8_R1: 5'-GCTGAGGAGGTGTGGAGAAG-3'
AcSp8_R2: 5'-GAGTGACCCAAACCGGAGTA-3'

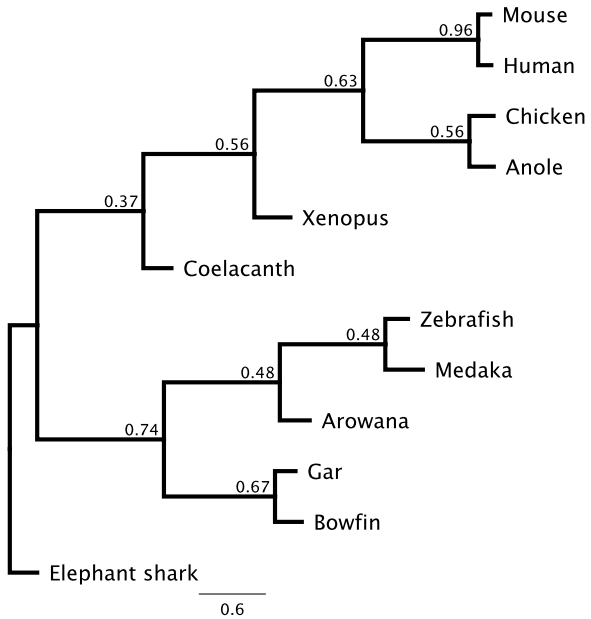
➔ AcSp8_F1 + AcSp8_R1 = 697 bp product
➔ AcSp8_F1 + AcSp8_R2 = 660 bp product in Bowfin, 330 bp product in Gar (for ISH)

9.3 Genomic comparison of the *fgf8* gene region. The *fgf8* gene regions of bowfin (AmiCal1; Aca_scaf_19), zebrafish (GRCz11; *fgf8a* chr13, fgf8b chr1), stickleback (BROADS1; *fgf8a* groupVI, *fgf8b* group IX), and mouse (GRCm38; chr19) were aligned against the spotted gar *fgf8* region (LepOcu1; LG5), spanning from *btrc* upstream and *poll* downstream of *fgf8*. Repeat-masked sequences and gene annotations were downloaded from Ensembl [Release 100] and aligned with mVISTA[84] using Shuffle-Lagan[85] (calculation window: 100bp, minimum conservation width: 100bp; conservation identity: 70%). Conserved non-coding regions in the VISTA plot (Supplementary Fig. 17) were compared against known *fgf8* region gene regulatory elements previously identified in zebrafish and mouse[90-93]. Genomic locations of these elements in bowfin and gar along with information on overlap with open chromatin regions in bowfin embryo stages are given in Supplementary Table 22.

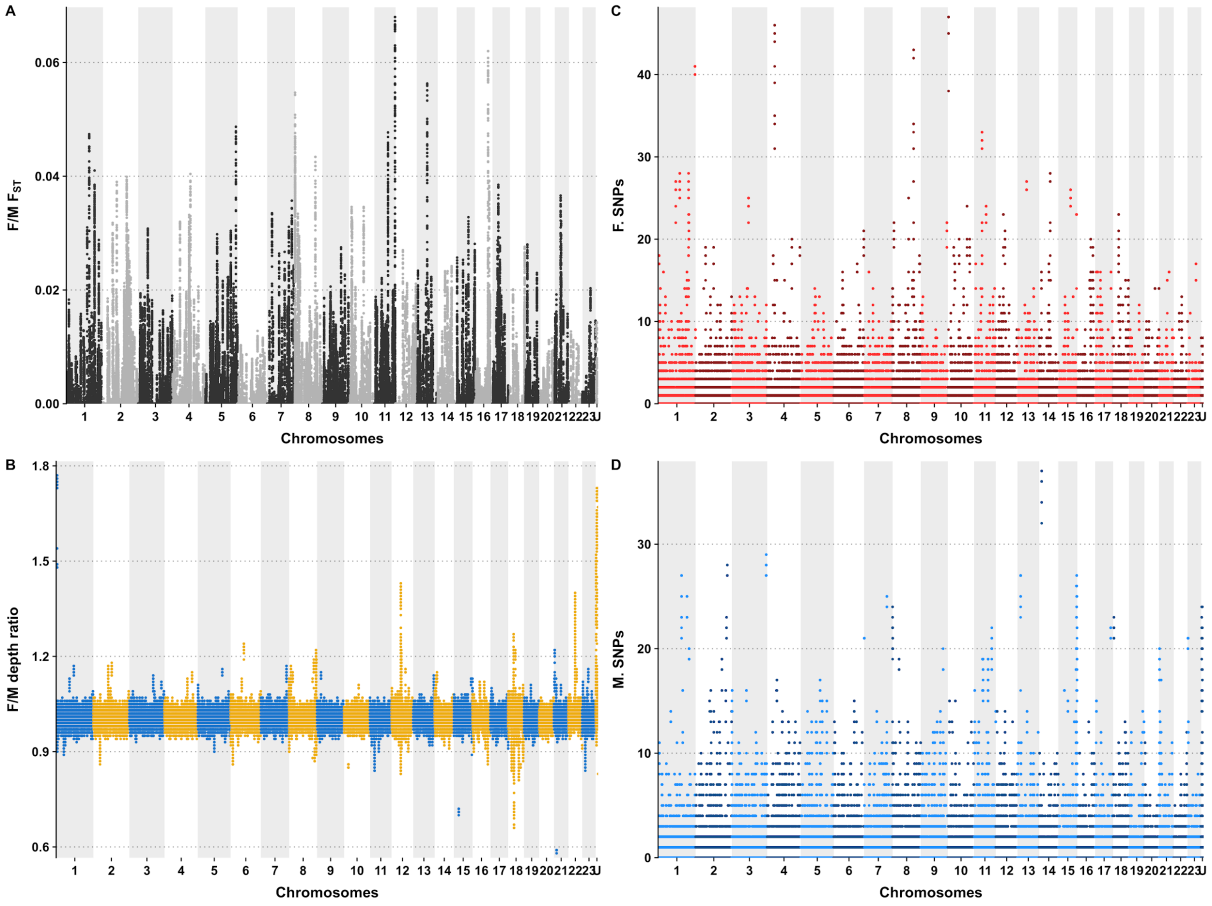
Supplementary Figures



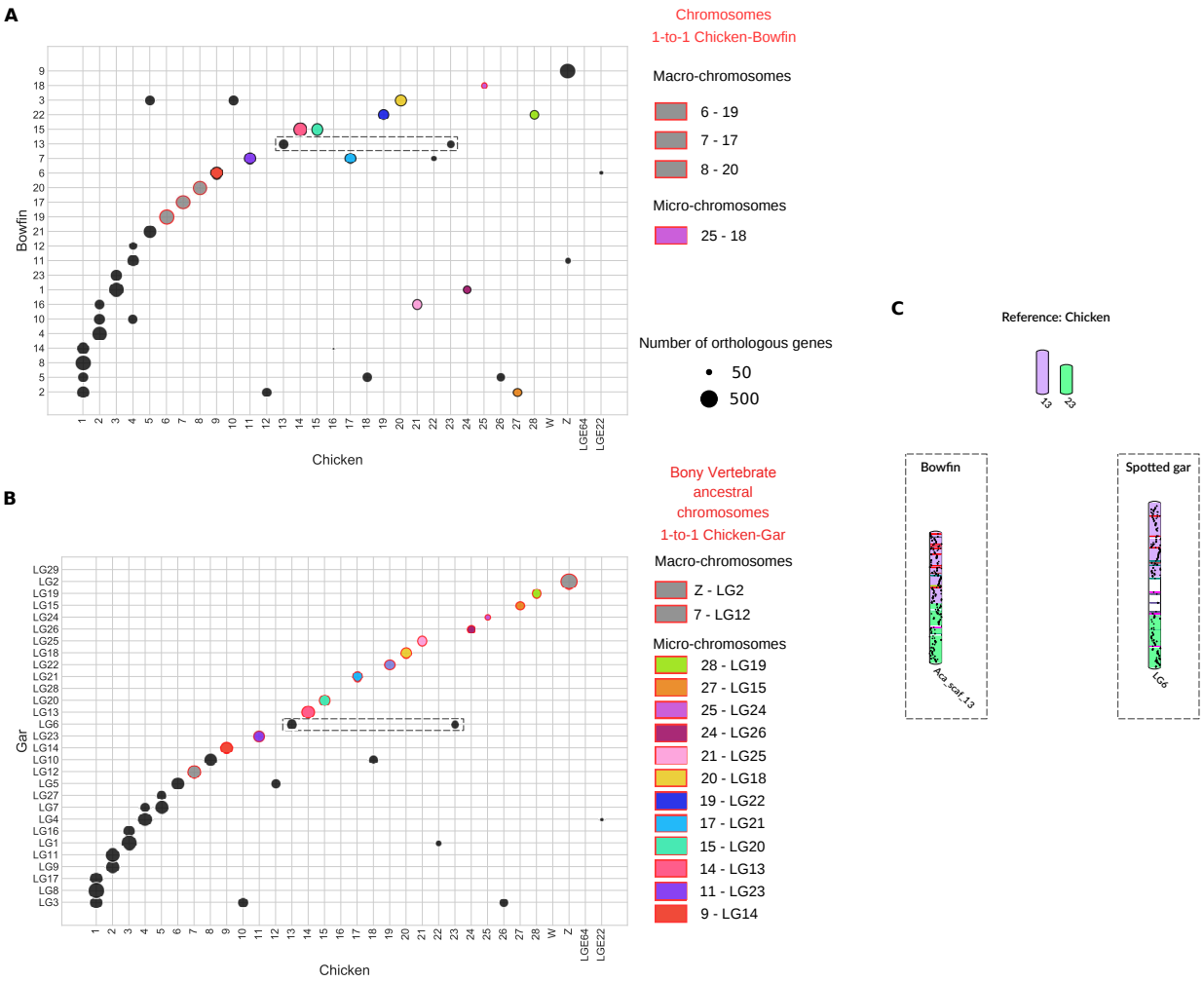
Supplementary Fig. 1. Size distribution of the 30 largest scaffolds in the AmiCal1 assembly. Superscaffolds Aca scaf1 to Aca scaf23 match the number of chromosomes in the bowfin genome[5,6]. Locations of the four bowfin Hox gene clusters A-D are indicated by colored circles.



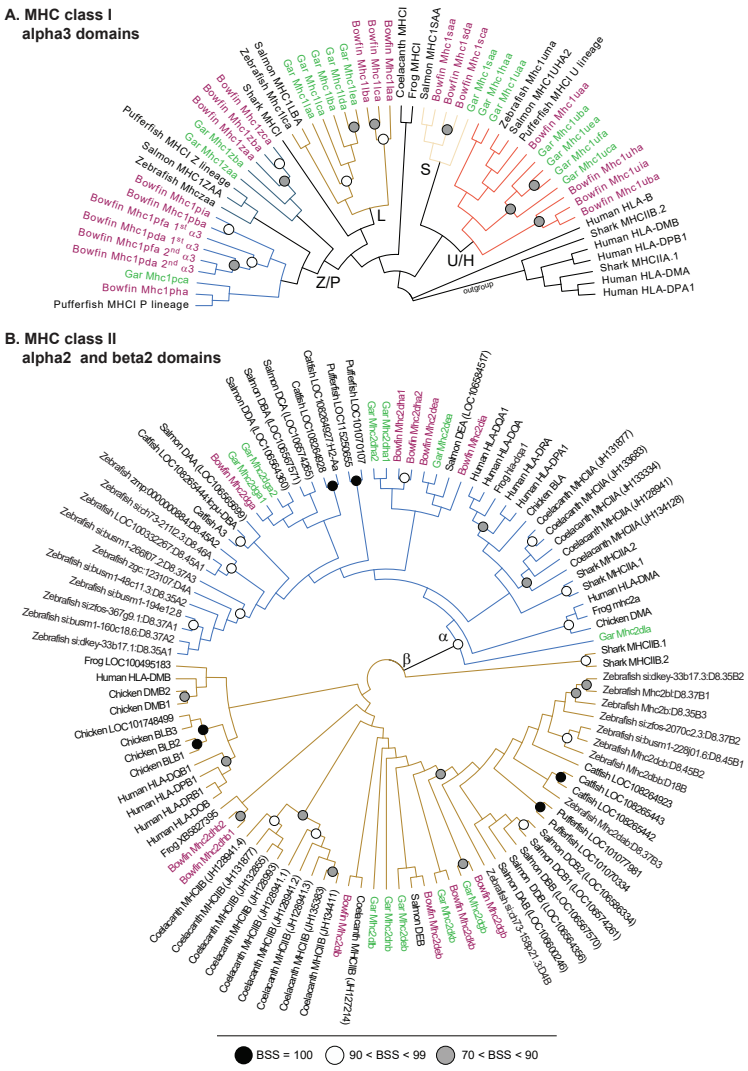
Supplementary Fig. 2. OrthoFinder species tree. Node support values are the proportion of individual estimates of the species tree that contain that bipartition. Branch lengths are the average lengths for each bipartition in the individual estimates of the species tree. The tree supports the monophyly of holosteans (gar + bowfin).



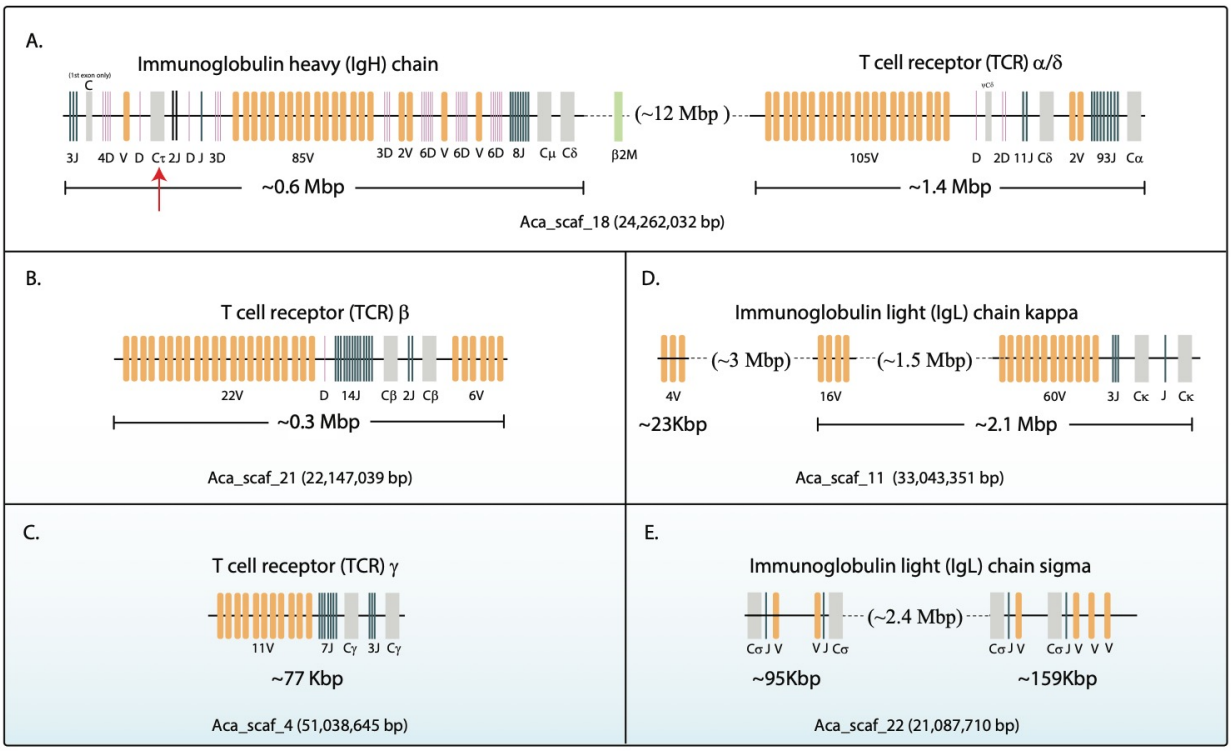
Supplementary Fig. 3. Pool-sequencing of male vs. female bowfin. A) Female/male F_{ST}; B) female/male sequencing depth ratio; C) female-specific SNPs; D) male-specific SNPs along the bowfin superscaffolds (chromosomes). No clear sex-specific region in bowfin was detected.



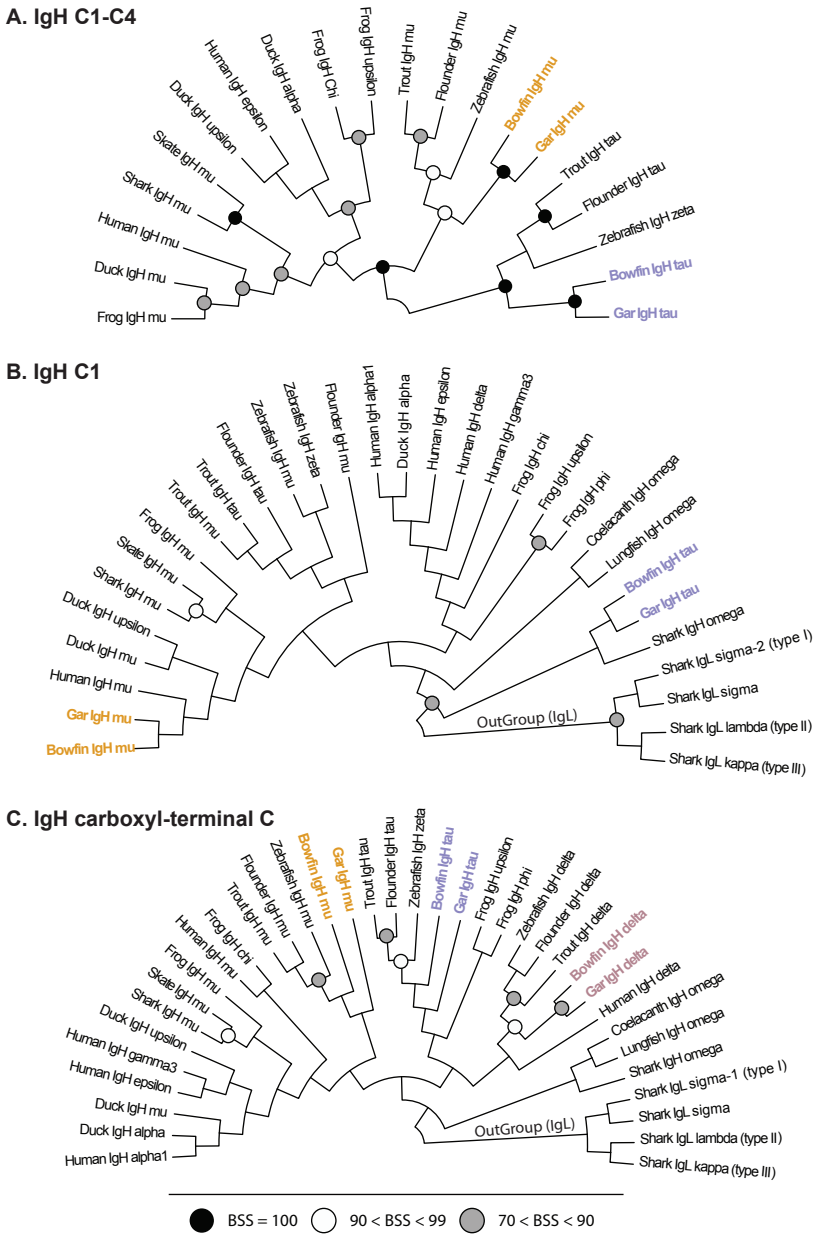
Supplementary Fig. 4. Orthology of holostean and chicken karyotypes. A) Orthologies between the bowfin and chicken chromosomes. B) Orthologies between the gar and chicken chromosomes. Circles represent the number of orthologous genes shared between pairs of chromosomes (if in excess compared to random expectations). One-to-one orthologous chromosome pairs are circled in red and summarized on the right. The dotted rectangle highlights a potential shared fusion in bowfin and gar. C) Putative holostean-specific chromosome fusion. Bowfin chromosome 13 and spotted gar LG6 are painted according to chicken chromosomes 13 and 23.



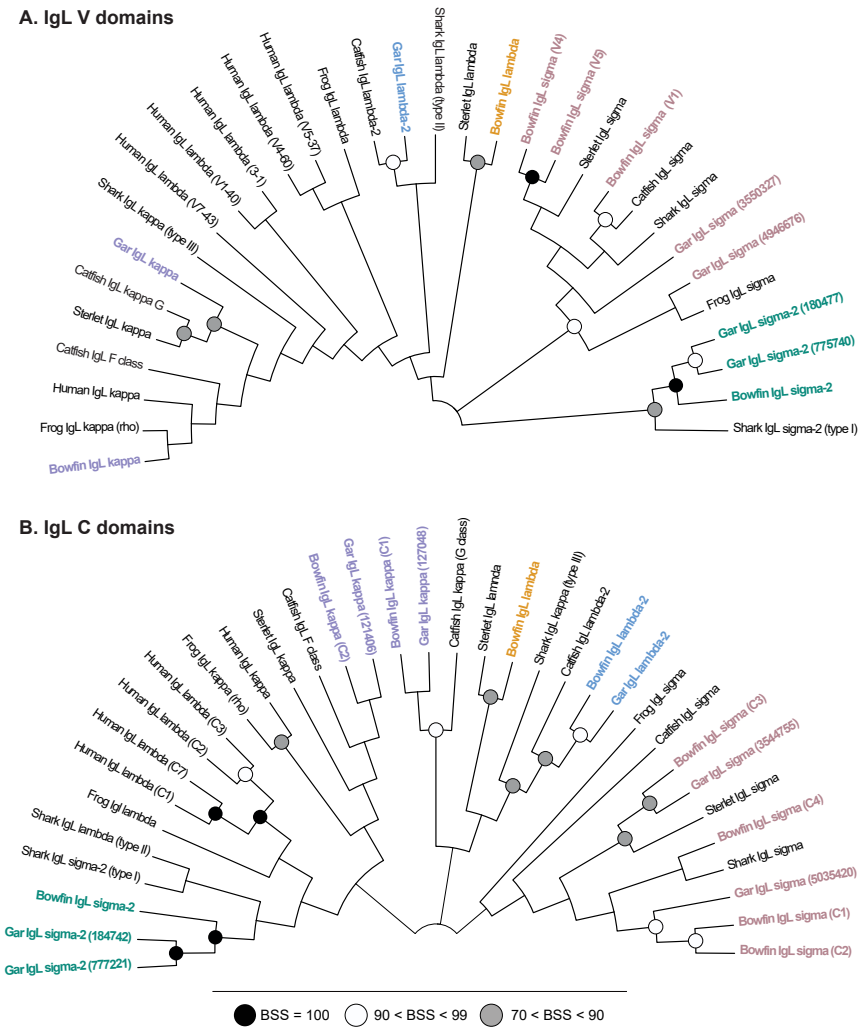
Supplementary Fig. 5. Cladograms depicting the evolutionary relationships among bowfin and representative MHC Class I and Class II sequences from other vertebrates. A) Bowfin MHC class I alpha 3 domains were compared to those of other species. Class I lineages are color coded: U/H = red, S = yellow, Z = navy blue, P = blue, L = gold. Select class II sequences are included as an outgroup. Note that of all H lineages evaluated, only the spotted gar possesses an alpha3 domain and could be included in this analysis. A bowfin *mhc1pda* transcript encodes two alpha 3 domains which are indicated as Mhc1pda.1 and Mhc1pda.2. Two bowfin Mhc1pfa alpha 3 domains (scaf12:9,769,801-9,769,529 and scaf12:9,764,651-9,764,379) are also included although they may represent a pseudogene. B) Bowfin MHC class II alpha 2 and beta 2 domains were compared to those of other species. Class IIa and IIb sequences are indicated by blue and gold branches, respectively. The holostean DG lineage may be orthologous to teleost DA/DB lineages which include salmon DA/DB/DC/DE/DF genes. Bowfin and gar sequences are indicated with purple and green text, respectively. Sequence and species identifiers are listed in Supplementary Tables 6-8. BSS = % bootstrap support. Best-fit substitution model: A) WAG+G and B) JTT+G.



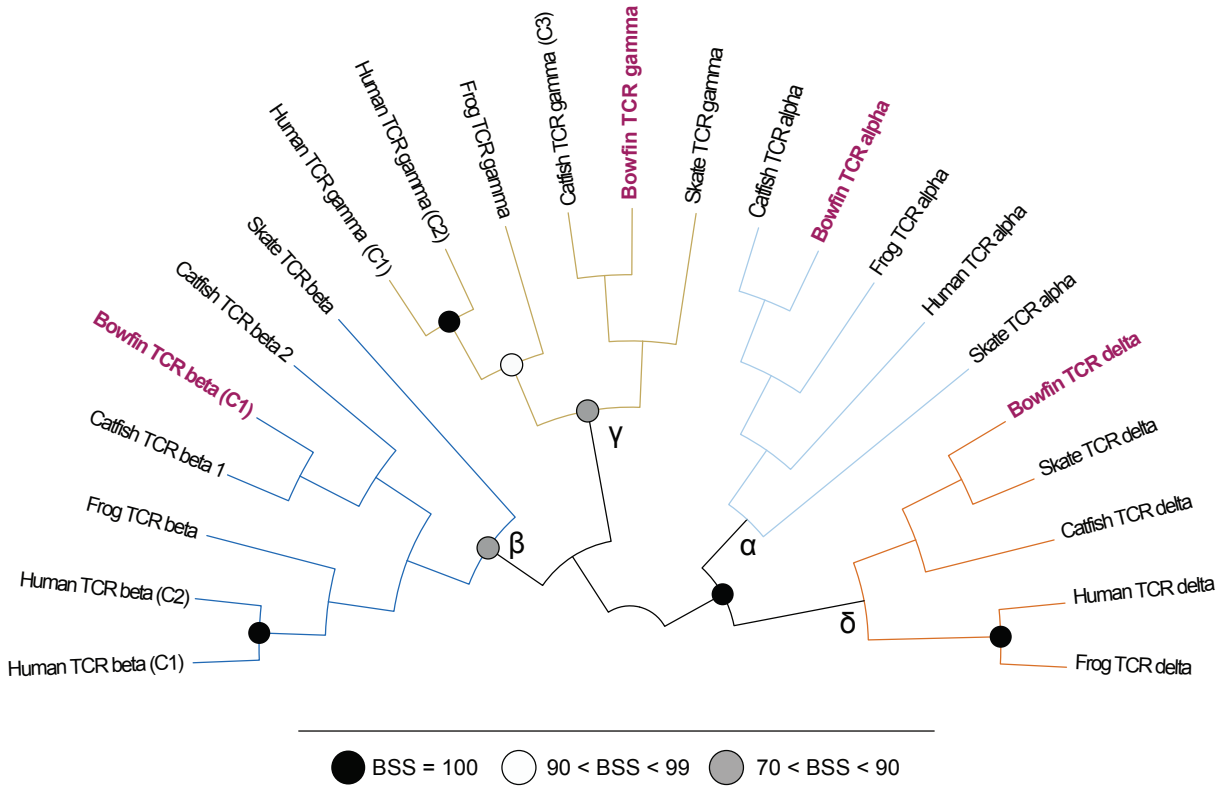
Supplementary Fig. 6. Genomic organization of bowfin immunoglobulin and TCR genes. A) The immunoglobulin heavy (IgH) chain and T cell receptor (TCR) α/δ loci are separated by 12 Mbp on Aca_scaf_18. The IgH locus encodes 3 classes of Ig constant domains, Cτ, Cμ and Cδ. The red arrow indicates Cτ, which was previously thought to be teleost-specific. Note that the IgL sigma-2 locus is present in the middle of V gene segment cluster of the IgH locus (not shown). B) The TCRβ locus is encoded on Aca_scaf_21. C) The TCRγ locus is located on Aca_scaf_4. D) The Ig light (IgL) chain kappa locus is present on Aca_scaf_11. E) The IgL chain sigma loci are encoded on Aca_scaf_22. Multiple genes (including pseudogenes) other than Ig and TCR are present in these scaffolds, but not shown. V domains are shaded orange, D segments purple, J segments blue, and constant (C) domains gray.



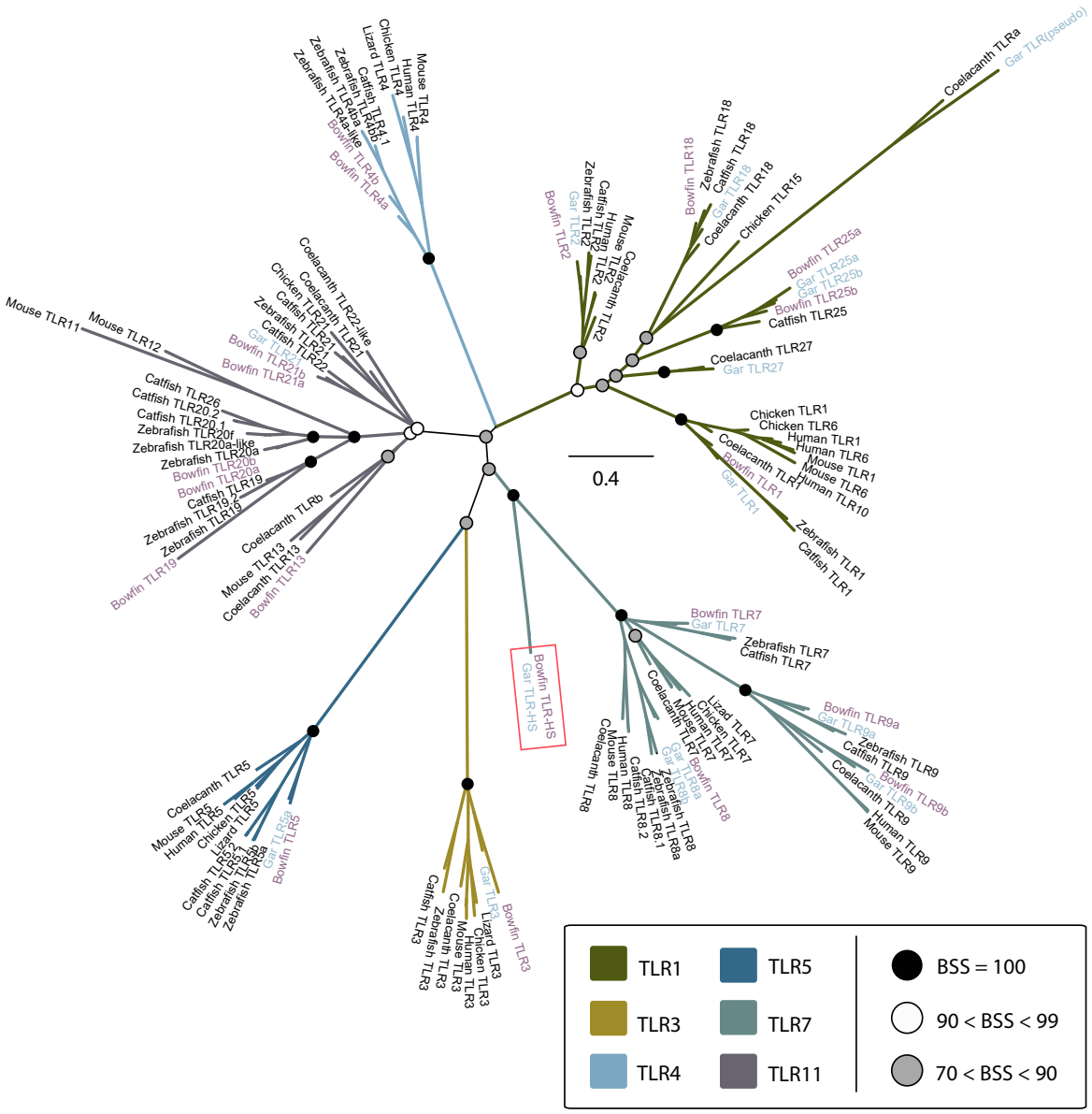
Supplementary Fig. 7. Cladograms depicting the evolutionary relationships among bowfin and diverse IgH sequences. A) IgH sequences that include four constant Ig domains (C1-C2-C3-C4) were compared. IgH sequences not consisting of four constant Ig domains, such as IgHδ (delta), were excluded from this analysis due to complications associated with domain duplications. B) The sequences of the first constant domain (C1) of IgH which forms disulfide bonds with IgL chain were compared. Since the IgD heavy chain of teleosts (and most likely of holosteans) utilize the first domain of IgHμ (mu) for the first domain of IgHδ (delta) peptides, the IgD sequences from teleosts and holostei were not included in this analysis. C) The carboxyl-terminal constant domain of IgH sequences were compared. Sequence and species identifiers are listed in Supplementary Table 9. Holostean IgHμ (mu), IgHτ (tau) and IgHδ (delta) sequences are in yellow, purple, and red text, respectively. Note that shark IgHμ (mu) and IgL (outgroup) sequences are from horn shark, whereas the shark Igω (omega) sequence is from sandbar shark. BSS = % bootstrap support (200 replicates). Best fit substitution models: WAG+G+I model for A) and WAG+G model for B) and C).



Supplementary Fig. 8. Cladograms depicting the evolutionary relationships among bowfin and diverse IgL sequences. A) IgL variable (V) domains from a range of species. B) IgL constant (C) domains. Sequence and species identifiers are listed in Supplementary Table 10. Following Guselnikov et al. (2018) [42], bowfin and gar sequences representing the five light chain isotypes are indicated by: kappa = purple, lambda = orange, lambda-2 = blue, sigma = red, sigma-2 = green. Note that bowfin IgL lambda and lambda-2 sequences likely reflect pseudogenes. BSS = % bootstrap support (200 replicates). Best fit substitution model: WAG+G+I.



Supplementary Fig. 9. Cladogram depicting the evolutionary relationships among bowfin and diverse T cell receptors (TCRs). TCR constant domains were compared to sequences from other vertebrate lineages. Sequence and species identifiers are listed in Supplementary Table 11. TCRα (alpha), TCRβ (beta), TCRγ (gamma) and TCRδ (delta) are indicated with light blue, dark blue, yellow, and orange branches, respectively. Bowfin sequences are indicated with purple text. BSS = % bootstrap support (200 replicates). Best fit substitution model: LG+G+I.



Supplementary Fig. 10. **Evolutionary relationships among vertebrate TLRs.** TIR domains identified in bowfin TLR sequences were compared to TLR sequences from a range of vertebrate species. Bowfin TLRs were annotated based on these relationships. Sequence and species identifiers are listed in Supplementary Table 12. Bowfin and spotted gar have a holostean-specific TLR-HS (red box). Coelacanth sequences were reported previously and include possible pseudogenes ENSLACG00000006376 (TLRa) and ENSLACG00000004773 (TLRb)[94]. Branches are color coded based on the six major families of TLRs. BSS = % bootstrap support.

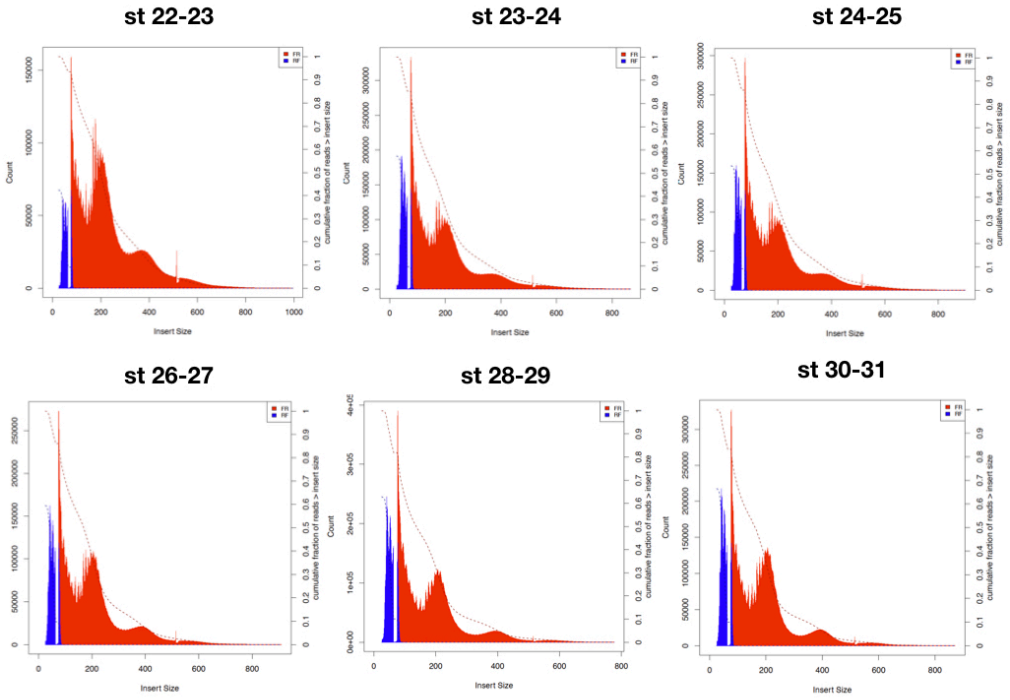
A Alignment of amino acid sequences encoded by *scpp5* genes

```
                exon2              exon3                  exon4
Gasterosteus    MKLVIFCLCLASTACAAP S--IFHYLPHYAGSRQQVPPSQ VGNPFTAG-QSLPPPGAAGAYSVEL
Danio           MWTSLLCLLLAGAVSAAP LSPFFNYLPHYGSPR------Q GNTGGFQGMPSQPHPAMNAPISMEI
Astyanax        MWSSILCLSFVTMVSSAP ISPFFNYLPHYGNPRP-GPSAQ GTNDFFS---TNGQPAFNAPISMEI
Ictalurus       MWSAVFCFSIISAVSAAP LYSF---LQHYGNPMQSGPSNQ AANDMFS--PLHPHTGMTTPISMEI

                exon5
Gasterosteus    IYPHRVAGGVGGTNAG-Q ---------------------- ------------------
Danio           IFPPRFPANPAGGAAGTS ---------------------- ------------------
Astyanax        VFPPQFQGSPVGGAGAGP ---------------------- ------------------
Ictalurus       LLPPRFPGSAGGQGSG-P GNSMFPGLPSHLQPGVNTPISIE LFHPGFQGTAAGGQGSRT

                exon6                  exon7      exon8            exon9
Gasterosteus    SF---GFIKYSIPQPPGRQSVEV YYPYDFSQQR- IMTNLPPMTNSPQMPN VFPFEYPPQNIPQQ-IPN
Danio           SFPTQAFIKYSLPKAPGRKSVEI FYPYDFGRA-Q DQP---NVPLIPQLPN IFPFDLMPQTVPQQ-PPV
Astyanax        AFPSQGFIKYSLPKVPGRKSVEI YYPYDFTQQRQ MMPSVPMMPAVPQLPG LVTFDNPPHNVPQQQPPR
Ictalurus       -----GLIKYSIPKAPGRKSIEI YYPYNFAQG-E VLP-----NILPQIPS IFPFNYLPQTPPQQQPPR

                exon10                                                 exon11
Gasterosteus    IPSFNPNALPSQDP-------MQPLQQDQPIQTSQ ---------------------- MPAKV
Danio           NPPFQDGAPQTQEPQQQTQPEQQQQQQQQQAQTGQ ---------------------- VSTRP
Astyanax        -----ASPPQSNDP-------HPQIMHDQPVQTGQ ---------------------- TPINL
Ictalurus       ----QAAPSQGNDQPFVF---SYPPQSGPQQQTPR AAPQANDPQQQIQQDPQVPAGQ P----
```

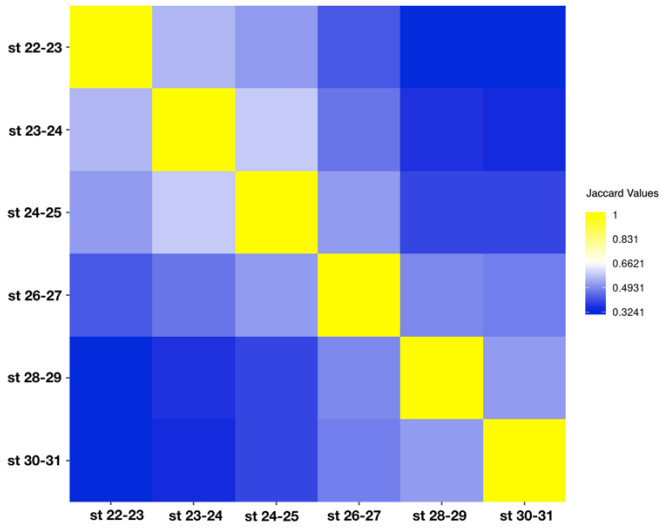
B Alignment of amino acid sequences encoded by *scpp1* genes

```
                exon2              exon3            exon4
Gasterosteus    MMVNIIMFCLMGAVFSNP ISFNAVLESNELDSNS TEN--LSSQSSENGTSAL--L
Danio           MKSALLILCLLGAACANP ILHKVAMEMIQHASNS SESSSISESSDQSNTSEPSEE
Astyanax        MKLTVVILCLLGATAANP ILHKVSMEIIDHASNS T--SSVSESSEESNTSE--HD
Ictalurus       MKLAFVILCLLGAAGANP ILHTDMMET---ASNS SQTSSMSASTEE--TVAIDQD

                exon5      exon6                  exon7
Gasterosteus    SSEESTSK-Q SQSSESSSESESESTSTESTSEDR KQTEEESNSLLDEKDV
Danio           KSEENVSDSN SSES------------LESESDEP ISKESESHSV---ESL
Astyanax        SSPENTSENI SSES------------DESKSDEQ TSDLSHSHSL---EER
Ictalurus       SSQENTSEDT TSES------------MESNSSEK TLETSQSNSL---EER

                exon8                 exon9
Gasterosteus    ADPM--AETTDNSMGSEENIRK SEGAAVSTNDSSES-----------------------
Danio           IGKSETALTADNTQSSKENIRR GWIYTLKWVQP--------------NNNIVQPTGQPH
Astyanax        FGTGEPGMTTDNSQGSQENMRK NWVHLINVKMASKED----------TEEVTDQPDEED
Ictalurus       FGNGEAGMTVDNSQGSTEIMRK NWIHVFSPQDISSEDNSTSASLALASSEISKSMESPE

                exon9 (continued)
Gasterosteus    -------------------------------TSSESNETSETSETS--------------
Danio           EN-------------------DIQSTSDASQISDSSESSESQEKVVVNKMQVDNSDDTS
Astyanax        KD----------------------------TTEDQTSESSESTEKPTPSSSSSSSTEDS-
Ictalurus       KNSKSISSSSESSESTEGQGNNSTSSSSESSESSESSESSESTENPEKNSNSSSSESSES

                exon9 (continued)
Gasterosteus    ------------------------------------------------------DSSDSSD
Danio           -------------------------------------ESAEHNGVVATEYSNSNSSSSSS
Astyanax        ---------------------------------------RAVVVDSSENSHSNSSSSS
Ictalurus       SESSESSESQEEKSSSSSESSESNENQQNASDSSSESKSVENRSTIDSDENSALKSSNSS

                exon9 (continued)
Gasterosteus    SSTSVEESEASDDALLGQLETKDCVNG--TQSCESE--EYLFQDIGDD-AHYSVDSLMVP
Danio           SSESTESKDSGTDSESR---STECVPGDDSQDCESE--ENLPQDIGDDGATDPFNGFLMP
Astyanax        SSESHESSHSTESTESQ---SKECPPGTDSNECDSD--EYQFHDVGDDGATDPFNGFHTT
Ictalurus       SSESHESTETTESTESKQSRSNECQPGADSQDCDSDSDEYVLQNVGDDGTNDPFDGFHVP

                exon9 (continued)
Gasterosteus    DEDEREFSLRR
Danio           DVAEP------
Astyanax        DNAGHEFAFKR
Ictalurus       DSTEREVTFKR
```

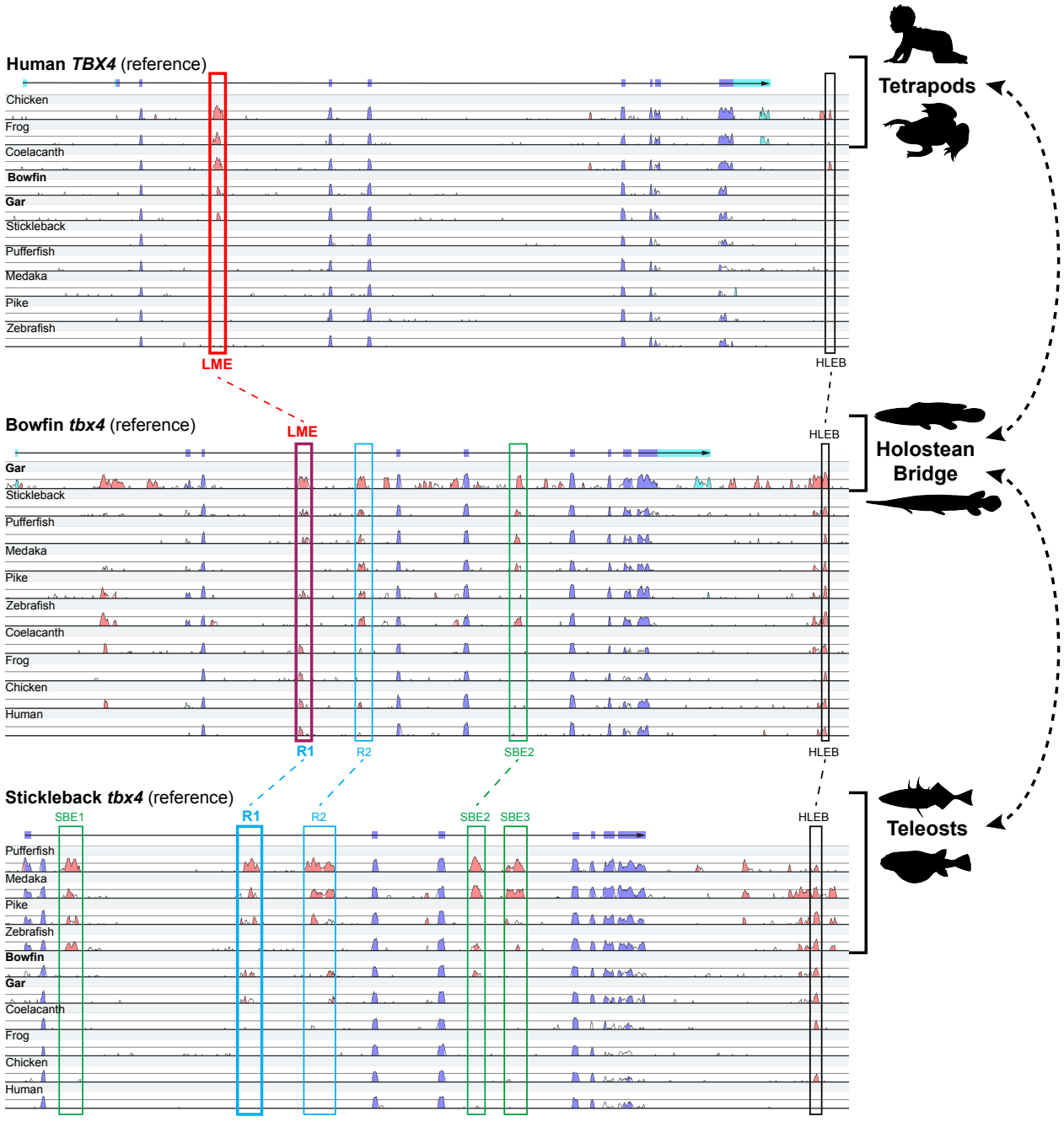
Supplementary Fig. 11. Teleost *scpp5* and *scpp1* genes. Alignment of three-spine stickleback (*Gasterosteus*), zebrafish (*Danio*), Mexican tetra (*Astyanax*), and channel catfish (*Ictalurus*) support the presence of A) *scpp5* in the stickleback (groupIX: 8,859,633-8,861,237) and channel catfish genomes (NC_030444.1 14,455,199-14,461,226) and B) *scpp1* in channel catfish (NC_030418.1: 4,433,039-4,437,245).



Supplementary Fig. 12. Nucleosome periodicity in ATAC-Seq data based on insert sizes of mapped reads.



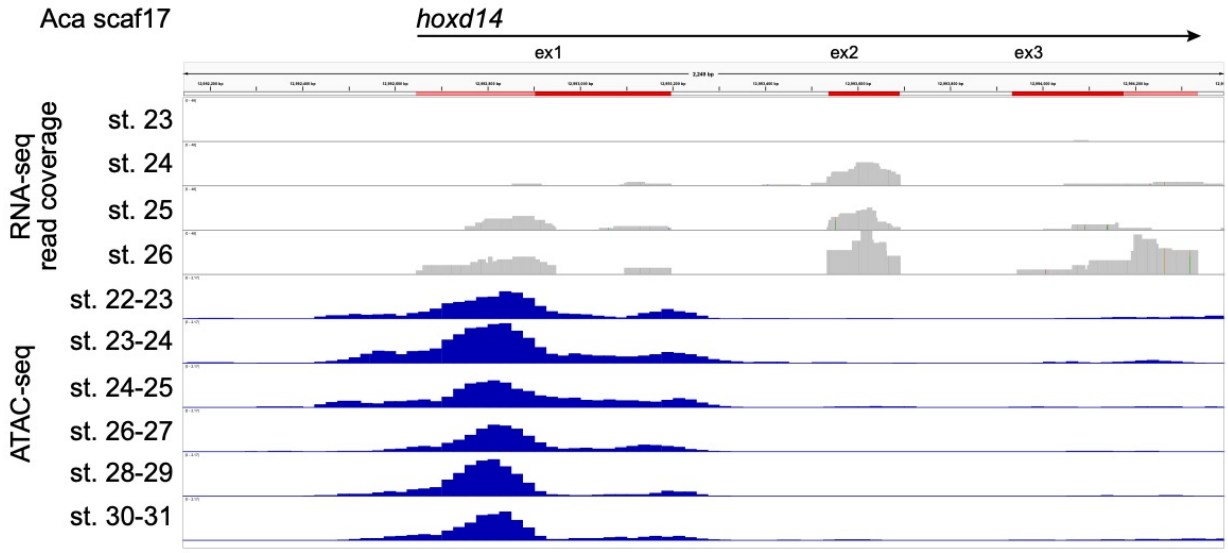
Supplementary Fig. 13. Heatmap of pairwise Jaccard distances among OCR profiles of bowfin developmental stages.



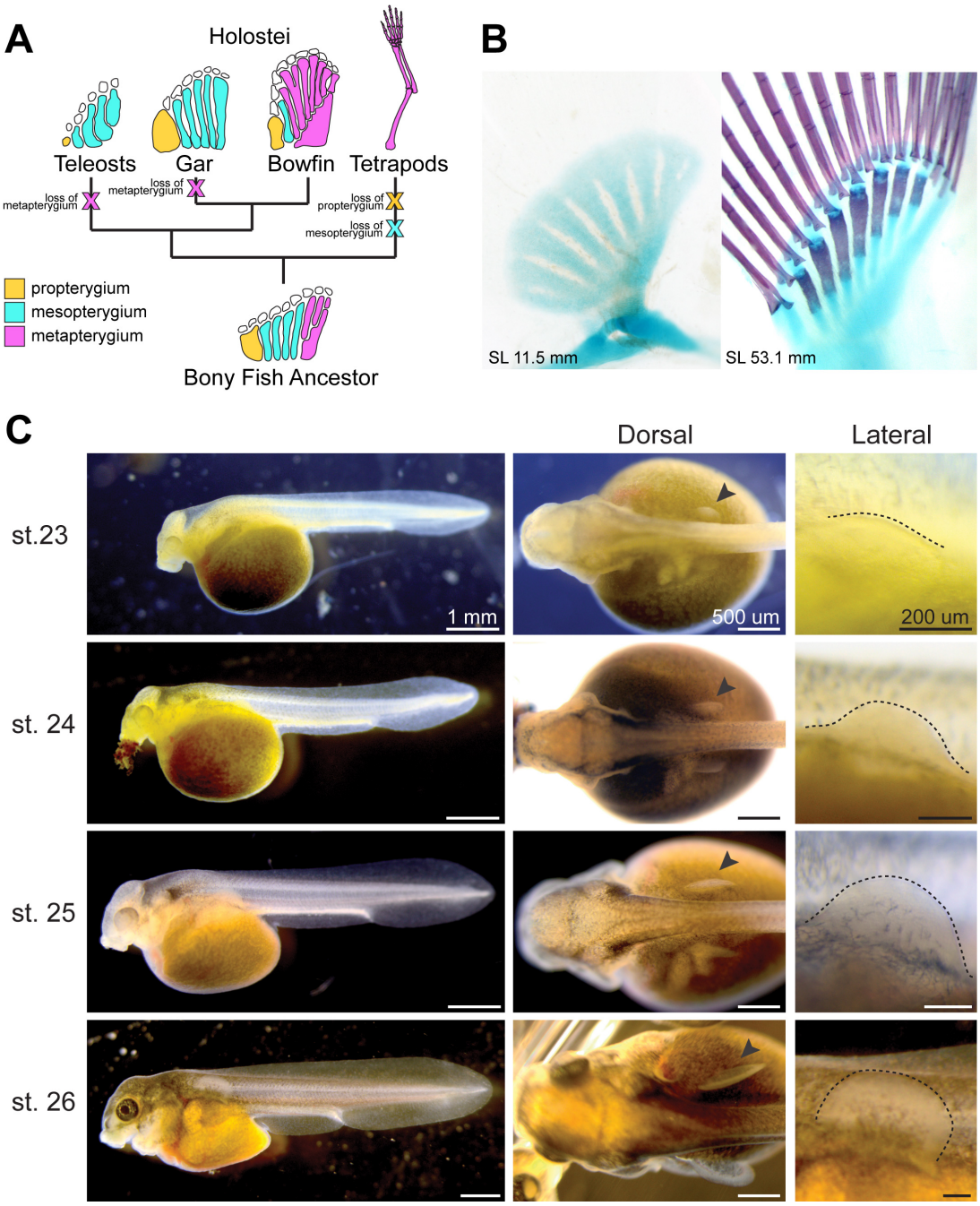
Supplementary Fig. 14. The holostean bridge connects vertebrate *Tbx4* gene enhancers. VISTA SLAGAN alignments were generated with human (top), bowfin (middle), and stickleback (bottom) as reference sequences.

Top: In the VISTA plot of the human *TBX4* gene, the intron 3 'lung mesenchyme enhancer' (LME, red box) that drives expression in the developing mammalian lung bud[78,80] shows conservation with other tetrapods, coelacanth, bowfin, and gar (and also with bichir[77] and some sharks[82], not shown).

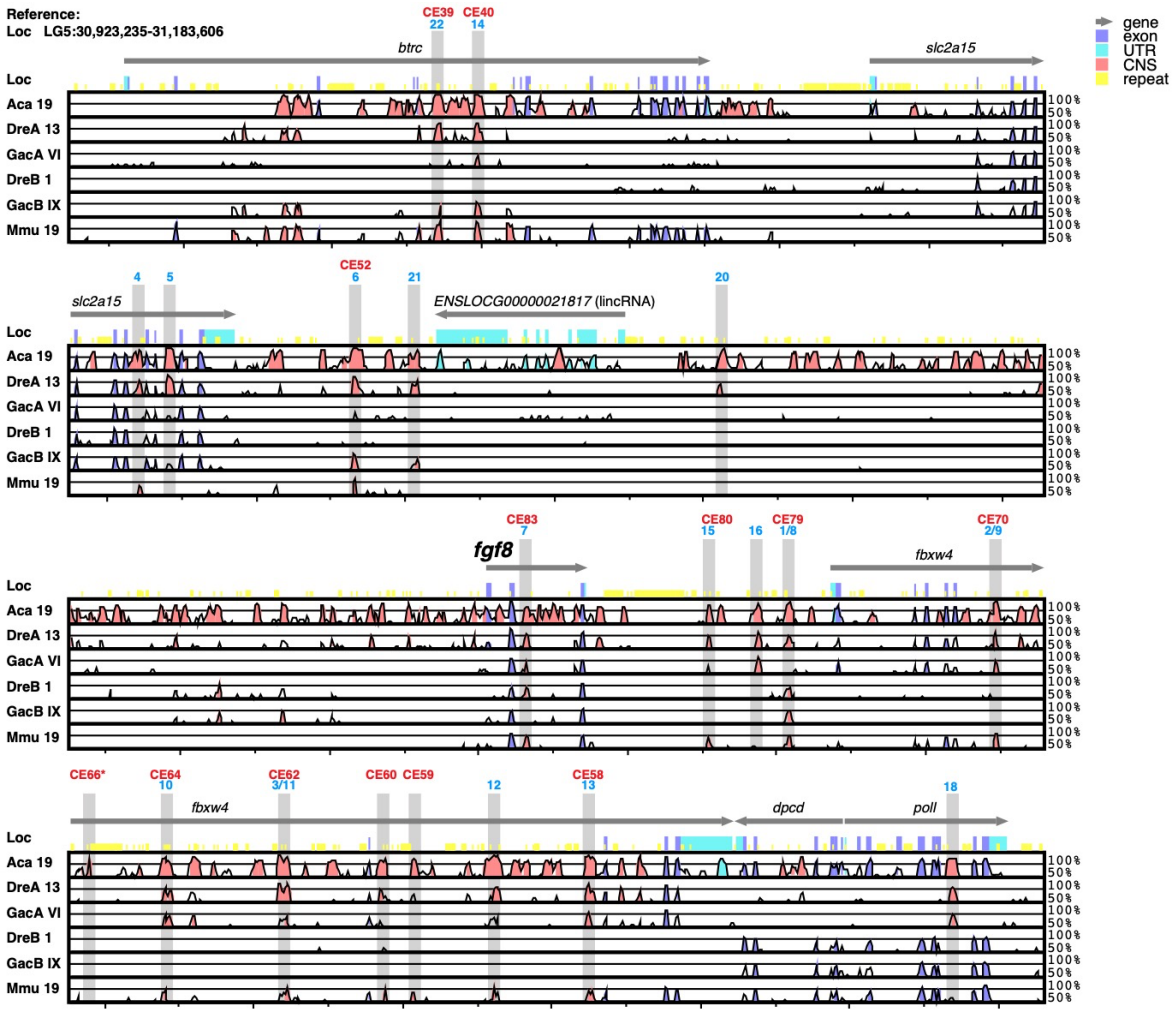
Conservation with teleosts is not observed. Hind limb enhancer B (HLEB, black box)[78] is indicated as well.
Bottom: VISTA plot of the stickleback *tbx4* gene, featuring three putative 'swimbladder enhancer' regions (SBE1-3, green boxes; as defined by Nikaido et al. 2014[79]), of which SBE2 is conserved with bowfin but not lobe-finned vertebrates. Two other regions in intron 3, regions R1 and R2 (blue boxes) appear as conserved among neopterygians (teleosts, bowfin, gar), but not with coelacanth or tetrapods. HLEB (black box) is also present.
Middle: VISTA plot of the bowfin *tbx4* gene identifies different conserved regions defined in the human- and stickleback-centric alignments (top, bottom): SBE2 (green box), R2, and HLEB. Importantly, the lobe-finned LME and the neopterygian R1 elements overlap within the bowfin genome, and are thus orthologous conserved non-coding elements, connected via the holosteans bridge (right). This bowfin LME/R1 region is also a bowfin developmental ncOCR (main text Fig. 4E).
Note that the *tbx4* genomic region is shown here in forward orientation, while it is located on the reverse strand in the bowfin genome assembly (as shown in main text Fig. 4E).
Fish silhouettes obtained from phylopic.org: human (credit: A. A. Farke), Western clawed frog (S. Miranda-Rottmann), bowfin (D. Raver), spotted gar, stickleback (both M. Tan), Japanese puffer (uncredited); license: https://creativecommons.org/licenses/by-nc-sa/3.0/.



Supplementary Fig. 15. Bowfin *hoxd14* pseudogene. IGV browser view of the *hoxd14* region on Aca scaf 17 shows read coverage from fin bud RNA-seq transcriptomes (paired-end samples) and whole embryo ATAC-Seq profiles of different developmental stages[89]. Predicted exons are indicated in red with dark red indicating parts confirmed by fin bud cDNA cloning and sequencing.



Supplementary Fig. 16. Bowfin pectoral fin Evo-Devo. A) Evolution of the pectoral appendage endoskeleton in bony fishes. Elements are colored based on which portion of the fin they comprise. From the tribasal ancestral condition, teleosts and gars independently lost the metapterygium, while tetrapods lost the propterygium and mesopterygium. Bowfin is the closest living relative of the teleosts that retains the metapterygium. B) Bowfin pectoral fins at two different developmental stages cleared and stained with alcian blue for cartilage and alizarin red for bone. Anterior to left, distal to top in all panels. C) Bowfin developmental stages[89] sampled for RNA-seq transcriptomics. Arrow in dorsal view (middle column) indicates developing pectoral fin, dotted line in lateral view (right column) indicates outline of the developing pectoral fin bud.



Supplementary Fig. 17. Genomic alignment of bony vertebrate *fgf8* gene regions. mVISTA plot showing sequence conservation of bowfin (Aca), zebrafish (Dre; *fgf8a* and *fgf8b* paralogons), stickleback (Gac; *fgf8a* and *fgf8b* paralogons), and mouse (Mmu) aligned against spotted gar (Loc) as reference species. Grey shaded areas highlight known regulatory elements for zebrafish (blue numbers)[90] and mouse (red CE numbers)[91]. All relevant elements, and in particular those known to be important for expression in the apical ectodermal ridge of the developing mouse limb[93], i.e. CE58, CE59, CE66, CE80, with the exception of CE61 (also absent from zebrafish), are found in gar and bowfin. CE66 (asterisk) was identified by sequence similarity to human.

Supplementary Tables

Supplementary Table 1. Genome assembly statistics for bowfin genome AmiCal1.

Assembly version	Sequenced read pairs	Scaffold N	Genome Size	N50	L50	N90	L90	CEGMA	BUSCO	% assembly in 23 superscaf.
de novo Assembly	433.5M	64,769	0.77 Gb	0.02 Mb	8,829					
Chicago Scaffolding	177M	1,846	0.78 Gb	11.7 Mb	22					
Chicago+ HiC Scaffolding	527M	1,958	0.831 Gb	41.2 Mb	9	22.5 Mb	20	98.3%: 244/248 eukaryotic CEGs	100%: 303/303 eukaryotic BUSCOs	99% (0.824/0.831Mb)

Supplementary Table 2. Repeat content in bowfin (AmiCal1) in comparison to spotted gar (LepOcu1)[9].

Element name	Bowfin			Gar		
	Genome Coverage (bp)	% Genome Coverage	# Copies	Genome Coverage (bp)	% Genome Coverage	# Copies
DNA (all)	**58896956**	**7.087**	**283729**	**33852877**	**3.544**	**139185**
DNA/EnSpm	1501171	0.181	10331	13259	0.001	94
DNA/Harbinger (all)	977597	0.118	4641	324845	0.034	2907
DNA/PIF-Harbinger	658249	0.079	2702	267478	0.028	816
DNA/hAT (all)	14508164	1.746	71543	2704250	0.283	18884
DNA/hAT-Ac	3534231	0.425	15119	176016	0.018	590
DNA/hAT-Buster	82903	0.010	143	653918	0.068	2088
DNA/hAT-Charlie	3177061	0.382	16235	3066787	0.321	16683
DNA/hAT-Tip100	1434199	0.173	5601	648362	0.068	2714
DNA/Helitron	634349	0.076	4173	1387	0.000	10
DNA/Kolobok	302794	0.036	1539	116	0.000	1
DNA/Mariner (all)	32042386	3.856	145570	2731980	0.286	7728
DNA/Merlin	43943	0.005	694	0	0.000	0
DNA/MuDr	200652	0.024	1754	406	0.000	1
DNA/PiggyBac	324070	0.039	1302	138983	0.015	567
DNA/Polinton	563313	0.068	2836	969065	0.101	1632
DNA/Sola	649039	0.078	3105	1833	0.000	4
DNA/TcMar (all)	31929994	3.842	145107	278427	0.029	1610
DNA/TcMar-MER6	7544	0.001	128	1532839	0.161	8845
DNA/TcMar-Pogo	259	0.000	4	243385	0.025	1312
DNA/TcMar-Tc1	31118535	3.745	141278	18223551	1.909	65097
DNA/TcMar-Tigger	378064	0.045	2249	1656906	0.174	6462
DNA/Zisupton	682217	0.082	6322	0	0.000	0
IntegratedVirus/Caulimovirus	**0**	**0.000**	**0**	**23078**	**0.002**	**319**
LINE (all)	**32650404**	**3.929**	**122893**	**59482068**	**5.757**	**202097**
LINE/CR1	14424539	1.736	53247	22706510	2.378	90057
LINE/L1	415886	0.050	2192	2269353	0.238	6838
LINE/L2	857499	0.103	2486	5170501	0.542	24198
LINE/Penelope	94490	0.011	627	2642775	0.277	9393
LINE/R2	6953	0.001	71	3300758	0.346	15582
LINE/R4	225	0.000	3	4185	0.000	12
LINE/Rex-Babar	67783	0.008	294	8336055	0.873	33565
LINE/Rex1	164118	0.020	628	3752335	0.393	7608
LINE/RTE (all)	4186616	0.504	19482	1159548	0.121	3910
LINE/RTE-BovB	15222	0.002	59	915697	0.096	2846
LINE/RTE-X	1163469	0.140	5645	1136602	0.119	6384
LINE/Vingi	45250	0.005	203	262949	0.028	1704
Low-complexity	**2183573**	**0.263**	**42965**	**4523626**	**0.474**	**113467**
LTR (all)	**18887718**	**2.273**	**58259**	**24380821**	**2.553**	**148279**
LTR/BEL	431888	0.052	1004	1844404	0.193	17715
LTR/Copia	108033	0.013	252	258542	0.027	420
LTR/ERV1	1166885	0.140	6342	2603185	0.273	7728
LTR/Gypsy (all)	7332735	0.882	17056	5295012	0.555	13025
LTR/Gypsy-Gmr1	279	0.000	1	108781	0.011	218
LTR/Ngaro	987717	0.119	5219	13075343	1.369	60441
Satellite/Simple-repeat	**17206706**	**2.071**	**382157**	**2948495**	**0.309**	**46452**
SINE (all)	**2698499**	**0.325**	**24013**	**37154654**	**2.769**	**161317**
SINE/rRNA (all)	73728	0.009	567	11124517	1.165	34827
SINE/5S	73491	0.009	564	11036055	1.156	34103
SINE/AFC	512	0.000	7	1479	0.000	21
SINE/Deu	152961	0.018	1231	6500648	0.681	31999
SINE/HPA	1197	0.000	24	35107	0.004	435
SINE/MIR	1811610	0.218	15820	2879141	0.302	19453
SINE/tRNA	559995	0.067	5216	21785	0.002	239
SINE/Unclassified	0	0.000	0	288	0.000	3
SINE/V	54117	0.007	624	5204795	0.545	37440
SINE?	1829	0.000	25	116960	0.012	739
Unknown/Other misc.	**51278478**	**6.171**	**255416**	**52882259**	**5.538**	**240704**
Total	**183802334**	**22.12**	**1169432**	**196387439**	**20.57**	**971268**

Supplementary Table 3. OrthoFinder results summary.

Number of genes	251621
Number of genes in orthogroups	234400
Number of unassigned genes	17221
Percentage of genes in orthogroups	93.2
Percentage of unassigned genes	6.8
Number of orthogroups	15916
Number of species-specific orthogroups	86
Number of genes in species-specific orthogroups	579
Percentage of genes in species-specific orthogroups	0.2
Mean orthogroup size	14.7
Median orthogroup size	12
G50 (assigned genes)	16
G50 (all genes)	15
O50 (assigned genes)	3745
O50 (all genes)	4306
Number of orthogroups with all species present	7532
Number of single-copy orthogroups	2079

Supplementary Table 4. OrthoFinder orthogroup statistics for bowfin and 11 other vertebrates.

	Bowfin	Gar	Arowana	Zebrafish	Medaka	Xenopus	Coelacanth	Mouse	Human	Chicken	Anole	Elephant Shark
Number of genes	21948	18341	25402	25902	19699	18442	19569	22585	23070	18346	18595	19722
Number of genes in orthogroups	19047	17887	24632	23758	18002	17620	18840	20670	21799	16518	17431	18196
Number of unassigned genes	2901	454	770	2144	1697	822	729	1915	1271	1828	1164	1526
Percentage of genes in orthogroups	86.8	97.5	97	91.7	91.4	95.5	96.3	91.5	94.5	90	93.7	92.3
Percentage of unassigned genes	13.2	2.5	3	8.3	8.6	4.5	3.7	8.5	5.5	10	6.3	7.7
Number of orthogroups containing species	13309	12637	12713	12751	11372	11356	12296	13237	13345	11485	11745	11962
Percentage of orthogroups containing species	83.6	79.4	79.9	80.1	71.5	71.3	77.3	83.2	83.8	72.2	73.8	75.2
Number of species-specific orthogroups	12	2	3	3	10	8	8	7	4	8	8	13
Number of genes in species-specific orthogroups	38	4	8	8	62	40	66	25	30	76	134	88
Percentage of genes in species-specific orthogroups	0.2	0	0	0	0.3	0.2	0.3	0.1	0.1	0.4	0.7	0.4

Supplementary Table 5. A reference-free k-mer analysis. Numbers for male and female specific k-mer count are listed for three different filtering conditions to determine a k-mer as sex specific. In all three filtering conditions, males consistently showed higher number of sex-specific k-mers, which could suggest male as the heterogametic sex in bowfin.

Minimal count in heterogametic sex	Maximal count in homogametic sex	Number of male-specific k-mers	Number of female-specific k-mers
25	5	95,928	68,857
15	2	546,314	495,804
15	0	546,314	495,804

Supplementary Table 6. Bowfin scaf_14 MHC region predicted genes.
[separate .pdf file]

Supplementary Table 7. Gar MHC genes.
[separate .xls file]

Supplementary Tables 8-12. Immune gene accessions.
[combined separate .pdf file]

Suppl. Table 8. MHC Sequence Accession Identifiers.

Suppl. Table 9. Immunoglobulin heavy chain (IgH) sequence accession identifiers.

Suppl. Table 10. Immunoglobulin light chain (IgL) sequence accession identifiers.

Suppl. Table 11. T cell receptor (TCR) sequence accession identifiers.

Suppl. Table 12. Bowfin Toll-like Receptor (TLRs).

Supplementary Table 13. Relative SCPP gene expression levels in zebrafish skin at three ages. Fragments Per Kilobase of transcript, per Million mapped reads (FPKM), showing relative abundance of each transcript among all transcripts in the zebrafish skin. FPKM_conf_lo/FPKM_conf_hi represent the lower/upper bound of the 95% confidence intervals of the abundance of each transcript. Relative expression levels of SCPP genes located on chromosomes Dre5 and Dre10 are high (>100 in FPKM) or significant (>10 in FPKM and >0 in FPKM_conf_lo). In contrast, relative expression level of *scpp5* is low (5.6 in the FPKM_conf_lo value only at 5 months of age).

Location		Dre1		Dre10				Dre5			
Gene		scpp5	scpp1	spp1	scpp8	scpp11b	scpp11a	scpp14	scpp13	gsp37	scpp12
5 months	FPKM	7.4	20.9	46.7	8.5	401.4	1340.3	128.0	187.0	16.5	26.9
	FPKM_conf_lo	5.6	18.6	43.6	5.6	381.8	1302.9	119.4	180.2	13.2	23.9
	FPKM_conf_hi	9.2	23.2	49.8	11.5	421.1	1377.8	136.5	193.8	19.7	29.8
24 months	FPKM	0.1	18.9	48.5	6.3	277.2	867.9	104.5	129.5	7.0	16.6
	FPKM_conf_lo	0.0	16.6	45.1	3.7	259.8	836.1	96.3	123.5	4.7	14.1
	FPKM_conf_hi	0.4	21.3	51.8	9.0	294.6	899.8	112.6	135.5	9.2	19.0
42 months	FPKM	0.4	9.4	24.1	14.3	377.4	1255.8	107.1	105.3	12.0	29.9
	FPKM_conf_lo	0.0	7.9	21.9	10.6	359.1	1220.9	99.6	100.3	9.3	26.9
	FPKM_conf_hi	0.8	10.8	26.2	18.0	395.8	1290.7	114.6	110.2	14.7	32.9

Supplementary Table 14. Number of OCRs found for each bowfin developmental stage.

Stage	OCRs per stage	Unique OCRs	ncOCRs per stage	Unique ncOCRs
22-23 (phylotypic)	67989	3584	51433	3229
23-24	82966	5356	63876	4690
24-25	81289	3648	63948	3589
26-27	90469	4716	71576	4611
28-29	114174	17728	86650	13941
30-31	103336	13138	77912	10553
All merged across stages	163771	-	132119	-

Supplementary Table 15. Number of bowfin OCRs in N developmental stages.

# Stage(s)	OCRs	ncOCRs
1 stage	48170	40613
2 stages	28075	23256
3 stages	20666	17732
4 stages	14463	12657
5 stages	14008	12067
6 stages	38042	25805

Supplementary Table 16. HOMER annotation of OCRs in each developmental stage based on nearest MAKER feature. TSS is transcript start site and TTS is transcript termination site. Log ratio and enrichment are based on the expected and observed proportions of peaks in a given feature.

Stage 22-23				
Annotation	**Number of peaks**	**Total size (bp)**	**Log2 Ratio (obs/exp)**	**LogP enrichment (+values depleted)**
TTS	1790	21856783	-0.008	0.895
Exon	5300	41761014	0.624	-462.176
Intron	17359	243178172	-0.207	263.725
Intergenic	37410	494804573	-0.124	343.482
TSS	6090	23322407	1.665	-2996.855
Stage 23-24				
Annotation	**Number of peaks**	**Total size (bp)**	**Log2 Ratio (obs/exp)**	**LogP enrichment (+values depleted)**
TTS	2245	21856783	0.031	-1.886
Exon	7117	41760890	0.762	-897.519
Intron	20983	243178189	-0.22	361.776
Intergenic	45584	494804573	-0.126	431.827
TSS	6987	23322407	1.576	-3129.045
Stage 24-25				
Annotation	**Number of peaks**	**Total size (bp)**	**Log2 Ratio (obs/exp)**	**LogP enrichment (+values depleted)**
TTS	2113	21851276	-0.027	1.619
Exon	6046	41760102	0.556	-424.558
Intron	20241	243158426	-0.243	425.312
Intergenic	46078	494594613	-0.08	181.056
TSS	6754	23316902	1.556	-2961.083
Stage 26-27				
Annotation	**Number of peaks**	**Total size (bp)**	**Log2 Ratio (obs/exp)**	**LogP enrichment (+values depleted)**
TTS	2325	21857884	-0.043	2.625
Exon	6994	41762201	0.612	-586.971
Intron	22989	243183740	-0.213	371.894
Intergenic	51082	494805956	-0.086	230.694
TSS	7027	23324609	1.459	-2755.25
Stage 28-29				
Annotation	**Number of peaks**	**Total size (bp)**	**Log2 Ratio (obs/exp)**	**LogP enrichment (+values depleted)**
TTS	3246	21857884	0.103	-10.576
Exon	12605	41762341	1.126	-3232.078
Intron	29346	243183736	-0.197	402.652
Intergenic	60724	494805956	-0.173	1075.038
TSS	8182	23324609	1.343	-2775.976
Stage 30-31				
Annotation	**Number of peaks**	**Total size (bp)**	**Log2 Ratio (obs/exp)**	**LogP enrichment (+values depleted)**
TTS	2793	21857884	0.03	-1.983
Exon	11392	41762825	1.124	-2912.155
Intron	26298	243183746	-0.211	415.628
Intergenic	54994	494805956	-0.172	962.896
TSS	7788	23324609	1.415	-2897.392

Supplementary Table 17. Overlap of bowfin OCRs with bowfin Ultra Conserved Elements (UCEs)[66] and gar-centric Conserved Non-coding Elements (CNEs)[9].

Total bowfin UCEs found in genome	
Bowfin UCEs	366
BLASTN hits in bowfin genome	364 (99%)
BLASTN hits overlapping bowfin OCRs all stages merged	144 (40%)
BLASTN hits overlapping bowfin ncOCRs all stages merged	122 (34%)
*Above results are the same for 1bp+ and 33%+ overlap	
Vertebrate CNEs detected in bowfin genome	
Gar centric WGA CNEs	156087
BLASTN hits to bowfin genome	64608 (41%)
BLASTN hits overlapping 1bp + bowfin ncOCRs	21257 (33%)
BLASTN hits overlapping 33%+ bowfin ncOCRs	21279 (33%)
Bowfin ncOCRs	132119
Bowfin ncOCR BLASTN hits in gar genome	52890 (40%)

Supplementary Table 18. Number of VISTA enhancers detected in bowfin and zebrafish genomes.

	VISTA enhancer mouse	VISTA enhancer human
VISTA enhancers	621	989
BLASTN hits in Zebrafish	47 (8%)	449 (45%)
BLASTN hits in Bowfin	97 (16%)	600 (61%)
BLASTN hits overlapping bowfin OCRs	71 (73%)	330 (55%)
BLASTN hits overlapping bowfin ncOCRs	65 (67%)	314 (52%)

Supplementary Table 19. Location of human VISTA enhancers in bowfin OCRs.
[separate .xls file]

Supplementary Table 20. Mouse OCRs[1] found in bowfin.

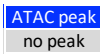
Genomic feature	Mouse OCRs	BLASTN hits in Bowfin genome	BLASTN hits overlapping bowfin OCRs	BLASTN hits overlapping bowfin ncOCRs
Exon	20979	5790 (28%)	1869 (32%)	226 (4%)
Intron	135807	6842 (5%)	2234 (33%)	1323 (19%)
Intergenic	90217	1752 (2%)	876 (50%)	814 (46%)
TSS	52981	3064 (6%)	1645 (54%)	514 (17%)
TTS	4516	101 (2%)	28 (28%)	17 (17%)
Total OCRs in mouse	304500	17549 (6%)	6652 (38%)	2894 (17%)
Total noncoding OCRs in mouse	226024	8594 (4%)	3110 (36%)	2137 (25%)
*Above results are the same for 1bp+ and 33%+ overlap				

Supplementary Table 21. Number of tissue-specific mouse OCRs[1] found in bowfin.

Tissue	Mouse OCRs	BLASTN hits in Bowfin genome	BLASTN hits overlapping bowfin OCRs	BLASTN hits overlapping bowfin ncOCRs
Allantois	793	31 (4%)	8 (26%)	8 (26%)
Cardiomyocytes	1519	56 (4%)	34 (61%)	21 (38%)
Endothelium	2492	82 (3%)	48 (59%)	31 (38%)
Erythroid	1566	78 (5%)	16 (21%)	10 (13%)
ExE endoderm	354	11 (3%)	3 (27%)	1 (9%)
Forebrain	1550	209 (13%)	165 (79%)	153 (73%)
Gut	1284	75 (6%)	41 (55%)	26 (35%)
Mesenchyme	1726	36 (2%)	16 (44%)	14 (39%)
Mid Hindbrain	1032	190 (18%)	150 (79%)	141 (74%)
Mixed mesoderm	588	50 (9%)	30 (60%)	23 (46%)
Neural crest	1156	76 (7%)	49 (64%)	46 (61%)
NMP	2107	217 (10%)	144 (66%)	124 (57%)
Notochord	149	8 (5%)	5 (63%)	3 (38%)
Paraxial mesoderm	1166	91 (8%)	63 (69%)	50 (55%)
Phayngeal mesoderm	974	64 (7%)	39 (61%)	38 (59%)
Somitic mesoderm	2029	132 (7%)	82 (62%)	64 (48%)
Spinal cord	2728	332 (12%)	251 (76%)	230 (69%)
Surface ectoderm	1245	68 (5%)	38 (56%)	27 (40%)
Ubiquitous	857	73 (9%)	66 (90%)	25 (34%)
**Above results are the same for 1bp+ and 33%+ overlap*				

Supplementary Table 22. Putative *fgf8* gene regulatory regions in bowfin and their ATAC-Seq profile.

Region		Location Gar	Location Bowfin	ATAC-seq peak (Ballard stages)					
Mouse (ref. 1)	**Zebrafish** (ref. 2)			22-23	23-24	24-25	26-27	28-29	30-31
CE39	fgf.dr22	LG5:30947677-30948088	Aca_scaf_19:20039103-20039514	■	■	■	■	■	■
CE40	fgf.dr14	LG5:30950223-30950719	Aca_scaf_19:20036292-20036744			■			
	fgf.dr4	LG5:30992807-30992954	Aca_scaf_19:19999288-19999434						
	fgf.dr5	LG5:30994709-30995125	Aca_scaf_19:19996896-19997335						
CE52	fgf.dr6	LG5:31007087-31007580	Aca_scaf_19:19985590-19986107			■	■	■	■
	fgf.dr21	LG5:31011442-31011566	Aca_scaf_19:19982750-19983108			■			
	fgf.dr20	LG5:31031807-31031940	Aca_scaf_19:19964282-19964154						
CE83	fgf.dr7	LG5:31083803-31084013	Aca_scaf_19:19919190-19919363						
CE80	fgf.dr15	LG5:31096149-31096261	Aca_scaf_19:19907658-19907719						
	fgf.dr16	LG5:31099373-31099672	Aca_scaf_19:19904756-19905061						
CE79	fgf.dr1/8	LG5:31101477-31101648	Aca_scaf_19:19902701-19902861	■	■	■	■		
CE70	fgf.dr2/9	LG5:31115298-31115473	Aca_scaf_19:19893350-19893526						
CE66		LG5:31119662-31119764	Aca_scaf_19:19889478-19889663						
CE64	fgf.dr10	LG5:31124736-31125137	Aca_scaf_19:19885965-19886447						
CE62	fgf.dr3/11	LG5:31132406-31133156	Aca_scaf_19:19878997-19879728	■	■				
CE59		LG5:31141354-31141592	Aca_scaf_19:19873851-19874071			■		■	■
CE60		LG5:31146201-31147373	Aca_scaf_19:19874776-19875263						
	fgf.dr12	LG5:31146831-31147276	Aca_scaf_19:19870268-19870878						
CE58	fgf.dr13	LG5:31153080-31153525	Aca_scaf_19:19867354-19867788	■	■	■		■	■
	fgf.dr18	LG5:31177489-31177812	Aca_scaf_19:19852686-19853018	■	■	■	■		



References:
1. Marinic M, Aktas T, Ruf S, & Spitz F (2013) Dev Cell 24(5):530-542.
2. Komisarczuk AZ, Kawakami K, & Becker TS (2009) Dev Biol 336(2):301-312.

Supplementary References

1 Pijuan-Sala, B. *et al.* Single-cell chromatin accessibility maps reveal regulatory programs driving early mouse organogenesis. *Nature Cell Biology* **22**, 487-497, doi:10.1038/s41556-020-0489-9 (2020).

2 Chapman, J. A. *et al.* Meraculous: de novo genome assembly with short paired-end reads. *PLoS One* **6**, e23501, doi:10.1371/journal.pone.0023501 (2011).

3 Putnam, N. H. *et al.* Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. *Genome Res* **26**, 342-350, doi:10.1101/gr.193474.115 (2016).

4 Hardie, D. C. & Hebert, P. D. N. Genome-size evolution in fishes. *Canadian Journal of Fisheries and Aquatic Sciences* **61**, 1636-1646 (2004).

5 Ohno, S. *et al.* Microchromosomes in holocephalian, chondrostean and holostean fishes. *Chromosoma* **26**, 35-40, doi:10.1007/BF00319498 (1969).

6 Majtanova, Z., Symonova, R., Arias-Rodriguez, L., Sallan, L. & Rab, P. "Holostei versus Halecostomi" Problem: Insight from Cytogenetics of Ancient Nonteleost Actinopterygian Fish, Bowfin Amia calva. *J Exp Zool B Mol Dev Evol* **328**, 620-628, doi:10.1002/jez.b.22720 (2017).

7 Simao, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210-3212, doi:10.1093/bioinformatics/btv351 (2015).

8 Parra, G., Bradnam, K. & Korf, I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **23**, 1061-1067, doi:10.1093/bioinformatics/btm071 (2007).

9 Braasch, I. *et al.* The spotted gar genome illuminates vertebrate evolution and facilitates human-teleost comparisons. *Nat Genet* **48**, 427-437, doi:10.1038/ng.3526 (2016).

10 Bohne, A. *et al.* Zisupton--a novel superfamily of DNA transposable elements recently active in fish. *Mol Biol Evol* **29**, 631-645, doi:10.1093/molbev/msr208 (2012).

11 Holt, C. & Yandell, M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* **12**, 491, doi:10.1186/1471-2105-12-491 (2011).

12 Bateman, A. *et al. The Pfam Protein Families Database.* http://www.sanger.ac.uk/Software/Pfam/ (2000).

13 Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol* **20**, 238, doi:10.1186/s13059-019-1832-y (2019).

14 Emms, D. M. & Kelly, S. STAG: Species Tree Inference from All Genes. *bioRxiv*, 267914, doi:10.1101/267914 (2018).

15 Kottler, V. A. *et al.* Independent Origin of XY and ZW Sex Determination Mechanisms in Mosquitofish Sister Species. *Genetics* **214**, 193-209, doi:10.1534/genetics.119.302698 (2020).

16 Pan, Q. *et al.* Identification of the master sex determining gene in Northern pike (Esox lucius) reveals restricted sex chromosome differentiation. *PLoS Genet* **15**, e1008013, doi:10.1371/journal.pgen.1008013 (2019).

17 Wen, M. *et al.* Sex chromosome and sex locus characterization in goldfish, Carassius auratus (Linnaeus, 1758). *BMC Genomics* **21**, 552, doi:10.1186/s12864-020-06959-3 (2020).

18 Zhang, A. *et al.* Computational identification of Y-linked markers and genes in the grass carp genome by using a pool-and-sequence method. *Sci Rep* **7**, 8213, doi:10.1038/s41598-017-08476-y (2017).

19 Darolti, I. *et al.* Extreme heterogeneity in sex chromosome differentiation and dosage compensation in livebearers. *Proc Natl Acad Sci U S A* **116**, 19031-19036, doi:10.1073/pnas.1905298116 (2019).

20 Li, S. *et al.* A new approach for comprehensively describing heterogametic sex chromosomes. *DNA Res* **25**, 375-382, doi:10.1093/dnares/dsy010 (2018).

21 Morris, J., Darolti, I., Bloch, N. I., Wright, A. E. & Mank, J. E. Shared and Species-Specific Patterns of Nascent Y Chromosome Evolution in Two Guppy Species. *Genes (Basel)* **9**, doi:10.3390/genes9050238 (2018).

22 Moore, E. C. & Roberts, R. B. Polygenic sex determination. *Curr Biol* **23**, R510-512, doi:10.1016/j.cub.2013.04.004 (2013).

23 Baroiller, J. F., D'Cotta, H. & Saillant, E. Environmental effects on fish sex determination and differentiation. *Sex Dev* **3**, 118-135, doi:10.1159/000223077 (2009).

24 Kamiya, T. *et al.* A trans-species missense SNP in Amhr2 is associated with sex determination in the tiger pufferfish, Takifugu rubripes (fugu). *PLoS Genet* **8**, e1002798, doi:10.1371/journal.pgen.1002798 (2012).
25 Hauptmann, G. & Bahram, S. Genetics of the central MHC. *Curr Opin Immunol* **16**, 668-672, doi:10.1016/j.coi.2004.07.001 (2004).
26 Trowsdale, J. The MHC, disease and selection. *Immunol Lett* **137**, 1-8, doi:10.1016/j.imlet.2011.01.002 (2011).
27 Ohta, Y. *et al.* Primitive synteny of vertebrate major histocompatibility complex class I and class II genes. *Proc Natl Acad Sci U S A* **97**, 4712-4717, doi:10.1073/pnas.97.9.4712 (2000).
28 Grimholt, U. MHC and Evolution in Teleosts. *Biology (Basel)* **5**, 6, doi:10.3390/biology5010006 (2016).
29 Yamaguchi, T. & Dijkstra, J. M. Major Histocompatibility Complex (MHC) Genes and Disease Resistance in Fish. *Cells* **8**, 378, doi:10.3390/cells8040378 (2019).
30 Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *Journal of molecular biology* **215**, 403-410, doi:10.1016/S0022-2836(05)80360-2 (1990).
31 Thompson, J. D., Higgins, D. G. & Gibson, T. J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22**, 4673-4680, doi:10.1093/nar/22.22.4673 (1994).
32 Kumar, S., Stecher, G., Li, M., Knyaz, C. & Tamura, K. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Mol Biol Evol* **35**, 1547-1549, doi:10.1093/molbev/msy096 (2018).
33 Grimholt, U. *et al.* A comprehensive analysis of teleost MHC class I sequences. *BMC Evol Biol* **15**, 32, doi:10.1186/s12862-015-0309-1 (2015).
34 Grimholt, U., Tsukamoto, K., Hashimoto, K. & Dijkstra, J. M. Discovery of a Novel MHC Class I Lineage in Teleost Fish which Shows Unprecedented Levels of Ectodomain Deterioration while Possessing an Impressive Cytoplasmic Tail Motif. *Cells* **8**, 1056, doi:10.3390/cells8091056 (2019).
35 Swann, J. B., Holland, S. J., Petersen, M., Pietsch, T. W. & Boehm, T. The immunogenetics of sexual parasitism. *Science* **369**, 1608-1615, doi:10.1126/science.aaz9445 (2020).
36 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 Evol Biol* **13**, 260, doi:10.1186/1471-2148-13-260 (2013).
37 Dirscherl, H., McConnell, S. C., Yoder, J. A. & de Jong, J. L. The MHC class I genes of zebrafish. *Dev Comp Immunol* **46**, 11-23, doi:10.1016/j.dci.2014.02.018 (2014).
38 Traver, D. & Yoder, J. A. in *The Zebrafish in Biomedical Research* (eds S. Cartner *et al.*) 191–216 (Academic Press, 2020).
39 Flajnik, M. F. A cold-blooded view of adaptive immunity. *Nat Rev Immunol* **18**, 438-453, doi:10.1038/s41577-018-0003-9 (2018).
40 Flajnik, M. F. & Kasahara, M. Origin and evolution of the adaptive immune system: genetic events and selective pressures. *Nat Rev Genet* **11**, 47-59, doi:10.1038/nrg2703 (2010).
41 Mirete-Bachiller, S., Olivieri, D. N. & Gambón-Deza, F. Immunoglobulin T genes in Neopterygii. *bioRxiv*, 2020.2005.2021.108993, doi:10.1101/2020.05.21.108993 (2020).
42 Guselnikov, S. V. *et al.* Diversity of Immunoglobulin Light Chain Genes in Non-Teleost Ray-Finned Fish Uncovers IgL Subdivision into Five Ancient Isotypes. *Front Immunol* **9**, 1079, doi:10.3389/fimmu.2018.01079 (2018).
43 Rast, J. P. *et al.* alpha, beta, gamma, and delta T cell antigen receptor genes arose early in vertebrate phylogeny. *Immunity* **6**, 1-11, doi:10.1016/s1074-7613(00)80237-x (1997).
44 Aoki, T., Hikima, J., Hwang, S. D. & Jung, T. S. Innate immunity of finfish: primordial conservation and function of viral RNA sensors in teleosts. *Fish Shellfish Immunol* **35**, 1689-1702, doi:10.1016/j.fsi.2013.02.005 (2013).
45 Fitzgerald, K. A. & Kagan, J. C. Toll-like Receptors and the Control of Immunity. *Cell* **180**, 1044-1066, doi:10.1016/j.cell.2020.02.041 (2020).

46 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, doi:10.1007/s00251-013-0694-9 (2013).
47 Wcisel, D. J., Ota, T., Litman, G. W. & Yoder, J. A. Spotted Gar and the Evolution of Innate Immune Receptors. *J Exp Zool B Mol Dev Evol* **328**, 666-684, doi:10.1002/jez.b.22738 (2017).
48 Sievers, F. & Higgins, D. G. Clustal Omega. *Current Protocols in Bioinformatics* **48**, 3.13.11-13.13.16, doi:10.1002/0471250953.bi0313s48 (2014).
49 Pasquier, J. *et al.* Gene evolution and gene expression after whole genome duplication in fish: the PhyloFish database. *BMC Genomics* **17**, 368, doi:10.1186/s12864-016-2709-z (2016).
50 Minh, B. Q. *et al.* IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol Biol Evol* **37**, 1530-1534, doi:10.1093/molbev/msaa015 (2020).
51 Minh, B. Q., Nguyen, M. A. & von Haeseler, A. Ultrafast approximation for phylogenetic bootstrap. *Mol Biol Evol* **30**, 1188-1195, doi:10.1093/molbev/mst024 (2013).
52 Hoang, D. T. *et al.* MPBoot: fast phylogenetic maximum parsimony tree inference and bootstrap approximation. *BMC Evol Biol* **18**, 11, doi:10.1186/s12862-018-1131-3 (2018).
53 Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res* **39**, W29-37, doi:10.1093/nar/gkr367 (2011).
54 Kawasaki, K. *et al.* SCPP Genes and Their Relatives in Gar: Rapid Expansion of Mineralization Genes in Osteichthyans. *J Exp Zool B Mol Dev Evol* **328**, 645-665, doi:10.1002/jez.b.22755 (2017).
55 Kawasaki, K. & Weiss, K. M. Mineralized tissue and vertebrate evolution: the secretory calcium-binding phosphoprotein gene cluster. *Proc Natl Acad Sci U S A* **100**, 4060-4065, doi:10.1073/pnas.0638023100 (2003).
56 Afgan, E. *et al.* The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Res* **44**, W3-W10, doi:10.1093/nar/gkw343 (2016).
57 Mansfeld, J. *et al.* Branched-chain amino acid catabolism is a conserved regulator of physiological ageing. *Nat Commun* **6**, 10043, doi:10.1038/ncomms10043 (2015).
58 Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120, doi:10.1093/bioinformatics/btu170 (2014).
59 Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21, doi:10.1093/bioinformatics/bts635 (2013).
60 Kapustin, Y., Souvorov, A., Tatusova, T. & Lipman, D. Splign: algorithms for computing spliced alignments with identification of paralogs. *Biol Direct* **3**, 20, doi:10.1186/1745-6150-3-20 (2008).
61 Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28**, 511-515, doi:10.1038/nbt.1621 (2010).
62 Liu, Z. *et al.* The channel catfish genome sequence provides insights into the evolution of scale formation in teleosts. *Nat Commun* **7**, 11757, doi:10.1038/ncomms11757 (2016).
63 Miyabe, K. *et al.* GSP-37, a novel goldfish scale matrix protein: identification, localization and functional analysis. *Faraday Discussions* **159**, 463-481, doi:10.1039/C2FD20051A (2012).
64 Buenrostro, J. D., Wu, B., Chang, H. Y. & Greenleaf, W. J. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. *Curr Protoc Mol Biol* **109**, 21 29 21-21 29 29, doi:10.1002/0471142727.mb2129s109 (2015).
65 Heinz, S. *et al.* Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Molecular Cell* **38**, 576-589, doi:10.1016/j.molcel.2010.05.004 (2010).
66 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**, doi:10.1371/journal.pone.0065923 (2013).
67 Visel, A., Minovitsky, S., Dubchak, I. & Pennacchio, L. A. VISTA Enhancer Browser - A database of tissue-specific human enhancers. *Nucleic Acids Research* **35**, doi:10.1093/nar/gkl822 (2007).
68 Quinlan, A. R. & Hall, I. M. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842, doi:10.1093/bioinformatics/btq033 (2010).
69 Babarinde, I. A. & Saitou, N. The Dynamics, Causes, and Impacts of Mammalian Evolutionary Rates Revealed by the Analyses of Capybara Draft Genome Sequences. *Genome Biol Evol* **12**, 1444-1458, doi:10.1093/gbe/evaa157 (2020).

70 Graham, J. B. *Air-Breathing Fishes*. (Academic Press, 1997).
71 Cass, A. N., Servetnick, M. D. & McCune, A. R. Expression of a lung developmental cassette in the adult and developing zebrafish swimbladder. *Evol Dev* **15**, 119-132, doi:10.1111/ede.12022 (2013).
72 Liem, K. F. Form and function of lungs: the evolution of air breathing mechanisms. *American Zoologist* **28**, 739-759 (1988).
73 Perry, S. F., Wilson, R. J., Straus, C., Harris, M. B. & Remmers, J. E. Which came first, the lung or the breath? *Comp Biochem Physiol A Mol Integr Physiol* **129**, 37-47 (2001).
74 Zheng, W. *et al.* Comparative transcriptome analyses indicate molecular homology of zebrafish swimbladder and mammalian lung. *PLoS One* **6**, e24019, doi:10.1371/journal.pone.0024019 (2011).
75 Funk, E., Lencer, E. & McCune, A. Dorsoventral inversion of the air-filled organ (lungs, gas bladder) in vertebrates: RNAsequencing of laser capture microdissected embryonic tissue. *J Exp Zool B Mol Dev Evol*, doi:10.1002/jez.b.22998 (2020).
76 Funk, E. C., Breen, C., Sanketi, B. D., Kurpios, N. & McCune, A. Changes in Nkx2.1, Sox2, Bmp4 and Bmp16 expression underlying the lung-to-gas bladder evolutionary transition in ray-finned fishes. *Evolution and Development* **in press** (2020).
77 Tatsumi, N. *et al.* Molecular developmental mechanism in polypterid fish provides insight into the origin of vertebrate lungs. *Sci Rep* **6**, 30580, doi:10.1038/srep30580 (2016).
78 Menke, D. B., Guenther, C. & Kingsley, D. M. Dual hindlimb control elements in the Tbx4 gene and region-specific control of bone size in vertebrate limbs. *Development* **135**, 2543-2553, doi:10.1242/dev.017384 (2008).
79 Nikaido, M. *et al.* Coelacanth genomes reveal signatures for evolutionary transition from water to land. *Genome Res* **23**, 1740-1748, doi:10.1101/gr.158105.113 (2013).
80 Zhang, W. *et al.* Spatial-temporal targeting of lung-specific mesenchyme by a Tbx4 enhancer. *BMC Biol* **11**, 111, doi:10.1186/1741-7007-11-111 (2013).
81 Brito, P. M., Meunier, F. J., Clement, G. & Geffard-Kuriyama, D. The Histological Structure of the Calcified Lung of the Fossil Coelacanth Axelrodichthys Araripensis (Actinistia: Mawsoniidae). *Palaeontology* **53**, 1281-1290, doi:Doi 10.1111/J.1475-4983.2010.01015.X (2010).
82 Hara, Y. *et al.* Shark genomes provide insights into elasmobranch evolution and the origin of vertebrates. *Nat Ecol Evol* **2**, 1761-1771, doi:10.1038/s41559-018-0673-5 (2018).
83 Onimaru, K. The evolutionary origin of developmental enhancers in vertebrates: Insights from non-model species. *Dev Growth Differ* **62**, 326-333, doi:10.1111/dgd.12662 (2020).
84 Frazer, K. A., Pachter, L., Poliakov, A., Rubin, E. M. & Dubchak, I. VISTA: computational tools for comparative genomics. *Nucleic Acids Res* **32**, W273-279, doi:10.1093/nar/gkh458 (2004).
85 Brudno, M. *et al.* Glocal alignment: finding rearrangements during alignment. *Bioinformatics* **19 Suppl 1**, i54-62, doi:10.1093/bioinformatics/btg1005 (2003).
86 Stanke, M. & Waack, S. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* **19 Suppl 2**, ii215-225, doi:10.1093/bioinformatics/btg1080 (2003).
87 Bian, C. *et al.* The Asian arowana (Scleropages formosus) genome provides new insights into the evolution of an early lineage of teleosts. *Sci Rep* **6**, 24501, doi:10.1038/srep24501 (2016).
88 Martin, K. J. & Holland, P. W. H. Diversification of Hox Gene Clusters in Osteoglossomorph Fish in Comparison to Other Teleosts and the Spotted Gar Outgroup. *J Exp Zool B Mol Dev Evol* **328**, 638-644, doi:10.1002/jez.b.22726 (2017).
89 Ballard, W. W. Stages and rates of normal development in the holostean fish, Amia calva. *Journal of Experimental Zoology* **238**, 337-354, doi:10.1002/jez.1402380308 (1986).
90 Komisarczuk, A. Z., Kawakami, K. & Becker, T. S. Cis-regulation and chromosomal rearrangement of the fgf8 locus after the teleost/tetrapod split. *Dev Biol* **336**, 301-312, doi:10.1016/j.ydbio.2009.09.029 (2009).
91 Marinic, M., Aktas, T., Ruf, S. & Spitz, F. An integrated holo-enhancer unit defines tissue and gene specificity of the Fgf8 regulatory landscape. *Dev Cell* **24**, 530-542, doi:10.1016/j.devcel.2013.01.025 (2013).
92 Gehrke, A. R. & Shubin, N. H. Cis-regulatory programs in the development and evolution of vertebrate paired appendages. *Semin Cell Dev Biol* **57**, 31-39, doi:10.1016/j.semcdb.2016.01.015 (2016).

93 Hörnblad, A., Langenfeld, K., Bastide, S., Langa Vives, F. & Spitz, F. Dissection of the Fgf8 regulatory landscape by in vivo CRISPR-editing reveals extensive inter- and intra-enhancer redundancy. *bioRxiv*, 2020.2003.2003.966796, doi:10.1101/2020.03.03.966796 (2020).
94 Boudinot, P. *et al.* A tetrapod-like repertoire of innate immune receptors and effectors for coelacanths. *J Exp Zool B Mol Dev Evol* **322**, 415-437, doi:10.1002/jez.b.22559 (2014).