Supplementary information for:

The genome of the platyfish, *Xiphophorus maculatus*

Manfred Schartl[1,2*+], Ronald B. Walter[3+], Yingjia Shen[3], Tzintzuni Garcia[3], Julian Catchen[4], Angel Amores[4], Ingo Braasch[1,4], Domitille Chalopin[5], Jean-Nicolas Volff[5], Klaus-Peter Lesch[6], Angelo Bisazza[7], Pat Minx[8], LaDeana Hillier[8], Richard K. Wilson[8], Susan Fuerstenberg[9], Jeffrey Boore[9], Steve Searle[10], John H. Postlethwait[4] and Wesley C. Warren[8*]

[1]Physiological Chemistry, University of Würzburg, Biozentrum, Am Hubland, and [2]Comprehensive Cancer Center, University Clinic Würzburg, Josef Schneider Straße 6, 97074 Würzburg, Germany, [3]Department of Chemistry and Biochemistry, 419 Centennial Hall, Texas State University, 601 University Drive, San Marcos, TX 78666, USA. [4]Institute of Neuroscience, University of Oregon, 1425 E. 13th Avenue, Eugene, OR 97403 USA, [5]Institut de Génomique Fonctionnelle de Lyon, Unité Mixte de Recherche 5242, Centre National de la Recherche Scientifique/Université de Lyon I/ Ecole Normale Supérieure de Lyon, 46 allée d'Italie Lyon, France, [6]Division of Molecular Psychiatry, Department of Psychiatry, Psychosomatics and Psychotherapy, University Clinic Würzburg, Füchsleinstraße 15, 97080 Würzburg, Germany, [7]Department of General Psychology, University of Padua, Via Venezia 8, 35131 Padua, Italy, [8]The Genome Institute, Washington University School of Medicine, 4444 Forest Park Blvd., St Louis, MO 63108, USA, [9]Genome Project Solutions, 1024 Promenade Street, Hercules, CA, USA, [10]European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton Cambridge CB10 1SD, UK.

+These authors have contributed equally to the work

*Corresponding author

Supplementary Material

Table of Contents
Supplementary Notes
Supplementary Tables 1 to 15
Supplementary Figures 1 to 9
Supplementary References

Table of Contents

SUPPLEMENTARY NOTE

Genome sequencing and assembly

Sequencing: The platyfish genome was sequenced using deep sequencing methods supplied through the Roche 454 and Illumina technologies. For this approach fragment and long insert paired libraries of 3 and 20kb were prepared according to Roche recommended methods. Illumina paired end libraries of insert size 200bp and 75 base pair read length were created according the manufacturers recommendations. In addition to the shotgun sequencing strategy a physical map indicating tiling paths of *Xiphophorus maculatus* contigs was constructed by generating fingerprints from the WLC-1247 BAC library. The physical map consists of 43,192 BAC clones.

Assembly: To assemble the platyfish genome first two independent assemblies were built using the Newbler and PCAP [1] algorithms from ~19.6X total sequence coverage. Then the Newbler and the PCAP assemblies were merged. As a result of this merge a total of 23,144 contigs were added to the Newbler assembly (~7Mb), thus increasing the final contig and supercontig length. Redundant reads found in contigs were removed and then all contigs were renamed. The final merged assembly (Supplementary Table 1), referred to as platyfish 4.4 contained 130,963 contigs with an N50 contig and supercontig length of 21kb and 1.1Mb, respectively (Supplementary Figure 8). The N50 contiguity statistic denotes the percentage of the assembled genome that is of that base length or greater. Platyfish 4.4 was screened for sequence contamination from other organisms by our group and NCBI that resulted in the removal of 410 contigs. A total of 669Mb was assembled in contigs. Previous estimates of genome size from flow cytometry range from 750 to 950 Mb [2,3].

Assembly consensus base error correction: The pyrosequencing chemistry method used by 454 Titanium instruments causes false insertions and deletions within homopolymer regions of the genome. To correct these, we generated from the same DNA source used for the reference assembly 101 million Illumina reads (75 base paired-end reads, insert size 200bp). These reads were trimmed using stringent criteria for sequence quality (less than 0.01 chance of error, containing no more than 2 ambiguous nucleotides, retaining a length after trimming of at least 20 bases). This retained a dataset of 94 million reads, for an average coverage of 7.4x of the assembled genome. To each of the contigs that comprised the set of 454 supercontigs, we aligned all trimmed reads using the Genomics Workbench v.4.03 software (CLC Bio). For any contig shorter than 8 kb, we included in these alignments the contigs themselves to ensure all regions were covered by the reference alignment by at least one read. Software limitations precluded this for contigs 8kb or longer, so these were shredded *in silico* into fragments of 60 bases, which were then included in the reference alignments for the same purpose along with the Illumina reads. A consensus sequence was then created that factored the quality scores of both the 454 assembly and the individual Illumina reads. Then these contigs were joined back together into the supercontigs as originally structured by joining them in the same order and orientation and adding back the same number of N's in the gaps as had been inferred for the supercontigs by the paired-end information. Based on manual examination of a small subset of the data, this process appears to have corrected about 373,000 bases within the Platyfish 454-based 4.4 assembly, mostly deletions in homopolymeric regions.

Assembly accuracy: To examine putative misassemblies due to false de novo joins during assembly graph construction we utilized a high density meiotic map described below in the section entitled: Assigning genome contigs to the genetic map. Using 14,391 marker sequences, we could reliably align 1,950 scaffolds to all linkage groups. Of these, 231

scaffolds mapped to multiple linkage groups, suggesting a misassembly event and were manually split. A total of 576 splitting events occurring on 231 of these scaffolds increased the final number of scaffolds to 2,288. The great majority of the splitting events, 163 of them, were split into two pieces, with a much smaller number of cases had to be split into multiple pieces. Overall these detected misassembly events (11%) are similar in scope and structure to what we have observed for other large genome projects.

Transcriptome sequencing and annotation

The *X. maculatus* transcriptome (version_4) was assembled from RNA isolated from tissues including heart, liver, brain, ovaries, and testes, as well as from embryonic stages 15 and 25. The raw reads were filtered using a custom filtration method[4] which resulted in 265,833,281paired-end reads. The short reads were then aligned to the genome contigs using Bowtie[4] [5], then assembled using the Velvet/Oases package by testing all odd k-mer sizes from 31 to 59 nt. A broad peak was observed in the N50 scores, when plotted as a function of k-mer size, for the assembled sequences over 500bp in the k-mer size range of 31 to 39bp. The k-mer size of 37 was on the upper end of this range before a more significant drop-off began so it was selected as a representative assembly. To begin to remove likely erroneous transcripts, short reads were mapped to the 495,520 assembled transcripts and only sequences with 5 or more mapped reads were kept. The resulting sequences contained 151,079 transcripts with an N50 of 2,978 bp and average length of 1705 bp. The transcriptome is available in genome browser format at http://avogadro.tr.txstate.edu/cgi-bin/gb2/gbrowse/XM_ncbi442/ and as a FASTA bulk sequence file at http://avogadro.tr.txstate.edu/Xiph_data_link/stable/Xm_transcriptome_v4.0/.

For the *X. hellerii* transcriptome, RNA from 1 month old whole fish, and from brain, liver, ovaries and testes of mature fishes was sequenced. After custom filtration [4], 173 million paired reads and 22 million singletons were used for transcriptome development. We then used Velvet [6] to guide the assembly using combined paired-end and singleton reads. We first used all odd k-mer sizes from 21 bases to 49 bases and compared assemblies produced from different k-mer sizes to identify the assembly with the longest N50 length. In the final result, we used a hash length (k-mer size) of 35 bases. We employed Oases (http://www.ebi.ac.uk/~zerbino/oases/) to perform the final assembly and reporting of putative transcripts and splice variants using a coverage cutoff of 4, an insert length estimate of 120, and other parameters at default values. The final assembly has 42,675 transcripts with an N50 of 3,280bp, an average length of 1,991bp and a total size of 483Mb.

Gene Models and annotation using PHRINGE

Gene identification analysis was performed within all corrected supercontigs. The primary method relied on evidence from large amounts of RNA-seq data. Reads from the transcriptome sequencing were trimmed to eliminate all reads with any ambiguous base calls (e.g., N) and any that failed to meet the prescribed length as defined by the number of cycles. The next step at gene finding requires sets of data of identical read length. These trimmed reads were then aligned to the sequence-corrected supercontigs using Bowtie [5], then this alignment was adjusted for the most likely exon-intron boundaries using TopHat [7], and then gene models created using Cufflinks [8]. This process created a large number of potential transcript sequences, ranging from 33,374 to 90,314 among the various conditions analyzed (Supplementary Table 13). Only those transcripts containing a complete ORF and a transcript read coverage of at least 3x were retained, and these were reconciled into a single set of

33,756 unique potential protein-encoding genes. Additional evidence was found by *ab initio* modeling using Augustus [9] that had been trained on the medaka gene set and on the alignment of full-length gene models of medaka and zebrafish (both from Ensembl) using BLATX [10]. All of these models can be viewed as separate tracks in genome browser format at http://avogadro.tr.txstate.edu/cgi-bin/gb2/gbrowse/XM_ncbi442/.

These gene models were further culled to a subset of 17,783 that are especially reliable and amenable to phylogenetic analysis for entry into a whole genome evolutionary interpretation using the PHRINGE (Phylogenetic Resources for the Interpretation of Genomes) system (http://genomeprojectsolutions.com/PHRINGE_pipeline.html) by eliminating any transcripts shorter than 300 nucleotides and retaining only the longest version of any splice variant at each locus. PHRINGE creates a graph with all inferred protein sequences as nodes, with edges formed from distance scores calculated from their full length alignments, then PHRINGE clusters these sequences into gene families using a method that considers the evolutionary relationships among the organisms (Supplementary Figure 9), then performs a phylogenetic analysis for each cluster (Supplementary Table 14). A separate analysis is conducted at each node of the tree of these species (called a "cluster level") so that the user can choose to see the sets of gene relationships at only the more shallow levels, where it may be more accurate, or at the deeper levels, where it will be more comprehensive. This procedure allows the most accurate possible assignment of orthologous and paralogous relationships, and reconstruction of gene duplications and losses. Users can see the multiple sequence alignments and phylogenetic trees of all genes, search using keywords, compare intron-exon structures, and see the relative arrangements of homologs across all genomes in the PHRINGE database. This comparison of the culled subset of 17,783 *Xiphophorus* genes with gene sets of 11 other animals is available at http://xiphophorus.genomeprojectsolutions-

databases.com/ and as FASTA bulk files at http://avogadro.tr.txstate.edu/Xiph_data_link/stable/Xm_JB_gene_models/.

From *Danio rerio* 28,630 genes, from *Gasterosteus aculeatus* 27,576 genes, from *Tetraodon nigroviridis* 27,918 genes, from *Oryzias latipes* 24,661genes, and from *Xiphophorus maculatus*17,783 genes were entered into the PHRINGE analysis. Of the *Xiphophorus* genes, PHRINGE found a total of 15,676 to be clearly homologous to one or more genes of another considered organism.

The comparison that included only the gene sets of *Xiphophorus maculatus* and just its closest considered relative *Oryzias latipes* (cluster level 11) identified 11,914 gene families. These 11,914 gene families contain 13,384 *Xiphophorus* genes (from 1 to 40 genes in each family) and 16,973 *Oryzias* genes (from 1 to 37 genes in each family except for a single family with 175 members). Of the 11,914 gene families, 8,280 gene families have exactly 1 gene in each of the two species. We consider these as most likely 1:1 orthologs.

By way of comparison, the analysis that includes only the genes sets of the *Gasterosteus aculeatus* gene set and its closest considered relative *Tetraodon nigroviridis* (cluster level 10) identified 16,238 gene families. This contains 23,609 *Gasterosteus* genes (from 1 to 84 genes in each family except for a single family with 244 members) and 18,945 Tetraodon genes (from 1 to 38 genes in each family). Of the 16,238 gene families, 10,207 gene families have exactly 1 gene in each of the two species the culled subset of 17,783.

At a deeper level (cluster level 9), the comparison that includes the gene sets of all four fish, with a clade of *Xiphophorus maculatus* plus *Oryzias latipes* being reciprocally the outgroup to the clade of *Gasterosteus aculeatus* plus *Tetraodon nigroviridis*, there are 22,249 gene families identified. This comparison contains 12,847 *Xiphophorus* genes (from 0 to 32 genes in each family), 19,910 *Oryzias* genes (from 0 to 56 genes in each family), 22,171

Gasterosteus genes (from 0 to 38 genes in each family except for three families with 70, 120, and 157 members), and 14,488 *Tetraodon* genes (from 0 to 53 genes in each family). Of the 22,249 gene families, 10,836 have a homolog in *Oryzias* and one or more of *Gasterosteus* and *Tetraodon* but not in *Xiphophorus*. Only 239 gene families have exactly 1 gene in each of the four species, perhaps owing to the rampant and differing gene losses after whole genome duplications.

Gene annotation using Ensembl genebuild

Assembly Xipmac4.4.2 (GenBank Assembly ID GCA_000241075.1; http://www.ncbi.nlm.nih.gov/genome/assembly/?term=GCA_000241075.1) was annotated for protein coding genes using the Ensembl genebuild procedure. Gene models were based on 1.) Genewise alignments of UniProt protein sequences from platyfish, 2.) models build from platyfish RNASeq data using the Ensembl RNASeq pipeline, 3.) Genewise alignments of UniProt protein sequences from other fish and vertebrate species, and 4.) exonerate alignments of Ensembl Stickleback and Zebrafish proteins from Ensembl release 65. The protein-coding models were extended into their untranslated regions using RNASeq models. In addition to the coding transcript models, non-coding RNAs and pseudogenes were annotated. This resulted in 20,366 protein-coding genes, 348 non-coding genes and 28 pseudogenes, giving a total of 20,817 gene transcripts. The annotated platyfish genome can be found athttp://www.ensembl.org/Xiphophorus_maculatus/Info/Index.

Estimation of gene number by transcriptome similarity

To estimate the number of protein coding genes in the platyfish genome, the Jp163A reference transcriptome was analyzed by performing reciprocal best-hit BLAST comparisons against assembled transcript sequences from *O. latipes*, *G. aculeatus*, *T. nigroviridis*, *T. rubripes*, *D. rerio*, and *H. sapiens* retrieved from the Ensembl database. This analysis produced a BLAST reciprocal best-hit (RBHB) reference library containing 20,105 cDNA sequences. This approach allows several representations of some genes, since the transcript assembly process may construct alternative transcript forms that may all have best hits to similar sequences in any of the other species. To control for this, we grouped alternatively assembled sequences that were reported by Oases[11] to have a common origin (i.e., the same "locus" number), and then only included the unique 'locus' numbers from a group. This grouping produced an estimated 15,431 unique genes.

Estimation of novel, platyfish-specific genes

We searched the *X. maculatus* genome for evidence of novel, active genes or those that may not be well enough conserved for identification using homology-based searches. We compared the assembled Jp163A transcriptome against medaka and stickleback transcriptomes as reference sets and created a set of transcripts found in both *X. maculatus* and either medaka or stickleback transcriptomes. We employed BLAST [12,13] searches in amino acid space using TBLASTX to detect divergent yet related sequences at a similarity threshold of E $<10^{-10}$ in both directions. This search identified 1,638 stickleback and 1,436 medaka sequences that did not have hits and thus may be absent from the *X. maculatus* transcriptome. In addition this search produced approximately 70,000 *X. maculatus* contigs that did not match either the medaka or stickleback reference set. This quantity of apparently

unique *X. maculatus* sequences is improbably large and may contain non-coding RNAs, misassemblies, fragments, or other artifacts that do not represent novel, active protein-coding genes. The reference set of similar sequences (RSS) included 81,188 *X. maculatus* transcripts similar to medaka and 80,761 transcripts similar to stickleback (many of these overlap) based on the BLAST hits.

To reduce the impact of artifact sequences, we grouped the *X. maculatus* sequences to create clusters that represented all the sequence alternatives for a common gene product. To group the *X. maculatus* sequences together, we performed a self-BLAST of the entire set of *X. maculatus* transcripts using TBLASTN with an E-value threshold of 10^{-10}. We then clustered sequences based on hits in these results. Any hit with an identity of 95% or greater was added to a similarity cluster with the hit sequence. Once the self-similar clusters were created, we sorted them based on whether the cluster was represented in the RSS. Sequences from any cluster with at least one *X. maculatus* sequence that hit a medaka or stickleback reference sequence were all taken to be related to that reference sequence, and were not considered candidates for novel, active *X. maculatus* genes. The *X. maculatus* clusters with no hits to medaka contained 31,286 sequences in 27,236 clusters and those with no hits to stickleback contained 30,444 sequences in 26,730 clusters. Clusters with hits to both medaka and stickleback contained 78,269 sequences in 24,550 clusters. Clusters with no hits to either set contained 28,487 sequences in 25,384 clusters; this is the "no hit" pool most likely to include novel, active *X. maculatus* genes.

To filter out sequences that were not likely to represent protein-coding transcripts we mapped all transcripts from the "no-hit" pool to the *X. maculatus* genome and predicted likely coding sequence regions. We used the GMAP program [14] to align each of the unmatched *X. maculatus* sequences to the genome. GMAP detects the location and boundaries of exons and

predicts a coding region. We used this to further screen the sequences that mapped to the genome by keeping only those transcripts that had a total coding sequence >300 bases and >99% of the transcript directly mapped to the genome. This filtration regimen reduced the candidate pool to 5,570 putative protein-coding transcripts with no match to medaka and 5,066 with no match to stickleback. 4,313 of these *X. maculatus* sequences are different from both medaka and stickleback, and those formed 3,964 clusters; these sequences are candidates to be novel, active genes.

We BLASTX searched the above 4,313 transcripts against the NCBI NR database and found hits for 639 sequences at an E-value threshold of 10^{-10}. Of these 639 sequences, 122 had been removed from GenBank due to "standard genome annotation processing" and 73 were duplicates. The remaining 404 NCBI NR hits gave 315 that carried at least one uncertainty term in its description (e.g., hypothetical, unnamed, or novel predicted). Thus, after removing the 639 sequences from the 4,313 *X. maculatus* sequences different from both medaka and stickleback, 3,674 transcripts in 3,416 clusters remained as an estimated pool of novel sequences in the *X. maculatus* transcriptome.

However, some of these sequences might represent parts of the same transcript that had not been joined properly during the transcriptome assembly. We therefore sought to cluster the remaining 3,674 transcriptome contigs differently, by physical distance in the genome. We examined inter-genic and inter-exonic distances in zebrafish, medaka, and stickleback (gene annotations retrieved from Ensembl, release 67) and found that, generally, 48-82% of inter-exonic distances were 1kb or less, but in all three fish, some distances were well over 100kb. Inter-genic distances, however, were such that 17% of neighboring zebrafish genes fall within 1kb of one another and 30% of neighboring genes are within a 2kb range. Thus, we selected a distance of 1kb to cluster the remaining *X. maculatus* sequences into likely gene groups, since

at the 1kb range the grouped sequences are more likely to be fragments of the same gene than they are to represent separate genes. This reasoning was further supported when the members of each cluster contained sequences with similar locus numbers as assigned by the Oases assembler. Similar locus numbers indicate the sequences are closely related by links in the DeBrujin graph representation, but may have been connected by a low-coverage region. Of the 156 clusters with 2 or more sequences, 136 had a locus number range of 3 or less. The quantity of clusters of different sizes (i.e., sequences per cluster) at a 1kb clustering distance is shown in Supplementary Table 15. In total, 3,482 distance-clusters were formed, with 156 having more than one member. This quantity is similar to the 3,416 clusters formed earlier through sequence similarity, and thus serve to corroborate the results.

Non-coding RNAs

A total of 1,464 non-coding RNAs were annotated by software prediction (Supplementary Table 11). Approximately half of the rRNAs, snRNAs and snoRNAs species could be confirmed from the transcriptome. 611 micro-RNAs were predicted. We identified 535 tRNA genes that represent 49 different anticodon loops. The number of tRNA genes per codon ranges from 1 to 137 (Supplementary Table 12).

Transposable elements

For annotation of transposable elements (TE) first a combined library containing 7257 sequences from both manual (119 sequences, from 366 to 20,400 nucleotides) and automatic annotation (7138 sequences, from 50 to 2,835 nucleotides) was produced. This inventory masked about 16% of the genome. Within the repeat fraction TEs make up about 5% of the

genome of *X. maculatus,* in accordance with a relatively compact genome (Supplementary Tables 2-3). This TE content is only marginally higher than the values from the compact genomes of pufferfishes and is in the range of the chicken [15,16]. Despite a high diversity of TE families, none of them is highly repeated in the genome, usually in the range of 1-10 complete copies. One of the highest repeated elements is the non-autonomous MIToy (more than 500 copies). A considerable fraction of these elements share more than 97% nucleotide (nt) similarity, suggesting a recent transposition activity of this element.

The foamy virus sequence of the platyfish was initially detected on the sex chromosome. The sequence includes LTRs and specific Gag, Pol and Env ORFs, with a total length of 17,027 nucleotides. At least two complete copies were identified in this region, with more than 95% sequence identity. The whole genome contains more than 30 copies (>85% identity), but only the copies in one region of the sex chromosome are complete. Looking more precisely at each copy and solo-LTR, a TG preference for insertion was determined. Searching for possible endogenous foamy virus sequences in other vertebrate genomes was performed on the Ensembl versions of the chicken, zebrafinch, green anole, Xenopus, and all teleost genomes. A new sequence was identified in the cod genome (contig 24163). The predicted protein corresponds to a foamy reverse transcriptase and clearly groups within the foamy branch. A more complete foamy virus sequence could not be identified because of short contig lengths in the cod genome.

Construction of a high-density meiotic map and anchoring of sequence contigs

To obtain a high density meiotic map the RAD-tag methodology was used from a mapping cross DNA panel consisting of 267 individuals. The sequence for each of the initial set of 18,119 polymorphic RAD-tags was mapped onto sequenced genome contigs and each of the

markers was annotated by the Synteny Database [17]. In addition, the set of sequences surrounding microsatellites from a previous genetic map [18] were identified in genomic contigs and those contigs containing both mapped microsatellites and mapped RAD-tags were associated to the RAD-tag map, thus providing anchor markers. Because JoinMap couldn't initially handle our huge number of markers, the data set was partitioned into two overlapping subsets, with each set containing the common anchor markers. The Kosambi mapping function and JoinMap's maximum likelihood mapping algorithm grouped markers at an initial LOD threshold of 15.0. After this initial mapping of markers to individual LGs, markers were divided into 3 groups of 8 different LGs each. Subsequent linkage mapping analysis for individual LGs was performed at a LOD=30.0. Suspicious double recombinants were reevaluated by visually inspecting reads in Stacks and manually correcting of the genotype if necessary.

Assigning genome contigs to the genetic map

After exclusion of 1794 conflicting markers a total 14391 markers were mapped. Markers from the genetic map were aligned against the platyfish genome by using GSnap [19] to align the nucleotide sequences of mapped RAD-tag markers against the genome contigs. Contigs that contained a single mapped RAD-tag marker were ordered along the chromosome, but with a single anchor point, their orientation was ambiguous and thus their orientation was randomly selected. When more than one adjacent mapped RAD-tag marker hit the same contig that contig was not only ordered on the chromosome but its orientation along the chromosome was certain. When two genetic markers from independent map positions aligned to the same scaffold, causing the map order to disagree with the physical order, we prioritized for map order. This type of discrepancy could be caused by two possible reasons: 1) there is a genome misassembly, or 2) we failed to detect a crossover event. The latter reason is very unlikely due to the rare occurrence of nearby double crossovers. Therefore, the scaffold

was broken into pieces so that the physical order will agree with the genetic map. To break the scaffold, the nearest contig boundaries to the genetic markers within the scaffold were identified (since scaffolds are composed of contigs and gaps) and the scaffold was broken at the contig scaffold boundary. 101 scaffolds mapped to multiple linkage groups. 231 scaffolds were split (total 576 splitting events; 2 fragments: 163 cases; 3 fragments: 44 cases; 4 fragments: 16 cases; 5 fragments: 2 cases; 6 fragments: 4 cases; 7 fragments: 2 cases). Manual examination of splitting causes revealed that most were due to repeat structures allowing long read pairs to erroneously connect contigs. In total 2288 scaffolds were mapped from 1950 original scaffolds resulting in 653,124,558bp total map length.

After ordering the physical genome using the genetic map, we consecutively enumerated nucleotides from the ordered contigs along each chromosome, and then aligned each assembled gene transcript to its genomic location using BLAT [20] and recorded the maximal start/end coordinates for the gene. Orthologs of each gene were called by the algorithms of the Synteny Database as described [17].

Analyses of viviparity genes

Thirty-four protein-coding genes known to function in yolk production, placenta-related characteristics, and zona pellucida structures were selected as candidate genes that may be involved in the evolution of viviparity among *Xiphophorus* fishes (Supplementary Table 5). For viviparity genes, our selection criteria for gene candidates were as follows: a) The gene is known to be involved in nutrient provision during early development in oviparous vertebrates and invertebrates; b) The gene has been identified as coding for an egg envelope protein in closely related but oviparous fishes; c) The gene is documented to be involved in mammalian placental development; d) The gene has been proposed to be a driving force in the evolution of placentation. Eighteen randomly selected genes (Supplementary Table 5) but which to our knowledge had not been associated in any study so far with viviparity were used for control.

Orthologous sequences for the 34 candidate genes and the 18 control genes from four fish species (*O. latipes*, *G. aculeatus*, *T. nigroviridis*, *D. rerio*) were retrieved from the Ensembl database. To identify orthologs in *Xiphophorus*, the cDNAs and genomic sequences to medaka orthologs were searched against the Jp163A transcriptome or genome using TBLASTX (E-value cutoff 10^{-30}). Reciprocal BLAST searching was performed for each candidate gene to ensure each *Xiphophorus* and medaka homologue were indeed the best hits. We then used the MAFFT translation alignment (Algorithm G-INS-I) in the Geneious software package (www.geneious.com/) to make codon-delimited alignments for each significant *Xiphophorus* hit and the other four fish species. PAML (version 4.4, linux 64bit) was implemented to detect positive selection in genes and sites along the *Xiphophorus* lineages (i.e., *X. maculatus* and *X. hellerii*) [21]. The likelihood of alternative hypothesis (genes positively selected; model = 2, NSsites = 2) was compared with a null model (fix_omega=1 and omega=1), where no positive selection is allowed, using likelihood ratio tests (*chi-square* distribution, df=1) and *Xiphophorus* sequences were used as foreground. Genes with p-value less than 0.05 from likelihood ratio tests were designated as positively selected in *Xiphophorus* and the Bayes empirical Bayes method [22] was further used to calculate the selection pressure at each site. Eight of the 34 viviparity-related genes showed significant values for evolution under positive selection, while none of the 18 control genes showed signs of positive selection (Supplementary Table 5). Genes identified with positive selection were searched further using interProscan for functional motifs. To profile expression of viviparity genes in *X. maculatus* ovary and liver, we sequenced 60mer pair-end reads using Illumina GAIIx sequencer. 22.3 and 32.5 million RNA-seq reads were then mapped to *X. maculatus* genome using Bowtie [5] and FPKM (Fragments Per Kilobase of transcript per Million mapped reads) values were calculated using Cufflink 1.30 [7]. Except for 4 placenta genes (*cdx1a, cdx4, gcm1, mash2*), which did also not show any sign of evolution under positive selection, expression values were in accordance with a proposed viviparity function (Supplementary Table 5).

Post-TGD analysis of gene families

To study whether the platyfish and other teleosts have retained specific categories of genes that are possibly involved in bringing about the highly complex suites of behavior noted in many fishes we established a list of cognition-related genes (Supplementary Table 7). Only those genes were included that are assumed (from human and mouse data) to be involved in brain development and core cognitive functions and - if dysregulated or mutated in neurodevelopmental, psychiatric or neurodegenerative disease – are connected to cognitive dysfunction, or have been shown by functional analysis to be critically involved in neural network connectivity and synaptic plasticity. To produce such a list of genes primarily involved in cognition, the following criteria were used: We heuristically considered genes as candidates for cognition, if at least two of the following four criteria were met: Category A, 1) the gene is known to be involved in the development of brain networks implicated in cognition (e.g. proliferation, migration and differentiation of glial and/or specific neuronal cells); 2) the gene has been demonstrated to participate in synapse formation and activity-dependent remodeling and/or in transsynaptic signaling (e.g. long-term potentiation/depression) and adult brain function, such as regulators of neurotransmitter systems relevant to cognition circuits (e.g. glutamate signaling); Category B, 3) the gene was validated as a modulator of cognitive processes in neuropsychological and psychophysiological paradigms or functional neuroimaging; and 4) the gene has been repeatedly identified by genome-wide screening approaches as a candidate gene for disorders featuring cognitive impairment (e.g. intellectual disabilities), deficits in social cognition (e.g. autism and schizophrenia spectrum disorders) or neurodegenerative disorders with cognitive dysfunction (e.g. Alzheimer's disease, Parkinson's disease). Genes were further prioritized when cognitive function was altered in a gene-targeted mouse or other animal model. Although the distinction is prone to a certain degree of arbitrariness, genes shown to be involved in other domains of brain function and complex behavior, for example those primarily moderating emotionality or reward-related behavior, were largely excluded. The gene lists for pigmentation and liver were derived from Braasch et al. [23] (and updated using criteria therein). The overlap of the three gene categories (cognition/pigmentation/liver) and their TGD paralog retention rates in Xiphophorus, zebrafish, Atlantic cod, stickleback, *Tetraodon*,

Takifugu, medaka, and the parsimony-inferred teleost ancestor are depicted in Supplementary Figure 6.

.

Potential biases in gene categories accounting for TGD paralog retention

We tested whether the differences in TGD paralog retention rate among the three functional categories (cognition/pigmentation/liver genes) could be accounted for by a bias in categories toward functional and molecular features that were shown previously to be enriched among genes retained after whole genome duplications (WGDs).

Dosage sensitivity: According to the dosage balance hypothesis [24] genes that are dosage sensitive tend to be retained after WGDs, as for example found for retained ohnologs from the two rounds of early vertebrate genome duplication [25]. The non-occurrence of copy number variation (CNV) of a gene is considered to indicate dosage-sensitivity [25]. Here we tested for the occurrence of CNVs in a post-TGD lineage (zebrafish) and, as a proxy to the pre-TGD condition, a lineage that diverged before the TGD (human). Following the strategy of Makino and McLysaght[25] we obtained copy number variation (CNV) data for human genome assembly GRCh37/hg19 from the Database of Genomic Variants (http://dgvbeta.tcag.ca/dgv/app/home; release date 2012-03-29; 462,611 CNV regions). Furthermore, we obtained more limited CNV data for zebrafish from Brown et al. [26] (4,852 CNV regions, lifted from genome assembly Zv8 to Zv9). Genomic coordinates of human and zebrafish protein-coding genes were obtained from Ensembl. We then identified those human and zebrafish "CNV genes" within our dataset (Supplementary Table 8) defined as those genes that were completely covered by one or more CNV regions [25]. For both, human and zebrafish, there was neither a significant difference in the amount of "CNV genes" between singletons and retained TGD paralogs in our dataset nor among the three different functional categories (cognition/pigmentation/liver genes) (chi-square tests).
Protein complex membership: A bias for retained ohnologs in terms of protein complex membership was previously noted [25]. A list of human protein complex members was obtained from Human Protein Reference Database (www.hprd.org; 1,521 protein complexes). In our dataset (Supplementary Table

9), there was no difference in protein complex membership between singletons (117/315 = 37.1%) and TGD paralogs (74/176 = 42.0%) (chi^2 = 1.142, d.f. = 1, p = 0.28523), or among the three functional categories, cognition (80/192 = 41.7%), pigmentation (40/129 = 31.0%), and liver (76/185 = 41.1%) genes (chi^2 = 4.37, d.f. = 2, p = 0.11248).

Protein length: Sato et al. [27] found in their dataset a significant bias of TGD-retained paralogs encoding long (>1,000 amino acids) rather than short (<200 amino acids) proteins using human proteins as proxies to the ancestral state. In our dataset (Supplementary Table 10), we found a similar trend comparing TGD-retained paralogs and singletons (chi^2 = 10.711, d.f. = 2, p = 0.00472) (Supplementary Fig. 7a). Furthermore, there is a significant difference in the distribution among protein length groups between the three functional categories (chi^2 = 14.132, d.f. = 4, p = 0.00688) based on the high portion of cognition gene belonging to the long protein category (Supplementary Fig. 7b). This suggests that the high TGD paralog retention rate of the cognition genes may be based on their bias toward long proteins. Of note, there is no statistical difference between pigmentation and liver genes with respect to their protein length distribution (chi^2 = 0.382, d.f. = 2, p = 0.82613), despite the fact that significantly higher retention of TGD paralog retention rate for pigmentation compared to liver genes [28].

SUPPLEMENTARY TABLES

Supplementary Table 1: Assembly statistics for the platyfish genome

Assembly metric	Contigs	Scaffolds
Total	130,963	84,533
Total bases	669 Mb	729 Mb[a]
Maximum length	203,919	7,293,446
N50 length[b]	21,642	1,102,127
N50 number	8843	155

[a] Base size estimate includes gaps

[b]N50 statistic represents 50% of the genome assembly that is the defined length or longer.

Supplementary Table 2: Diversity of transposable elements in fish

Species	Pufferfish (*T. rubripes*)	Pufferfish (*T. nigroviridis*)	Zebrafish (*D. rerio*)	Platyfish (*X. maculatus*)
Retrotransposons (Class I)				
LINE				
Restriction enzyme-like				
NeSL/Zebulon	X	X	X	-
R2	-	-	X	X
R4/Rex6	X	X	-	X
Apurinic/apyrimidic				
LINE1/TX1	X	X	X	X
RTE/Rex3	X	X	X	X
I/Bgr	F	X	X	-
LINE2/Maui	X	X	X	X
LINE3/CR1	X	X	X	X
Rex1/Babar	X	X	X	X
Nimb	ND	ND	X	ND
Hero	ND	ND	X	X
SINE	**X**	**ND**	**X**	**X**
LTR				
Ty3/Gypsy				
SURL	X	X	X	-
SURL-like	X	X	X	-
Jule	X	X	X	X
CsRn1	X	X	X	-
Sushi	X	X	X	X
Barthez	X	X	X	X
Gmr1	X	X	X	-
Rex8	X	X	X	X
Osvaldo	ND	ND	X	-
Ty1/Copia	X	X	X	-
BEL	X	X	X	X
Retroviruses	X	X	X	X
DIRS1	**X**	**X**	**X**	**X**
Penelope	**X**	**X**	**X**	**X**
DNA Transposons (Class II)				
Subclass I				
TIR				
Tc1-Mariner	X	X	X	X
hAT	X	X	X	X
Harbinger	X	ND	X	X
EnSpm	X	X	X	-
P	X	X	X	-
PiggyBac	X	-	X	X
MuDR	-	X	X	X
ISL2EU	ND	ND	X	-
MITE	ND	ND	X	X
Crypton	**-**	**-**	**-**	**-**
Subclass II				
Helitron	**-**	**-**	**X**	**X**
Mavericks/Polintons	**ND**	**ND**	**X**	**X**

The presence (X) or absence (-) of transposable element families is indicated for four fish species, two pufferfishes (*Takifugu rubripes* and *Tetraodon nigroviridis*) (data based on [29]), the zebrafish (Danio rerio) (data based on [29]) and the platyfish (*Xiphophorus maculatus*). Presence of transposable elements indicates that at least a trace or a fossil of the element has been detected. For some families, no data (ND) could be found. The two classes, retrotransposons and DNA transposons, are written in red, orders in green, superfamilies in blue and families in black.

Supplementary Table 3: Comparison of the genome size and the percentage of transposable elements (TEs) in different vertebrate species

Species	Genome size (Mb)	Percentage of TEs (%)	References
Takifugu rubripes Torafugu	400	2.7	[16]
***Xiphophorus maculatus* Platyfish**	**900**	**5**	
Oreochromis niloticus Nile tilapia	1200	14	[30]
Gallus gallus Chicken	1220	4.3	[15,31]
Danio rerio Zebrafish	1700	15-20	[32]
Mus musculus Mouse	2900	38.2	[33]
Salmo salar Atlantic salmon	3000	30-35	[34]
Xenopus laevis Clawed frog	3100	37	[35]
Homo sapiens Human	3400	44.8	[36]

Supplementary Table 4: Location of pigmentation genes on X chromosome.

Pigmentation genes located on the X chromosome (LG21) were identified during the course of the post-teleost genome duplication paralog retention analysis and the position of the respective scaffold in the genetic map-based genome assembly was determined.

gene	platyfish gene ID*	Linkage group	start	end	orien-tation	transcript ID*	scaffold	map position
asip2a	-	group21	4772651	4772764	-1	-	JH556882.1	LG21: 5.8-8.1 cM
egfrb/xmrk	G0266747	group21	21380952	21381682	1	T0024005	JH558217.1	LG21: 71.2 cM
myca	G0327718	group21	15128501	15132049	-1	T0029430	JHP00189.0	LG21: 35.8-42.7 cM
rps20	-	group21	4925435	4925789	1	-	JH557038.1	LG21: 8.1-8.9 cM
tfap2a	G0209431	group21	7585659	7586219	1	T0018927	JH557150.1	LG21: 15.1 cM
muted	-	group21	6925178	6925554	-1	-	JHP00191.0	LG21: 10.8-14.3 cM

*Genes without gene/transcript ID not represented in the transcriptome and were identified by tblastn searches against the genome assembly using medaka/stickleback proteins as query sequences.

Supplementary Table 5: Viviparity related genes tested for positive selection[a]

Gene	category	Species used[a]	alignment length(starting position[b])	dN/dS	2(lnl N-lnlA)	p-value	PS[c]	Liver FPKM	Ovary FPKM	GOs	Reference
alveolin	Zona Pellucida	no significant *Xiphophorus* hit									37
cathepsind	Yolk	ZF, M, SB, TD, XM, XH	957(247)	0.065	1.42	0.23	N	71.147	1.276	C:lysosome; P:response to bacterium; P:proteolysis; F:aspartic-type endopeptidase activity	38
cdx1a	placenta	ZF, M, SB,TD,XH	999(1)	0.107	0	1	N	0	0	P:facial nucleus development; P:facial nerve structural organization; F:transcription regulator activity; F:transcription factor activity; P:positive regulation of transcription from RNA polymerase II promoter; P:rhombomere 3 development; P:anatomical structure formation involved in morphogenesis; C:nucleus	39
cdx4	placenta	ZF, M, SB,XH	732(85)	0.06	0.2	0.65	N	0.02	0.128	P:anterior/posterior axis specification; P:embryonic foregut morphogenesis; P:retinoic acid receptor signaling pathway; F:transcription regulator activity; P:pancreas development; F:sequence-specific DNA binding; P:epithelial cell differentiation; F:transcription factor activity; P:epithelial cell proliferation; P:positive regulation of transcription,	39

Gene	category	Species used[a]	alignment length(starting position[b])	dN/dS	2(lnl N-lnlA)	p-value	PS[c]	Liver FPKM	Ovary FPKM	GOs	Reference
choriogeninH	Zona Pellucida	ZF, M, XH	1212(22)	0.09	1.3	0.24	N	174.578	0.725	-	40
choriogeninH-minor	Zona Pellucida	ZF, M, SB, TD, XM, XH	600(268)	0.19	13.56	2.30E-04	Y	52.835	0.152	-	40
choriogeninL	Zona Pellucida	ZF, M, SB, TD, XM, XH	1263(253)	0.246	0	1	N	176.218	0	-	40
choriolysinH	Zona Pellucida	ZF, M, SB, TD, XM, XH	804(1)	0.17	25.86	3.67E-07	Y	254.26	0.027	F:metalloendopeptidase activity; P:proteolysis; F:zinc ion binding	40
choriolysinL	Zona Pellucida	ZF, M, SB, TD, XM	840(1)	0.05	11.68	6.00E-04	Y	8.561	8.84	F:metalloendopeptidase activity; P:proteolysis; F:zinc ion binding	40
ets2	placenta	ZF, M, SB, TD, XM	555(760)	0.123	0	1	N	0.816	6.672	P:skeletal system development; P:angiogenesis; P:heart morphogenesis; P:response to antibiotic; P:hemangioblast cell differentiation; P:positive regulation of cellular component movement; P:immune response; P:negative regulation of cell cycle; P:negative regulation of cell proliferation; P:cell motility;	41
gcm1	placenta	ZF, M, SB, TD, XM, XH	288(1207)	0.058	0.74	0.39	N	0	0.157	P:regulation of transcription, DNA-dependent; P:cartilage development; P:organ morphogenesis; P:epidermal cell fate specification; F:DNA binding; C:nucleus	42

Gene	category	Species used[a]	alignment length(starting position[b])	dN/dS	2(lnl N-lnlA)	p-value	PS[c]	Liver FPKM	Ovary FPKM	GOs	Reference
hb58	placenta	M, SB,TD, XM,XH	1530(46)	0.142	1.58	0.21	N	6.354	35.373	F:protein kinase C binding; F:RNA binding; P:protein heterooligomerization; P:N-glycan processing; P:intracellular protein kinase cascade; F:calcium ion binding; P:protein amino acid N-linked glycosylation via asparagine; P:protein folding; P:innate immune response; C:alpha-glucosidase II complex; P:post-translational protein modification; C:endoplasmic reticulum lumen; P:phosphorylation; F:kinase activity; P:pronephros development; C:plasma membrane	43
hgfb	placenta	ZF, M, SB, TD, XM	582(1390)	0.096	0	1	N	0	7.754	P:hepatocyte growth factor receptor signaling pathway; P:neuron migration; P:myoblast proliferation; F:growth factor activity; P:regulation of branching involved in salivary gland morphogenesis by mesenchymal-epithelial signaling; P:organ regeneration; P:cell morphogenesis; C:extracellular space; F:protein heterodimerization activity; P:anti-apoptosis; P:liver development; F:catalytic activity; P:cerebellar granule cell differentiation; P:activation of MAPK activity	41

Gene	category	Species used[a]	alignment length(starting position[b])	dN/dS	2(lnl N-lnlA)	p-value	PS[c]	Liver FPKM	Ovary FPKM	GOs	Reference
hxt	placenta	ZF, M, SB, TD, XM, XH	624(1)	0.048	1.9	0.17	N	5.918	0.179	P:trophectodermal cell differentiation; P:angiogenesis; P:negative regulation of transcription from RNA polymerase II promoter; P:cardiac septum morphogenesis; P:ventricular cardiac muscle tissue morphogenesis; C:cytoplasm; P:heart looping; F:transcription coactivator activity; P:cardiac left ventricle formation; ; P:cardiac right ventricle formation; P:positive regulation of transcription from RNA polymerase II promoter; F:protein homodimerization activity; C:nucleus; F:bHLH transcription factor binding	44
igf2	placenta	ZF, TD, XM, PC	663(1)	0.16	3.89	0.05	Y	19.582	27.984	F:growth factor activity; C:extracellular space; F:hormone activity	
mash2	placenta	ZF, M, SB, TD, XM, XH	423(118)	0.046	0	1	N	0.044	0.107	-	45
ncoa6	placenta	ZF, M, SB, TD, XM, XH	4659(133)	0.148	80.76	0	Y	2.872	1.842	P:transcription initiation from RNA polymerase II promoter; P:brain development; C:transcription factor complex; P:heart development; P:labyrinthine layer blood vessel development; F:receptor binding; F:transcription coactivator activity; P:positive regulation of transcription, DNA-dependent; F:chromatin binding	46

Gene	category	Species used[a]	alignment length(starting position[b])	dN/dS	2(lnl N-lnlA)	p-value	PS[c]	Liver FPKM	Ovary FPKM	GOs	Reference
ppara1	placenta	ZF, M, SB, TD, XM, XH	930(460)	0.033	1.22	0.27	N	3.409	22.469	P:steroid hormone mediated signaling pathway; F:zinc ion binding; P:positive regulation of transcription from RNA polymerase II promoter; P:response to cold; F:lipid binding; F:sequence-specific DNA binding; F:transcription factor activity; F:steroid hormone receptor activity; C:nucleus	47
ppara2	placenta	ZF, M, SB, TD, XM, XH	1470(1)	0.083	0	1	N	45.163	71.416	P:steroid hormone mediated signaling pathway; F:zinc ion binding; P:regulation of transcription, DNA-dependent; P:response to cold; F:sequence-specific DNA binding; F:transcription factor activity; F:steroid hormone receptor activity; C:nucleus	47
pparab	placenta	ZF, M, SB, TD, XM, XH	1575(1)	0.083	0.66	0.42	N	17.268	16.551	P:steroid hormone mediated signaling pathway; F:zinc ion binding; P:regulation of transcription, DNA-dependent; F:sequence-specific DNA binding; F:transcription factor activity; F:steroid hormone receptor activity; C:nucleus	47
pparg	placenta	M, SB, TD, XH	1626(1)	0.122	5.92	0.015	Y	29.214	21.2	P:steroid hormone mediated signaling pathway; F:zinc ion binding; P:regulation of transcription, DNA-dependent; F:sequence-specific DNA binding; F:transcription factor activity; F:steroid hormone receptor activity; C:nucleus	47

Gene	category	Species used[a]	alignment length(starting position[b])	dN/dS	2(lnl N-lnlA)	p-value	PS[c]	Liver FPKM	Ovary FPKM	GOs	Reference
transglutaminase	Yolk	ZF, M, SB, TD, XH	2082(1)	0.138	0	1	N	15.599	28.840	P:blood coagulation; P:peptide cross-linking; F:protein-glutamine gamma-glutamyltransferase activity	48
vtg_receptor	Yolk	ZF, M, SB, TD, XM, XH	2460(52)	0.0453	0.015	0.71	N	0	5.149	P:ventral spinal cord development; C:coated pit; F:apolipoprotein binding; P:lipid transport; F:calcium ion binding; F:lipoprotein receptor activity; P:cholesterol metabolic process; C:integral to membrane; P:endocytosis; C:very-low-density lipoprotein particle	49
vtg1	Yolk	ZF, M, SB, XM	3276(1)	0.323	16.48	4.92E-05	Y	12438.587	0.189	F:lipid transporter activity; F:nutrient reservoir activity; P:lipid transport	50
vtg2	Yolk	ZF, M, TD, XM	5097(16)	0.248	0.98	0.32	N	9943.06	1.578	F:lipid transporter activity; F:nutrient reservoir activity; P:lipid transport	50
vtg3	Yolk	M, SB, TD, XM	1329(2488)	0.325	0.012	0.73	N	14835.1	0	F:lipid transporter activity; P:response to chemical stimulus; P:lipid transport	
zpax	Zona Pellucida	ZF, M, SB, TD, XM	2589(130)	0.188	0	1	N	0.034	124.674	F:molecular_function; P:biological_process; C:cellular_component	51
zpb	Zona Pellucida	ZF, M, SB, TD, XM, XH	873(214)	0.196	1.38	0.24	N	0	143.46	-	51
zpc1	Zona Pellucida	M, SB, TD, XM,XH	810(52)	0.19934	1.86	0.17	N	0	68.403	-	51

Gene	category	Species used[a]	alignment length(starting position[b])	dN/dS	2(lnl N-lnlA)	p-value	PS[c]	Liver FPKM	Ovary FPKM	GOs	Reference
zpc2	Zona Pellucida	ZF, M, SB, TD, XM, XH	810(67)	0.128	0	1	N	0	107.665	-	51
zpc3	Zona Pellucida	ZF, M, SB, TD, XM	1454(212)	0.248	0.60	0.46	N	0.1205	0.436	-	51
zpc4	Zona Pellucida	ZF, M, SB, TD, XM	1673(24)	0.086	0.59	0.49	N	8.596	15.955	-	51
zpc5	Zona Pellucida	M, TD, XM,XH	1110(224)	0.157	0	1	N	0	241.019	F:molecular_function; P:biological_process; C:cellular_component	51
zvep	Zona Pellucida	ZF, M, SB, TD, XM, XH	2433(94)	0.06	15	1.08E-04	Y	2.735	7.17	P:positive regulation of transcription, DNA-dependent; F:sequence-specific DNA binding; F:protein binding; C:nucleus; F:zinc ion binding	52

Gene	category	Species used[a]	alignment length(starting position[b])	dN/dS	2(lnl N-lnlA)	p-value	PS[c]	Liver FPKM	Ovary FPKM	GOs	Reference
actin2	control	M, SB, TD, XM, XH	879(19)	0.012	1.48	0.22	N	9.198	4.914	F:ATP binding; C:cytoplasm; C:cytoskeleton; F:nucleotide binding; P:embryonic heart tube development; P:skeletal muscle fiber development; P:heart contraction; P:actomyosin structure organization; P:ATP catabolic process; C:sarcomere; P:actin filament-based movement; C:striated muscle thin filament; P:apoptotic process; F:myosin binding; P:skeletal muscle thin filament assembly; C:actomyosin, actin part; F:ATPase activity; C:stress fiber; C:I band; P:cardiac muscle tissue morphogenesis; F:protein binding; P:cardiac myofibril assembly; C:actin filament; P:cardiac muscle contraction; C:soluble fraction; P:actin-myosin filament sliding	
CAMK1	control	ZF, M, SB, TD, XM, XH	489(432)	0.021	0	1	N	1.399	16.656	F:kinase activity; F:ATP binding; F:protein kinase activity; P:phosphorylation; F:nucleotide binding; P:protein phosphorylation; F:protein serine/threonine kinase activity; C:calcium- and calmodulin-dependent protein kinase complex; F:transferase activity, transferring phosphorus-containing groups	
Chaperone	control	ZF, M, SB, TD, XM	303(30)	0.096	0	1	N	3.452	2.473	F:molecular_function; C:integral to membrane; C:membrane; P:biological_process	

Gene	category	Species used[a]	alignment length(starting position[b])	dN/dS	2(lnl N-lnlA)	p-value	PS[c]	Liver FPKM	Ovary FPKM	GOs	Reference
cyclophilinB	control	ZF, M, SB, TD, XM, XH	522(136)	0.062	0.917	0.34	N	192.131	95.957	F:peptidyl-prolyl cis-trans isomerase activity; P:protein folding; P:protein peptidyl-prolyl isomerization; F:isomerase activity; C:cellular_component; C:endoplasmic reticulum lumen; C:endoplasmic reticulum; C:melanosome; F:peptide binding	
Esterase	control	ZF, M, XM, XH	807(1)	0.069	0	1	N	0	0	C:mitochondrion	
GSTCD	control	M, SB, TD, XM, XH	786(1105)	0.092	0	1	N	3.534	3.45	P:rRNA processing; C:cytoplasm; F:transferase activity; P:rRNA methylation; F:rRNA methyltransferase activity; P:biological_process; F:protein binding; C:cellular_component; F:methyltransferase activity	
tubulin	control	ZF, M, SB, TD, XM, XH	1065(247)	0.019	0	0.996	N	1.994	2.152	P:microtubule-based movement; P:microtubule-based process; P:protein polymerization; P:GTP catabolic process; C:microtubule; F:nucleotide binding; C:cytoplasm; F:GTP binding; C:protein complex; F:GTPase activity; F:structural molecule activity; C:cytoskeleton; P:spindle assembly; F:structural constituent of cytoskeleton	

Gene	category	Species used[a]	alignment length(starting position[b])	dN/dS	2(lnl N-lnlA)	p-value	PS[c]	Liver FPKM	Ovary FPKM	GOs	Reference
angioprotein	control	ZF, M, SB, XM, XH	666(615)	0.07	2.76	0.097	N	1.312	0	P:transmembrane receptor protein tyrosine kinase signaling pathway; F:receptor binding; C:extracellular space; P:blood vessel development; P:signal transduction; P:Tie receptor signaling pathway; P:positive regulation of peptidyl-tyrosine phosphorylation; P:regulation of satellite cell proliferation; endothelial cell migration; C:membrane raft; P:multicellular organismal development; P:positive regulation of protein ubiquitination;	
collagenIX	control	ZF, M, SB, TD,XH	867(1122)	0.073	2.43	0.119	N	0.095	0.777	C:collagen; F:molecular_function; C:collagen type IX; P:biological_process; C:cellular_component; C:extracellular region; P:axon guidance; P:skeletal system development; F:extracellular matrix structural constituent conferring tensile strength	

Gene	category	Species used[a]	alignment length(starting position[b])	dN/dS	2(lnl N-lnlA)	p-value	PS[c]	Liver FPKM	Ovary FPKM	GOs	Reference
Hsp90aa1	control	ZF, M, SB, TD,XH	1407(772)	0.027	1.4	0.236	N	0.462	1.944	C:melanosome; P:response to stress; F:protein homodimerization activity; F:nucleotide binding; C:cytoplasm; F:nitric-oxide synthase regulator activity; F:ATP binding; P:protein import into mitochondrial outer membrane; P:positive regulation of nitric oxide biosynthetic process; C:intracellular; ; P:protein folding; F:TPR domain binding; F:unfolded protein binding; C:perinuclear region of cytoplasm; P:cell cycle; P:gonad development; P:defense response; P:dauer larval development; F:identical protein binding; P:reproduction; P:nematode larval development; P:embryo development ending in birth or egg hatching; P:determination of adult lifespan; F:protein binding; P:hermaphrodite genitalia development	
UHMK1	control	ZF, M, SB, TD, XM, XH	648(421)	0.07	0	1	N	8.794	4.174	F:nucleic acid binding; F:ATP binding; F:protein kinase activity; F:nucleotide binding; P:protein phosphorylation; F:transferase activity, transferring phosphorus-containing groups; C:cellular_component; F:RNA binding; C:nucleus; C:neuronal ribonucleoprotein granule; F:kinase activity; F:protein serine/threonine kinase activity; P:peptidyl-serine phosphorylation; P:regulation of protein export from nucleus; P:positive regulation of translational initiation; C:dendrite cytoplasm;	

Gene	category	Species used[a]	alignment length(starting position[b])	dN/dS	2(lnl N-lnlA)	p-value	PS[c]	Liver FPKM	Ovary FPKM	GOs	Reference
PLCE1	control	ZF, M, SB, TD,XH	2358(120)	0.047	0.49	0.484	N	0.198	6.115	P:intracellular signal transduction; F:phosphatidylinositol phospholipase C activity; P:lipid metabolic process; F:guanyl-nucleotide exchange factor activity; F:phosphoric diester hydrolase activity; P:signal transduction; F:phospholipase C activity; C:intracellular; F:signal transducer activity; P:small GTPase mediated signal transduction;	
PNN	control	ZF, M, SB, TD, XM, XH	552(1)	0.063	0	1	N	15.698	32.626	C:catalytic step 2 spliceosome; P:cell-cell adhesion; P:mRNA processing; C:desmosome; C:nucleus; F:DNA binding; P:RNA splicing; C:cell junction; C:cytoplasm; P:regulation of transcription, DNA-dependent; C:spliceosomal complex; P:transcription, DNA-dependent; C:nuclear speck	
PK2	control	ZF, M, SB, TD, XM, XH	1164(1)	0.045	0.75	0.386	N	3.631	8.006	P:transforming growth factor beta receptor signaling pathway; P:positive regulation of cell proliferation; P:regulation of transcription, DNA-dependent; C:PML body; P:adult walking behavior; C:nuclear body;;	
NDP	control	ZF, M, SB, TD, XM, XH	747(88)	0.053	0	1	N	5.178	18.428	C:cytoplasm; C:cytoskeleton; C:spindle; C:microtubule organizing center; P:transport; C:microtubule; P:retrograde axon cargo transport; P:regulation of neuron projection development; P:cell migration; C:chromosome; P:cell differentiation; P:nuclear envelope	

Gene	category	Species used[a]	alignment length(starting position[b])	dN/dS	2(lnl N-lnlA)	p-value	PS[c]	Liver FPKM	Ovary FPKM	GOs	Reference
Ankyrin	control	ZF, M, SB, TD, XM, XH	765(1108)	0.263	0	1	N	0.918	1.821	F:zinc ion binding; C:intracellular; F:metal ion binding	
GCS	control	ZF, M, SB, TD, XM, XH	1047(100)	0.023	0.213	0.644	N	5.41	22.255	F:ceramide glucosyltransferase activity; P:glycosphingolipid metabolic process; F:transferase activity; F:transferase activity, transferring glycosyl groups; C:integral to membrane; C:membrane; P:lipid metabolic process; P:lipid biosynthetic process; C:Golgi apparatus; P:sphingolipid metabolic process; P:glycosphingolipid biosynthetic process; C:Golgi membrane; P:glucosylceramide biosynthetic process; C:membrane fraction; P:epidermis development	
Hsp27	control	ZF, M, XM	606(1)	0.101	0	1	N	12.708	8.207	P:response to stress; P:response to heat; C:cytoplasm; P:response to arsenic-containing substance; C:nucleus; C:nucleolus	

[a] Likelihood values were calculated by using PAML v4.4. A likelihood ratio test applied to a chi-square distribution with degree of freedom =1. A p-value <0.05 is considered as positively selected.

[b] ZF: zebrafish; M: medaka; SB: stickleback; TD: *Tetraodon*; XM: *X. maculatus*; XH: *X. hellerii*; PL: *P. lucida*

[c] PS: positive selection; N: no; Y: yes

Supplementary Table 6: Test in mammals for selection of viviparity genes positively selected in livebearing fish[a]

Gene	Alignment Length [bp]	Species used	2(lnlN-lnlA)	p-value	PS[b]
igf2	555	mouse, human, cow, wallaby [c]	1,4	0,24	N
pparag	1425	cow, human, mouse, opossum, wallaby [c]	0	1	N
ncoa6	6255	cow, human, mouse, opossum [c]	0,15	0,7	N
vtg1	n.a.	missing in all species	-	-	-
choriogeninHminor	1749	human, mouse, opossum, wallaby	31	2,58E-08	Y
choriolysinH	894	cow, human, mouse, wallaby	1,8	0,18	N
choriolysinL	NA	missing in marsupials	-	-	-
zvep	2085	cow, human, mouse, opossum [c]	0	1	N

[a] Likelihood values were calculated by using PAML v4.4. A likelihood ratio test applied to a chi-square distribution with degree of freedom =1. A p-value <0.05 is considered as positively selected.

[b] PS: positive selection; N: no; Y: yes

[c] inclusion of platypus led to shorter aligment lengths (partly due to the more fragmented genome assembly), but did not change the result.

Supplementary Table 7: List of cognition, pigmentation and liver genes used for post-TGD retention rate analyses

(TGD duplicated genes are in bold)

Cognition				Pigmentation				Liver			
gene number	dupl. gene number	gene	human Ensembl ID	gene number	dupl. gene number	gene	human Ensembl ID	gene number	dupl. gene number	gene	human Ensembl ID
1	**1**	**ABCD3**	**ENSG00000117528**	1		abhd11	ENSG00000106077	1		Aacs	ENSG00000081760
2	**2**	**ALDOA**	**ENSG00000149925**	**2**	**1**	**adam17**	**ENSG00000151694**	2		Abce1	ENSG00000164163
3	**3**	**ANK3**	**ENSG00000151150**	3		adamts20	ENSG00000173157	3		Abhd11	ENSG00000106077
4	**4**	**APOE**	**ENSG00000130203**	4		ap3b1	ENSG00000132842	4		Acadm	ENSG00000117054
5		ARFGAP2	ENSG00000149182	5		ap3d1	ENSG00000065000	5		Acvr1	ENSG00000115170
6		ARFGEF2	ENSG00000124198	6		apc	ENSG00000134982	6		Ada	ENSG00000196839
7		ASTN2	ENSG00000148219	7		asip	ENSG00000101440	7		Apc	ENSG00000134982
8		ATIC	ENSG00000138363	**8**	**2**	**asip2**	**not detected in human**	**8**	**1**	**Arf6**	**ENSG00000165527**
9		ATP1A2	ENSG00000018625	9		atoh7	ENSG00000179774	9		Asl	ENSG00000126522
10	**5**	**ATP1A3**	**ENSG00000105409**	10		atox1	ENSG00000177556	10		Ass1	ENSG00000130707
11	**6**	**ATP2A2**	**ENSG00000174437**	**11**	**3**	**atp6ap1**	**ENSG00000071553**	11		Atg7	ENSG00000197548
12		ATXN10	ENSG00000130638	12		atp6ap2	ENSG00000182220	12		Atp5o	ENSG00000241837
13	**7**	**BAI1**	**ENSG00000181790**	**13**	**4**	**atp6v0c**	**ENSG00000185883**	**13**	**2**	**Atp6ap1**	**ENSG00000071553**
14	**8**	**BAIAP2**	**ENSG00000175866**	14		atp6v0d1	ENSG00000159720	14		Atp6ap2	ENSG00000182220
15	**9**	**CACNA2D1**	**ENSG00000153956**	**15**	**5**	**atp6v1e1**	**ENSG00000131100**	**15**	**3**	**Atp6v0c**	**ENSG00000185883**

Cognition				Pigmentation				Liver			
gene number	dupl. gene number	gene	human Ensembl ID	gene number	dupl. gene number	gene	human Ensembl ID	gene number	dupl. gene number	gene	human Ensembl ID
16		CACNA2D2	ENSG00000007402	16		atp6v1f	ENSG00000128524	16		Atp6v1f	ENSG00000128524
17		CACNA1C	ENSG00000151067	17		atp6v1h	ENSG00000047249	17		Atp6v1h	ENSG00000047249
18		CAMK2D1	ENSG00000145349	18		atp7a	ENSG00000165240	**18**	**4**	**Bmp2**	**ENSG00000125845**
19	**10**	**CAPN1**	**ENSG00000014216**	19		atp7b	ENSG00000123191	19		Bysl	ENSG00000112578
20		CC2D1A	ENSG00000132024	20		atrn	ENSG00000088812	20		C1orf109	ENSG00000116922
21		CDH13	ENSG00000140945	21		bloc1s3	ENSG00000189114	21		C11orf2	ENSG00000149823
22		CHD1L	ENSG00000131778	22		cno	ENSG00000186222	22		Cacna1c	ENSG00000151067
23	**11**	**CHL1**	**ENSG00000134121**	**23**	**6**	**creb1**	**ENSG00000118260**	23		Cad	ENSG00000084774
24	**12**	**CHRNA7**	**ENSG00000175344**	**24**	**7**	**csf1r**	**ENSG00000182578**	24		Cebpg	ENSG00000153879
25	**13**	**CIT**	**ENSG00000122966**	25		dac/fbxw4	ENSG00000107829	**25**	**5**	**Cadm1**	**ENSG00000182985**
26	**14**	**CNTN1**	**ENSG00000018236**	26		dct	ENSG00000080166	26		Ccnd1	ENSG00000110092
27		CNTN2	ENSG00000184144	27		drd2	ENSG00000149295	27		Ccne	ENSG00000105173
28	**15**	**CNTN4**	**ENSG00000144619**	**28**	**8**	**dtnbp1**	**ENSG00000047579**	28		Ccdc49	ENSG00000108296
29		CNTNAP1	ENSG00000108797	29		ebna1bp2	ENSG00000117395	29		Ccm2	ENSG00000136280
30	**16**	**CNTNAP2**	**ENSG00000174469**	30		ece1	ENSG00000117298	30		Ccna2	ENSG00000145386
31		COMT	ENSG00000093010	31		eda	ENSG00000158813	31		Cct5	ENSG00000150753
32		COX6B1	ENSG00000126267	**32**	**9**	**edn3**	**ENSG00000124205**	**32**	**6**	**Cdc37**	**ENSG00000105401**
33		CSMD2	ENSG00000121904	**33**	**10**	**ednrb1**	**ENSG00000136160**	33		Cdipt	ENSG00000103502

Cognition				Pigmentation				Liver			
gene number	dupl. gene number	gene	human Ensembl ID	gene number	dupl. gene number	gene	human Ensembl ID	gene number	dupl. gene number	gene	human Ensembl ID
34		CST3	ENSG00000101439	34		ednrb2	not detected in human	**34**	**7**	**Cdx1**	**ENSG00000113722**
35		CTNNA1	ENSG00000044115	**35**	**11**	**egfr**	**ENSG00000146648**	35		Cdx4	ENSG00000131264
36		CTNNA2	ENSG00000066032	**36**	**12**	**en1**	**ENSG00000163064**	36		Cebpa	ENSG00000245848
37	**17**	**CTNNB1**	**ENSG00000168036**	**37**	**13**	**erbb3**	**ENSG00000065361**	**37**	**8**	**Chmp6**	**ENSG00000176108**
38	**18**	**CTNND1**	**ENSG00000198561**	38		fgfr2	ENSG00000066468	38		Cited2	ENSG00000164442
39	**19**	**CTNND2**	**ENSG00000169862**	39		fig4	ENSG00000112367	**39**	**9**	**Clint1**	**ENSG00000113282**
40		DGKB	ENSG00000136267	40		foxd3	ENSG00000187140	**40**	**10**	**Cltc**	**ENSG00000141367**
41		DIRAS2	ENSG00000165023	**41**	**14**	**frem2**	**ENSG00000150893**	41		Cnot1	ENSG00000125107
42		DISC1	ENSG00000162946	42		fzd4	ENSG00000174804	42		Cpsf1	ENSG00000071894
43		DLAT	ENSG00000150768	43		gart	ENSG00000159131	**43**	**11**	**Cpsf3**	**ENSG00000119203**
44		DLD	ENSG00000091140	44		gch1	ENSG00000131979	44		Cstf3	ENSG00000176102
45	**20**	**DLGAP2**	**ENSG00000198010**	45		gch2	not detected in human	**45**	**12**	**Ctnnb1**	**ENSG00000168036**
46	**21**	**DNM1**	**ENSG00000106976**	46		gchfr	ENSG00000137880	46		Ctdp1	ENSG00000060069
47	**22**	**DNM2**	**ENSG00000079805**	47		gfpt1	ENSG00000198380	47		Ddx18	ENSG00000088205
48	**23**	**DNM3**	**ENSG00000197959**	**48**	**15**	**ghr**	**ENSG00000112964**	48		Ddx27	ENSG00000124228
49		DOCK2	ENSG00000134516	**49**	**16**	**gja5**	**ENSG00000143140**	49		Def	ENSG00000117597
50		DPP10	ENSG00000175497	**50**	**17**	**gnaq**	**ENSG00000156052**	50		Tsr2	ENSG00000158526
51	**24**	**DPP6**	**ENSG00000130226**	**51**	**18**	**gna11**	**ENSG00000088256**	51		Eif3d	ENSG00000100353

Cognition				Pigmentation				Liver			
gene number	dupl. gene number	gene	human Ensembl ID	gene number	dupl. gene number	gene	human Ensembl ID	gene number	dupl. gene number	gene	human Ensembl ID
52	**25**	**DRD1**	**ENSG00000184845**	52		gpc3	ENSG00000147257	**52**	**13**	**Ep300**	**ENSG00000100393**
53	**26**	**DUSP3**	**ENSG00000108861**	53		gpnmb	ENSG00000136235	53		Exosc4	ENSG00000178896
54	**27**	**EPHA4**	**ENSG00000116106**	54		gpr143	ENSG00000101850	54		Erbb2	ENSG00000141736
55	**28**	**ERBB4**	**ENSG00000178568**	55		gpr161	ENSG00000143147	55		Eya1	ENSG00000104313
56		FGF2	ENSG00000138685	**56**	**19**	**hdac1**	**ENSG00000116478**	**56**	**14**	**Fgfr1**	**ENSG00000077782**
57		FLNA	ENSG00000196924	57		hps1	ENSG00000107521	57		Fam32a	ENSG00000105058
58		FMR1	ENSG00000102081	58		hps3	ENSG00000163755	58		Fbl	ENSG00000105202
59		FOXP2	ENSG00000128573	59		hps4	ENSG00000100099	59		Fen1	ENSG00000168496
60	**29**	**GABBR1**	**ENSG00000204681**	60		hps5	ENSG00000110756	**60**	**15**	**Fgf10**	**ENSG00000070193**
61	**30**	**GABBR2**	**ENSG00000136928**	61		hps6	ENSG00000166189	61		Foigr	ENSG00000168538
62		GABRA1	ENSG00000022355	62		ikbkg	ENSG00000073009	62		Foxm1	ENSG00000111206
63		GABRA4	ENSG00000109158	**63**	**20**	**irf4**	**ENSG00000137265**	63		Gata4	ENSG00000136574
64		GABRG1	ENSG00000163285	**64**	**21**	**itgb1**	**ENSG00000150093**	64		Gata6	ENSG00000141448
65		GABRG2	ENSG00000113327	65		kcnj13	ENSG00000115474	65		Gins3	ENSG00000181938
66	**31**	**GABRG3**	**ENSG00000182256**	**66**	**22**	**kit**	**ENSG00000157404**	66		Gle1	ENSG00000119392
67		GAP43	ENSG00000172020	**67**	**23**	**kitlg**	**ENSG00000049130**	67		Gnl3	ENSG00000163938
68		GDI1	ENSG00000203879	68		lef1	ENSG00000138795	68		Gnl3l	ENSG00000130119
69	**32**	**GJA1**	**ENSG00000152661**	69		lmx1a	ENSG00000162761	69		Gnpnat1	ENSG00000100522

Cognition				Pigmentation				Liver			
gene number	dupl. gene number	gene	human Ensembl ID	gene number	dupl. gene number	gene	human Ensembl ID	gene number	dupl. gene number	gene	human Ensembl ID
70	**33**	**GJA8**	**ENSG00000121634**	70		ltk	ENSG00000062524	70		Gtpbp4	ENSG00000107937
71	**34**	**GLUD1**	**ENSG00000148672**	71		lyst	ENSG00000143669	71		Hdac3	ENSG00000171720
72	**35**	**GNA13**	**ENSG00000120063**	72		mbtps1	ENSG00000140943	72		Heatr1	ENSG00000119285
73	**36**	**GNAI1**	**ENSG00000127955**	**73**	**24**	**mchr1**	**ENSG00000128285**	**73**	**16**	**Hes1**	**ENSG00000114315**
74	**37**	**GNAI2**	**ENSG00000114353**	74		mchr2	ENSG00000152034	**74**	**17**	**Hgf**	**ENSG00000019991**
75	**38**	**GNAQ**	**ENSG00000156052**	75		mc1r	ENSG00000198211	75		Hhex	ENSG00000152804
76	**39**	**GNAS**	**ENSG00000087460**	**76**	**25**	**mcoln3**	**ENSG00000055732**	**76**	**18**	**Hinfp**	**ENSG00000172273**
77		GNPAT	ENSG00000116906	**77**	**26**	**mgrn1**	**ENSG00000102858**	77		Hnf1a	ENSG00000135100
78	**40**	**GPHN**	**ENSG00000171723**	**78**	**27**	**mitf**	**ENSG00000187098**	**78**	**19**	**Hnf1b**	**ENSG00000108753**
79	**41**	**GPI**	**ENSG00000105220**	**79**	**28**	**mlph**	**ENSG00000115648**	**79**	**20**	**Hspa8**	**ENSG00000109971**
80	**42**	**GRIA1**	**ENSG00000155511**	80		mreg	ENSG00000118242	80		Hspa9b	ENSG00000113013
81	**43**	**GRIA2**	**ENSG00000120251**	**81**	**29**	**myc**	**ENSG00000136997**	81		Icmt	ENSG00000116237
82	**44**	**GRIA3**	**ENSG00000125675**	82		mycbp2	ENSG00000005810	82		Kars	ENSG00000065427
83	**45**	**GRIA4**	**ENSG00000152578**	**83**	**30**	**myo5a**	**ENSG00000197535**	83		Kri1	ENSG00000129347
84	**46**	**GRIN1**	**ENSG00000176884**	**84**	**31**	**myo7a**	**ENSG00000137474**	**84**	**21**	**Jag1**	**ENSG00000101384**
85	**47**	**GRIN2A**	**ENSG00000183454**	**85**	**32**	**nsf**	**ENSG00000073969**	**85**	**22**	**Jag2**	**ENSG00000184916**
86	**48**	**GRIN2B**	**ENSG00000150086**	86		oca2	ENSG00000104044	**86**	**23**	**Jarid2**	**ENSG00000008083**
87	**49**	**GRIN2D**	**ENSG00000105464**	**87**	**33**	**pabpc1**	**ENSG00000070756**	87		Klf1	ENSG00000105610

Cognition				Pigmentation				Liver			
gene number	dupl. gene number	gene	human Ensembl ID	gene number	dupl. gene number	gene	human Ensembl ID	gene number	dupl. gene number	gene	human Ensembl ID
88	**50**	**GRM2**	**ENSG00000164082**	88		paics	ENSG00000128050	88		Lsr	ENSG00000105699
89	**51**	**GRM5**	**ENSG00000168959**	**89**	**34**	**pax3**	**ENSG00000135903**	89		Ltv1	ENSG00000135521
90	**52**	**GRM7**	**ENSG00000196277**	**90**	**35**	**pax7**	**ENSG00000009709**	90		Man2a1	ENSG00000112893
91	**53**	**GSK3B**	**ENSG00000082701**	91		pcbd1	ENSG00000166228	91		Mars	ENSG00000166986
92	**54**	**HADHA**	**ENSG00000084754**	92		pcbd2	ENSG00000132570	92		Mcm3ap	ENSG00000160294
93		HFE2	ENSG00000168509	93		pldn	ENSG00000104164	93		Mcm7	ENSG00000166508
94	**55**	**HOMER1**	**ENSG00000152413**	94		pmch	ENSG00000183395	94		Med1	ENSG00000125686
95		HOMER2	ENSG00000103942	**95**	**36**	**pomc**	**ENSG00000115138**	95		Med12	ENSG00000184634
96		HSD17B4	ENSG00000133835	96		pts	ENSG00000150787	96		Med14	ENSG00000180182
97		HSPD1	ENSG00000144381	**97**	**37**	**qdpr**	**ENSG00000151552**	**97**	**24**	**Mkl2**	**ENSG00000186260**
98	**56**	**HTR1A**	**ENSG00000178394**	98		rab27a	ENSG00000069974	98		Mybbp1a	ENSG00000132382
99		ITGA10	ENSG00000143127	**99**	**38**	**rab38**	**ENSG00000123892**	**99**	**25**	**Mybl2**	**ENSG00000101057**
100	**57**	**KALRN**	**ENSG00000160145**	**100**	**39**	**rab32**	**ENSG00000118508**	100		Nar	ENSG00000160917
101	**58**	**KCNQ2**	**ENSG00000075043**	101		rabggta	ENSG00000100949	101		Ncl	ENSG00000115053
102	**59**	**KIF1A**	**ENSG00000130294**	102		rpl24	ENSG00000114391	102		Ndufs5	ENSG00000168653
103	**60**	**KIF5B**	**ENSG00000170759**	103		rps19	ENSG00000105372	**103**	**26**	**Nf1**	**ENSG00000196712**
104		KRAS	ENSG00000133703	104		rps20	ENSG00000008988	104		Nf2	ENSG00000186575
105		L1CAM	ENSG00000198910	105		scarb2	ENSG00000138760	105		Nfyc	ENSG00000066136

Cognition				Pigmentation				Liver			
gene number	dupl. gene number	gene	human Ensembl ID	gene number	dupl. gene number	gene	human Ensembl ID	gene number	dupl. gene number	gene	human Ensembl ID
106	**61**	**LPHN1**	**ENSG00000072071**	106		sfxn1	ENSG00000164466	106		Nkap	ENSG00000189134
107	**62**	**LPHN3**	**ENSG00000150471**	**107**	**40**	**silver**	**ENSG00000185664**	107		Nmd3	ENSG00000169251
108		LRPPRC	ENSG00000138095	108		skiv2l2	ENSG00000039123	108		Noc3l	ENSG00000173145
109		LSAMP	ENSG00000185565	**109**	**41**	**slc24a4**	**ENSG00000140090**	109		Nol10	ENSG00000115761
110		MACROD1	ENSG00000133315	110		slc24a5	ENSG00000188467	110		Nop10	ENSG00000182117
111		MACROD2	ENSG00000172264	111		slc45a2	ENSG00000164175	111		Notch2	ENSG00000134250
112	**63**	**MAGI2**	**ENSG00000187391**	**112**	**42**	**smtl**	**not detected in human**	112		Notch3	ENSG00000074181
113		MAOA	ENSG00000189221	113		snai2	ENSG00000019549	113		Nup205	ENSG00000155561
114	**64**	**MAP1A**	**ENSG00000166963**	**114**	**43**	**sox9**	**ENSG00000125398**	114		Onecut1	ENSG00000169856
115		MAP1B	ENSG00000131711	**115**	**44**	**sox10**	**ENSG00000100146**	115		Onecut2	ENSG00000119547
116	**65**	**MAPT**	**ENSG00000186868**	116		sox18	ENSG00000203883	116		Paf1	ENSG00000006712
117		MECP2	ENSG00000169057	**117**	**45**	**spr**	**ENSG00000116096**	**117**	**27**	**Pbx4**	**ENSG00000105717**
118	**66**	**MYO5A**	**ENSG00000197535**	118		tfap2a	ENSG00000137203	118		Pcsk2	ENSG00000125851
119	**67**	**NCAM1**	**ENSG00000149294**	119		tpcn2	ENSG00000162341	119		Pes1	ENSG00000100029
120		NCAM2	ENSG00000154654	120		trim33	ENSG00000197323	**120**	**28**	**Pkm2**	**ENSG00000067225**
121		NDUFA2	ENSG00000131495	**121**	**46**	**trpm1**	**ENSG00000134160**	121		Polr1c	ENSG00000171453
122	**68**	**NDUFS1**	**ENSG00000023228**	122		trpm7	ENSG00000092439	122		Polr1d	ENSG00000186184
123		NDUFS3	ENSG00000213619	123		muted	ENSG00000188428	123		Polr3f	ENSG00000132664

Cognition				Pigmentation				Liver			
gene number	dupl. gene number	gene	human Ensembl ID	gene number	dupl. gene number	gene	human Ensembl ID	gene number	dupl. gene number	gene	human Ensembl ID
124		NDUFS7	ENSG00000115286	**124**	**47**	**tyr**	**ENSG00000077498**	**124**	**29**	**Ppm1k**	**ENSG00000163644**
125	**69**	**NEFL**	**pseudogene**	**125**	**48**	**tyrp1**	**ENSG00000107165**	125		Ppp1r12a	ENSG00000058272
126		NIPA2	ENSG00000140157	126		vps11	ENSG00000160695	126		Prpf3	ENSG00000117360
127	**70**	**NLGN2**	**ENSG00000169992**	127		vps18	ENSG00000104142	127		Psmb1	ENSG00000008018
128	**71**	**NLGN3**	**ENSG00000196338**	128		vps33a	ENSG00000139719	128		Psmc6	ENSG00000100519
129		NRG1	ENSG00000157168	129		vps39	ENSG00000166887	129		Qars	ENSG00000172053
130	**72**	**NRXN1**	**ENSG00000179915**	130		wnt1	ENSG00000125084	130		Rae1	ENSG00000101146
131	**73**	**NRXN3**	**ENSG00000021645**	131		wnt3a	ENSG00000154342	131		Rb1cc1	ENSG00000023287
132	**74**	**PCDH1**	**ENSG00000156453**	132		xdh	ENSG00000158125	132		Rbm19	ENSG00000122965
133	**75**	**PDE4B**	**ENSG00000184588**	**133**	**49**	**zic2**	**ENSG00000043355**	133		Rbm42	ENSG00000126254
134		PDE8A	ENSG00000073417					134		Rbp4	ENSG00000138207
135	**76**	**PDHA1**	**ENSG00000131828**					135		Rcl	ENSG00000112667
136		PEX11B	ENSG00000131779					136		Rela	ENSG00000173039
137		PGK1	ENSG00000102144					137		Rnf113a	ENSG00000125352
138		PHGDH	ENSG00000092621					138		Rpgrip1l	ENSG00000103494
139	**77**	**PI4KA**	**ENSG00000241973**					139		Rpl9	ENSG00000163682
140	**78**	**PLD5**	**ENSG00000180287**					140		Rpl11	ENSG00000142676
141	**79**	**PLP1**	**ENSG00000123560**					141		Rpl12	ENSG00000197958

Cognition				Pigmentation				Liver			
gene number	dupl. gene number	gene	human Ensembl ID	gene number	dupl. gene number	gene	human Ensembl ID	gene number	dupl. gene number	gene	human Ensembl ID
142		POLR3C	ENSG00000186141					142		Rplp1	ENSG00000137818
143	**80**	**POLR3GL**	**ENSG00000121851**					143		Rprd1b	ENSG00000101413
144		POU6F2	ENSG00000106536					144		Rps11	ENSG00000142534
145		PRKAB2	ENSG00000131791					145		Rpsa	ENSG00000168028
146		PRKCG	ENSG00000126583					146		Rrp1	ENSG00000160208
147	**81**	**PRODH**	**ENSG00000100033**					147		Runx1	ENSG00000159216
148	**82**	**PTEN**	**ENSG00000171862**					148		Ruvbl2	ENSG00000183207
149		PTPN11	ENSG00000179295					**149**	**30**	**Rxra**	**ENSG00000186350**
150		RAB1A	ENSG00000138069					150		Sars	ENSG00000031698
151		RAB3GAP2	ENSG00000118873					151		Scarb2	ENSG00000138760
152		RAPGEF2	ENSG00000109756					152		Sdad1	ENSG00000198301
153		RELN	ENSG00000189056					153		Sec61a	ENSG00000058262
154	**83**	**ROCK2**	**ENSG00000134318**					154		Sf1	ENSG00000168066
155		**SEC22B**	**ENSG00000223380**					155		Skp1a	ENSG00000113558
156		SEMA5A	ENSG00000112902					**156**	**31**	**Slc25a5**	**ENSG00000005022**
157	**84**	**SHANK2**	**ENSG00000162105**					157		Slco1b1	ENSG00000134538
158	**85**	**SHANK3**	**ENSG00000251322**					158		Smarca5	ENSG00000153147
159	**86**	**SLC1A3**	**ENSG00000079215**					159		Sod2	ENSG00000112096

Cognition				Pigmentation				Liver			
gene number	dupl. gene number	gene	human Ensembl ID	gene number	dupl. gene number	gene	human Ensembl ID	gene number	dupl. gene number	gene	human Ensembl ID
160		SLC25A12	ENSG00000115840					**160**	**32**	**Sox9**	**ENSG00000125398**
161		SLC25A4	ENSG00000151729					161		Sp1	ENSG00000185591
162		SLC6A4	ENSG00000108576					**162**	**33**	**Sp3**	**ENSG00000172845**
163	**87**	**SNAP25**	**ENSG00000132639**					163		Sptan1	ENSG00000197694
164		ST8SIA2	ENSG00000140557					164		Spns1	ENSG00000169682
165		STX1A	ENSG00000106089					**165**	**34**	**Stat5**	**ENSG00000126561**
166	**88**	**STXBP5**	**ENSG00000164506**					166		Surf6	ENSG00000148296
167	**89**	**SV2A**	**ENSG00000159164**					167		Taf1b	ENSG00000115750
168		SYN1	ENSG00000008056					168		Taf2	ENSG00000064313
169	**90**	**SYNE1**	**ENSG00000131018**					169		Taf8	ENSG00000137413
170	**91**	**SYNGAP1**	**ENSG00000197283**					**170**	**35**	**Tbx16**	**not detected in human**
171	**92**	**SYP**	**ENSG00000102003**					171		Tgfbr3	ENSG00000069702
172	**93**	**SYT1**	**ENSG00000067715**					172		Thoc2	ENSG00000125676
173		TAF1C	ENSG00000103168					173		Ttk	ENSG00000112742
174		TPH2	ENSG00000139287					174		Tubg2	ENSG00000037042
175	**94**	**TPI1**	**ENSG00000111669**					175		Ufd1l	ENSG00000070010
176	**95**	**TSC1**	**ENSG00000165699**					176		Uhrf1	ENSG00000034063

Cognition				Pigmentation				Liver			
gene number	**dupl. gene number**	**gene**	**human Ensembl ID**	**gene number**	**dupl. gene number**	**gene**	**human Ensembl ID**	**gene number**	**dupl. gene number**	**gene**	**human Ensembl ID**
177	**96**	**TSPAN7**	**ENSG00000156298**					177		Utp11l	ENSG00000183520
178		UBA1	ENSG00000130985					178		Vps18	ENSG00000104142
179	**97**	**UBE2N**	**ENSG00000177889**					179		Vps39	ENSG00000166887
180		UBE2M	ENSG00000130725					180		Vwf	ENSG00000110799
181		UBE2V2	ENSG00000169139					181		Wnt2b	ENSG00000134245
182		UBE3C	ENSG00000009335					182		Wdr33	ENSG00000136709
183		UBE4A	ENSG00000110344					183		Wdr36	ENSG00000134987
184		UBL4A	ENSG00000102178					184		Wdr46	ENSG00000204221
185		UBR4	ENSG00000127481					**185**	**36**	**Wdr68**	**ENSG00000136485**
186		UBXN6	ENSG00000167671					**186**	**37**	**Zcchc7l**	**ENSG00000147905**
187		UCHL1	ENSG00000154277					187		Zmat2	ENSG00000146007
188		VCP	ENSG00000165280								
189	**98**	**YWHAB**	**ENSG00000166913**								
190	**99**	**YWHAE**	**ENSG00000108953**								

Supplementary Table 8A: Human genes that show copy number variations (CNV)

	CNV genes	total	%
all genes genome	8890	20038	44,4
all genes this study	177	491	36,0
all TGD genes this study	65	177	36,7
singletons this study	112	314	35,7
cognition genes	63	192	32,8
cognition genes TGD	31	99	31,3
cognition singletons	32	93	34,4
pigmentation genes	48	129	37,2
pigmentation genes TGD	20	47	42,6
pigmentation singletons	28	82	34,1
liver genes	71	185	38,4
liver genes TGD	16	36	44,4
liver singletons	55	149	36,9

Supplementary Table 8B: Zebrafish genes that show copy number variations (CNV)

	CNV genes	total	%
all genes genome	2681	25325	10,6
all genes this study	31	658	4,7
all TGD genes this study	18	343	5,2
all genes singleton	13	315	4,1
cognition genes	14	286	4,9

	CNV genes	total	%
cognition genes TGD	11	194	5,7
cognition singletons	3	92	3,3
pigmentation genes	5	173	2,9
pigmentation genes TGD	4	91	4,4
pigmentation singletons	1	82	1,2
liver genes	13	218	6,0
liver genes TGD	4	68	5,9
liver singletons	9	150	6,0

Supplementary Table 9: Human genes being members of protein complexes that have corresponding TGD paralogs or singletons in the three fish functional categories

	Protein complex member	total	%
all genes genome	2665	20038	13,3
all genes this study	191	491	38,9
all TGD genes this study	74	176	42,0
all genes singleton	117	315	37,1
cognition genes	80	192	41,7
cognition genes TGD	45	99	45,5
cognition singletons	35	93	37,6
pigmentation genes	40	129	31,0
pigmentation genes TGD	20	47	42,6
pigmentation singletons	20	82	24,4

	Protein complex member	total	%
liver genes	76	185	41,1
liver genes TGD	11	36	30,6
liver singletons	65	149	43,6

Supplementary Table 10: Lengths of human proteins that have corresponding TGD paralogs or singletons in the three fish functional categories

	TGD paralogs		singletons		
	genes	%	genes	%	
short (<200aa)	9	5,14	42	13,46	
middle (200-1000aa)	120	68,57	208	66,67	
long (>1000aa)	46	26,29	62	19,87	
total	175		312		
	cognition		**pigmentation**		**liver**
	genes	%	genes	%	genes %
short (<200aa)	11	5,79	19	27,41	23 12,57
middle (200-1000aa)	124	65,26	86	67,19	128 69,95
long (>1000aa)	55	28,95	23	17,97	32 17,49
total	190		128		183

Supplementary Table 11. Non-coding RNAs in the genome and in the transcriptome

ncRNA type	Genome	Transcriptome
tRNA	535	n.a.
rRNA	51	21
miRNA	611	n.a.
snRNA	38	17
snoRNA	229	109

Supplementary Table 12. tRNA statistics in the genome assembly

Amino Acids	anticodon loops	Numbers[1]
Ala	TGC	19
	GGC	1
	CGC	3
	AGC	7
Arg	CCT	8
	TCT	10
	CCG	3
	ACG	7
	TCG	8
Asn	GTT	12
Asp	ATC	8
	GTC	137
Cys	GCA	12
Gln	CTG	6
	TTG	2
Glu	TTC	8
	CTC	15
Gly	CCC	2
	TCC	8
	GCC	11
His	GTG	12
Ile	TAT	14
	AAT	30
	GAT	1
Leu	AAG	8
	TAG	3
	TAA	8
	CAG	9
	CAA	4
Lys	CTT	11
	TTT	9
Met	CAT	16
Phe	GAA	5
Pro	AGG	10
	TGG	14
	CGG	5
Ser	AGA	7
	CGA	5
	GCT	8
	TGA	2
	ACT	1
Thr	CGT	3
	TGT	16
	AGT	11
Trp	CCA	11
Tyr	GTA	9
Val	TAC	5
	CAC	5
	AAC	6

[1] number of tRNA genes per codons

Supplementary Table 13. Transcript reads used for Tophat gene models

	Heart	Liver	Brain	Mixed 60mers[1]	Mixed 76mers[2]
Format	SIPES	SIPES	SIPES	SIPES	Single end
Number of reads	71,954,940	68,992,472	73,224,886	154,390,166	44,178,317
Read length	60	60	60	60	76
Number nucleotides	4,317,296,400	4,139,548,320	4,393,493,160	9,263,409,9600	3,357,552,000
Number potential transcripts	156,433	119,681	170,818	241,093	292,258
No coverage cutoff					
Number potential transcripts	64,021	33,374	72,008	90,314	43,524
-3x coverage cutoff					

[1] From adult ovary and testes and from embryos at stage 15 and stage 25 of development
[2] From five day and one month old mixed sexes, nine month old females, and 15 month old males

Supplementary Table 14. Tree taxon IDs and cluster codes

Taxon ID	Organism
1	*Strongylocentrotus purpuratus*
2	*Ciona intestinalis*
3	*Xiphophorus maculatus*
4	*Oryzias latipes*
5	*Gastersteus aculatus*
6	*Tetraodon nigroviridis*
7	*Danio rerio*
8	*Xenopus tropicalis*
9	*Gallus gallus*
10	*Monodelphis domestica*
11	*Mus musculus*
12	*Homo sapiens*

Cluster level	Outgroup(s)	Ingroup(s)
1	1	2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12
2	2	3, 4, 5, 6, 7, 8, 9, 10, 11, 12
3	8	9, 10, 11, 12
4	9	10, 11, 12
5	10	11, 12
6	11	12
7	8, 9, 10, 11, 12	3, 4, 5, 6, 7
8	7	3, 4, 5, 6
9	5,6	3, 4
10	5	6
11	4	3
12	1	1
13	2	2
14	8	8
15	9	9
16	10	10
17	11	11
18	12	12
19	7	7
20	5	5
21	6	6
22	4	4
23	3	3

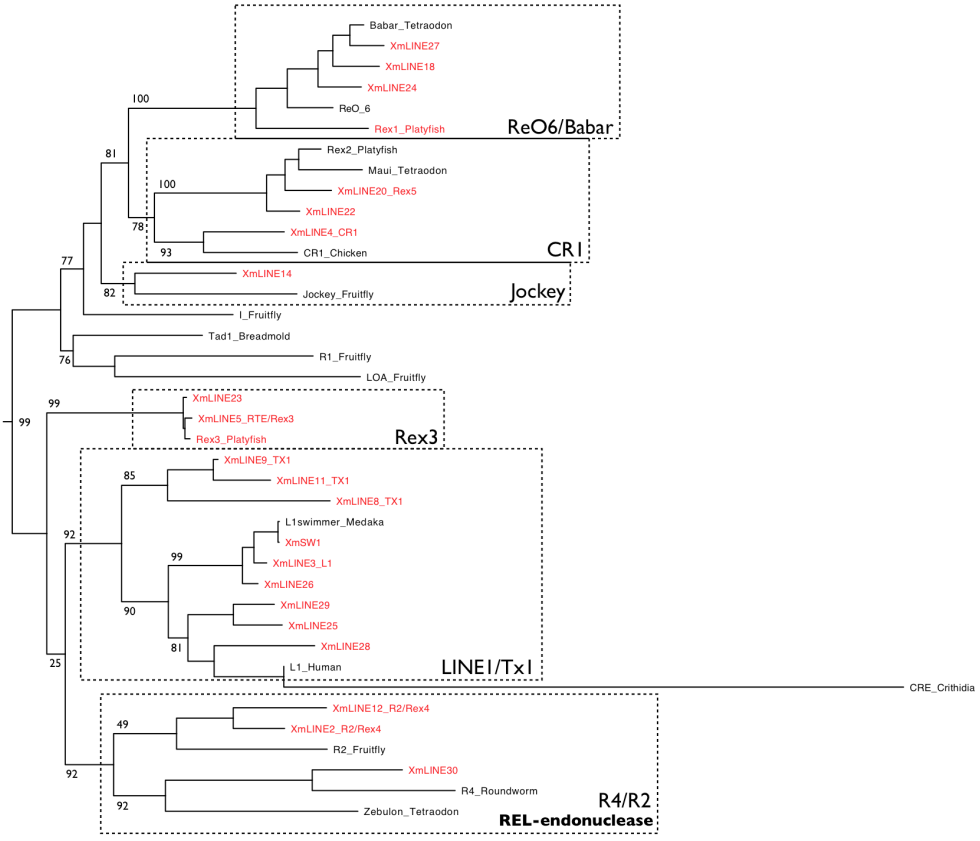
Supplementary Table 15. Quantity of clusters of different sizes when clustered by a distance of 1kb

Cluster size	Members
1	3326
2	127
3	24
4	4
6	1

SUPPLEMENTARY FIGURES

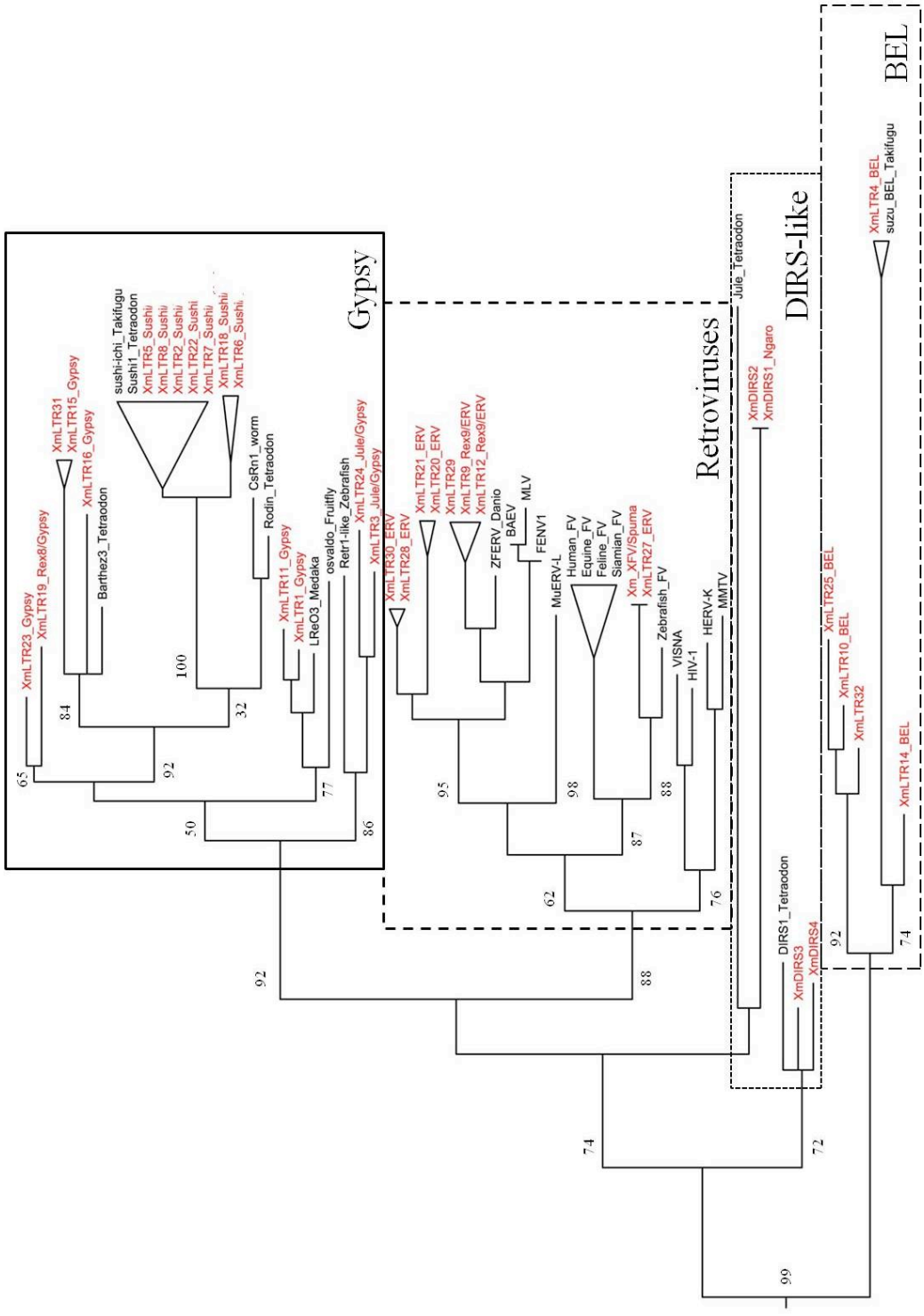
Supplementary Figure 1: Phylogenetic tree of *X. maculatus* Long Interspersed Nuclear Elements (LINE) based on reverse transcriptase alignment

Protein sequences were aligned with ClustalW and a phylogenetic tree was constructed on 136 amino acids with the PhyML package using maximum likelihood methods [53] with default bootstrap calculation (shown at the beginning of branches). Platyfish elements are written in red and names of the elements start with "Xm" prefix. Boxes highlight the different apurinic-apyrimidic endonuclease (Jockey, Maui, CR1, ReO6/Babar, L1, Rex3/RTE and Tx-1) and Restriction Enzyme-Like (REL)-endonuclease (R2 and R4).



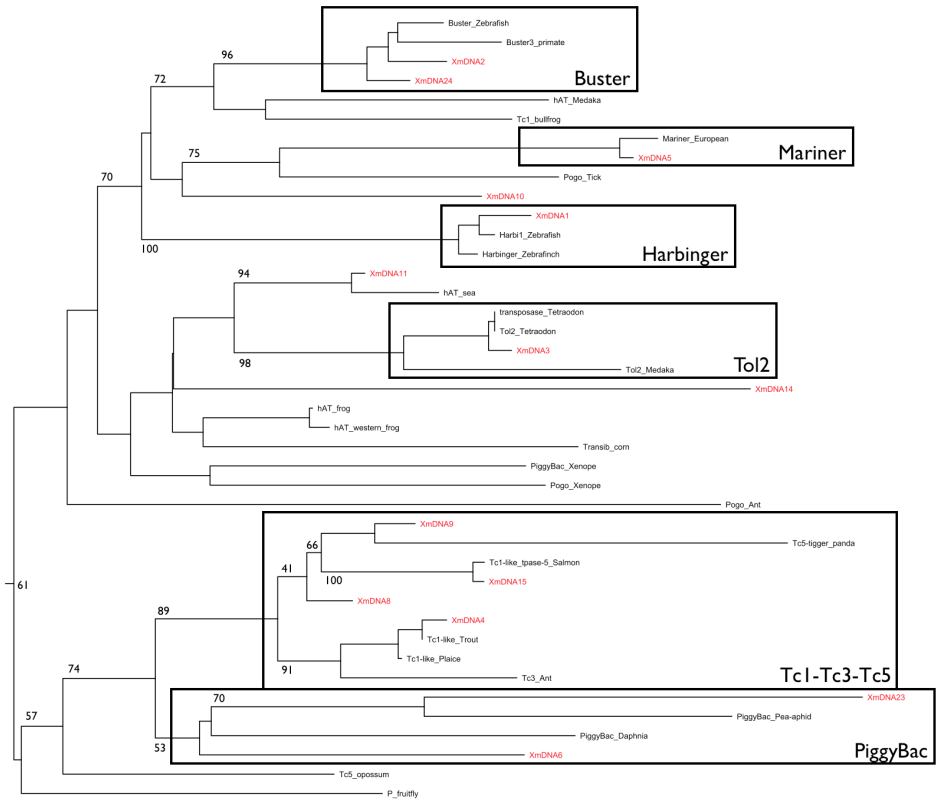
Supplementary Figure 2: Phylogenetic tree of *X. maculatus* Long Terminal Repeat (LTR) retroelements based on reverse transcriptase alignment

Protein sequences were aligned with ClustalW (91 amino acids) and the phylogenetic tree was constructed with the PhyML package using maximum likelihood methods [53] with default bootstrap calculation (shown at the beginning of branches). Platyfish elements are written in red and the names of the elements start with the "Xm".



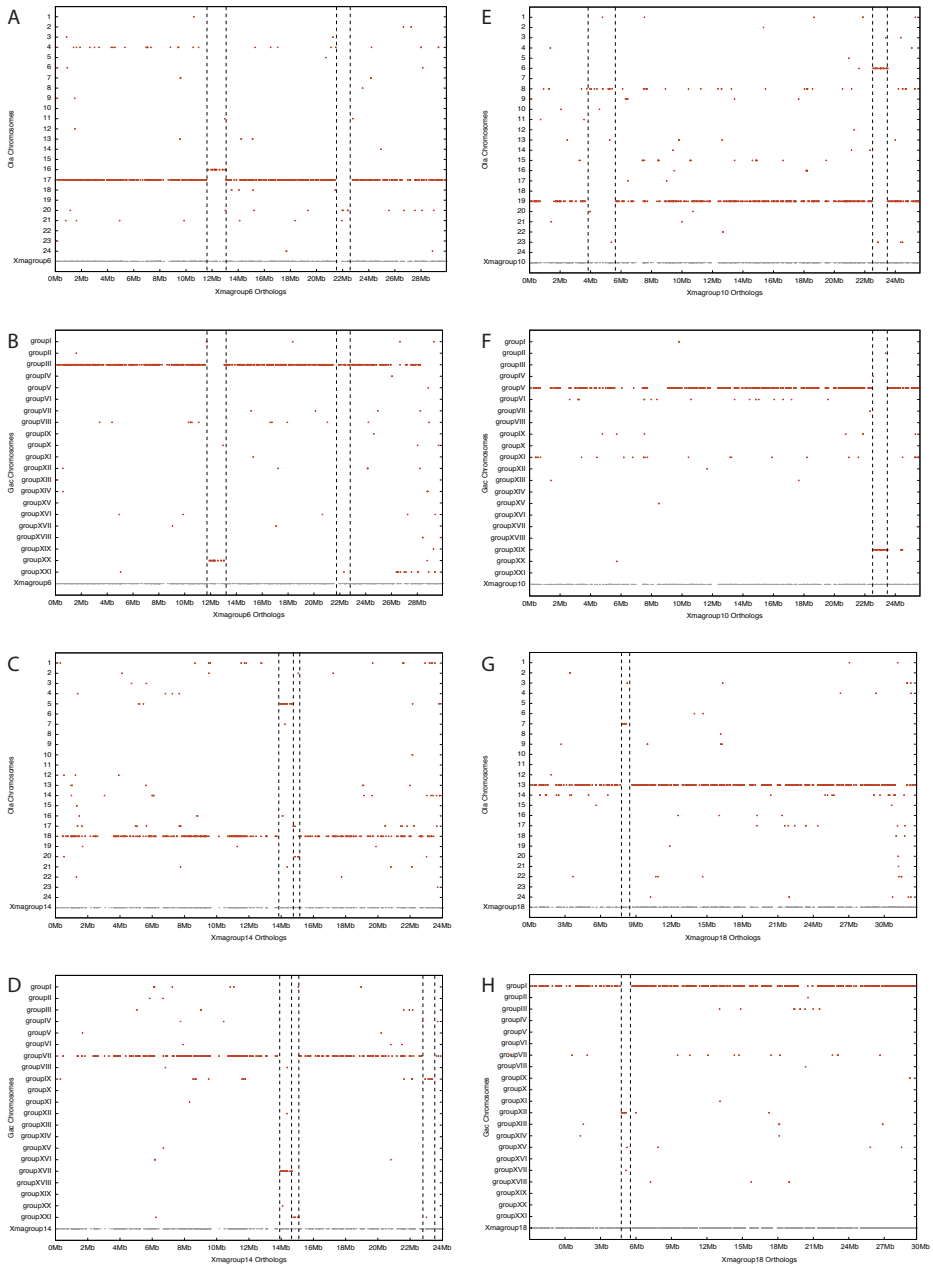
Supplementary Figure 3: Phylogenetic tree of *X. maculatus* DNA transposons based on transposase alignment

Protein sequences were aligned with ClustalW (307 amino acids) and a phylogenetic tree was constructed with the PhyML package using maximum likelihood methods [53] with default bootstrap calculation (shown at the beginning of branches). Platyfish elements are written in red and names of the elements start with the "Xm" prefix.



Supplementary Figure 4: Split conservation of chromosomal location between platyfish and medaka or stickleback chromosomes

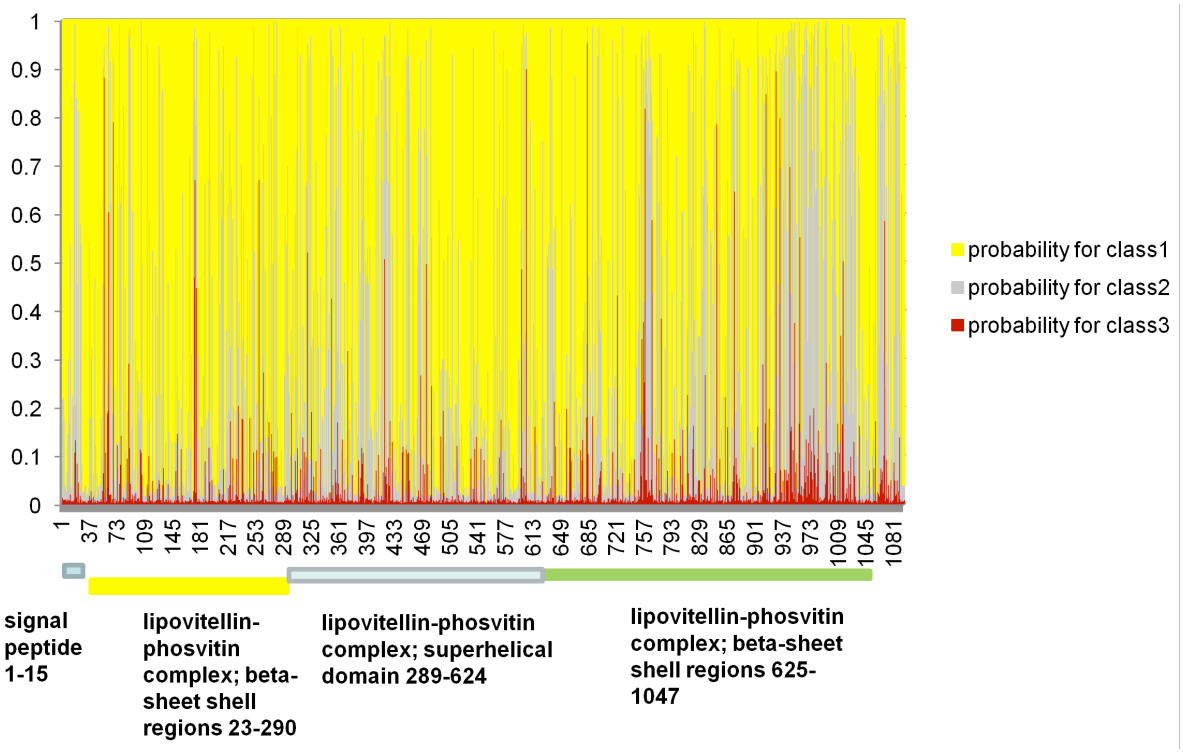
Dotplots reveal small discrepancies between platyfish (Xma) and medaka (Ola) or stickleback (Gac) chromosomes. Orthologs of genes on platyfish chromosomes Xma6, (**A, B**) Xma10 (**E, F**), Xma14 (**C, D**), Xma18 (**G, H**) are mostly on single other medaka or stickleback chromosomes but have small portions that reside on a second medaka or stickleback chromosome.



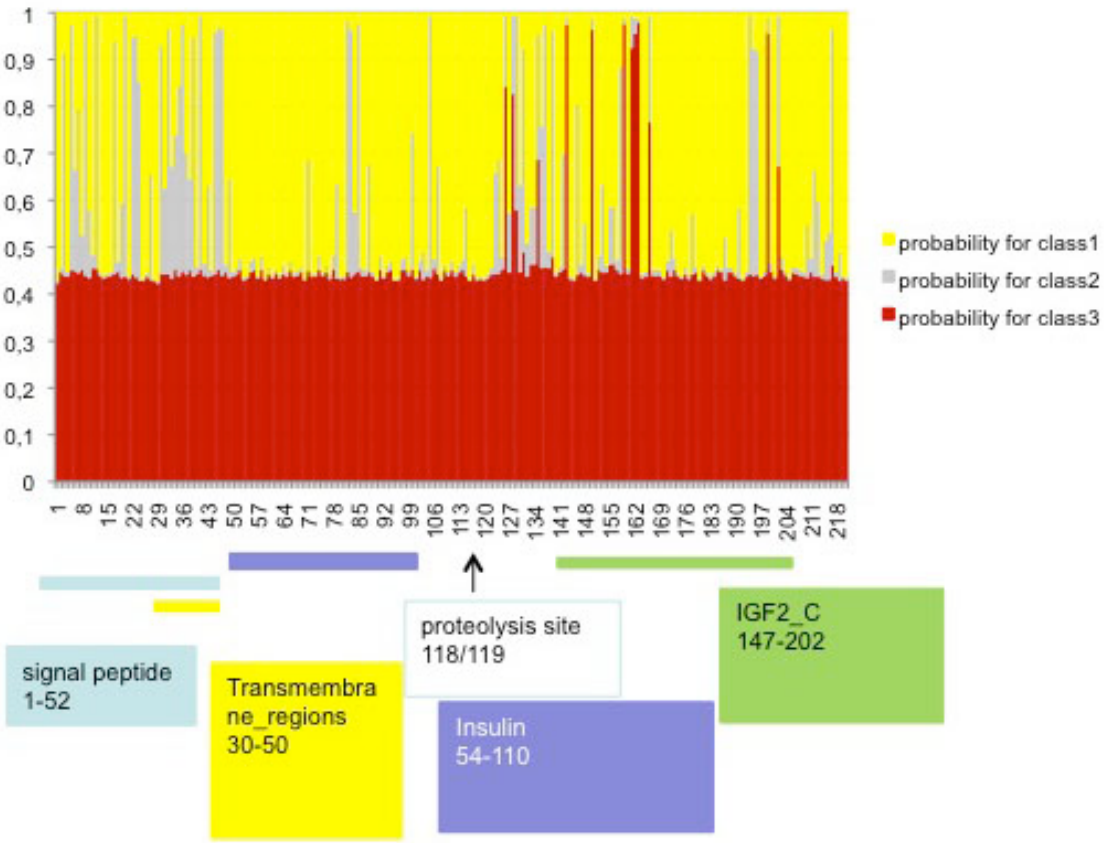
Supplementary Figure 5: Posterior probabilities for viviparity genes under positive selection using branch site model

For each amino acid site, there are four probabilities calculated by Bayes Empirical Bayes analysis. Class 1 (yellow) is the probability of this site being under purifying selection (ka/ks ratio about 0), class 2 (grey) is the probability of this site being under natural selection (ka/ks ratio about 1), class 3 (red) is the probability of this site being under positive selection in the analyzed Xiphophorus species. (A) *vitellogenin1*, (B) *igf2*, (C) *pparg*, (D) *ncoa6*, (E) *choriolysinL*, (F) *choriolysinH*, (G) *zvep, (H) caudal type homeobox 4* (*cdx4*). Known functional protein domains are indicated as bars.

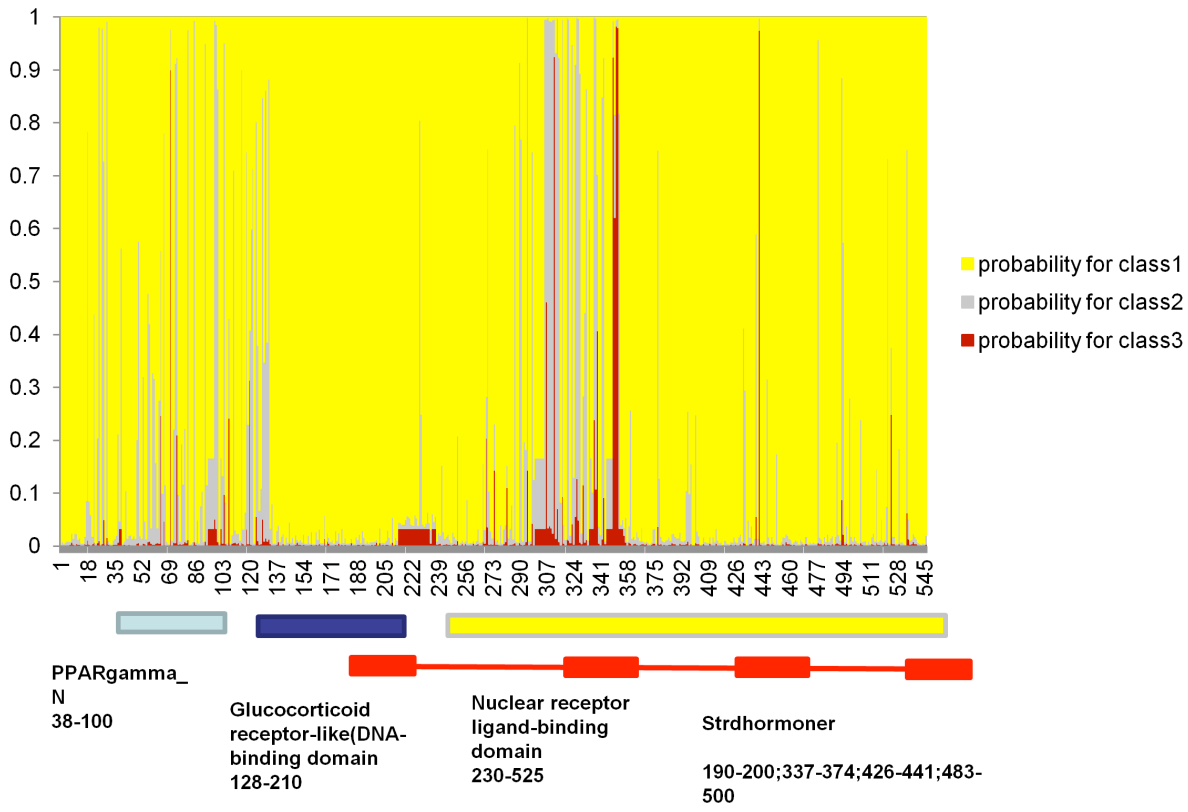
(A)



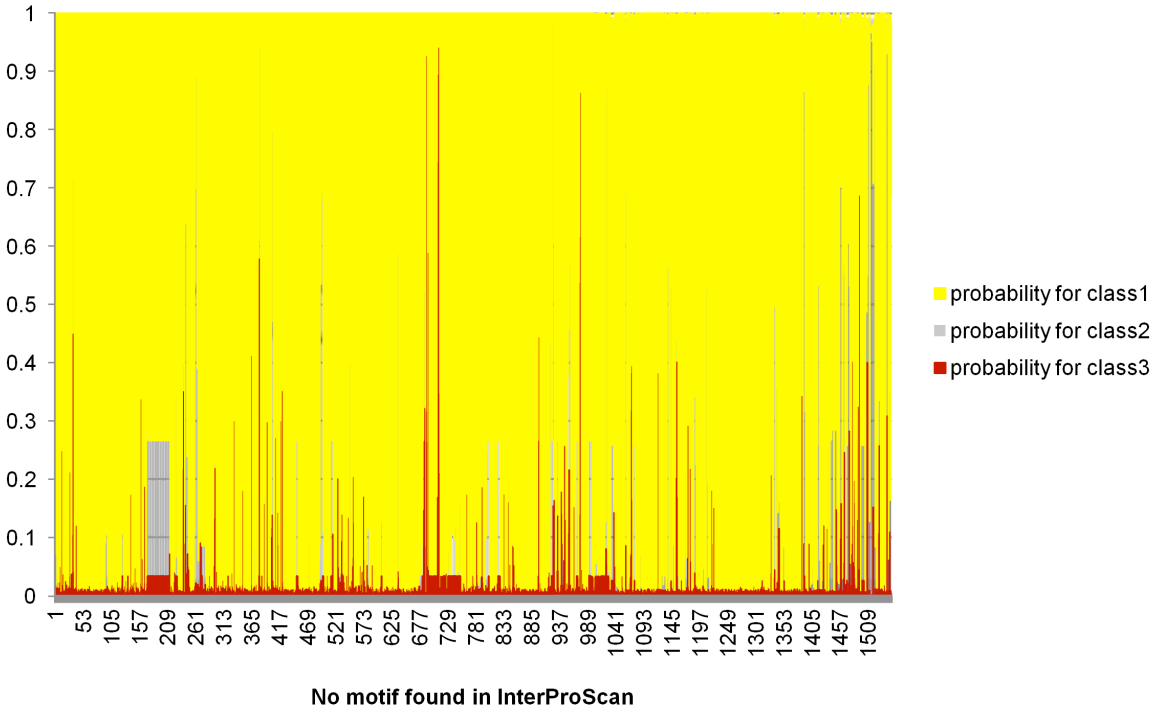
(B)


probability for class1
probability for class2
probability for class3
signal peptide
1-52
Transmembrane_regions
30-50
proteolysis site
118/119
Insulin
54-110
IGF2_C
147-202

(C)

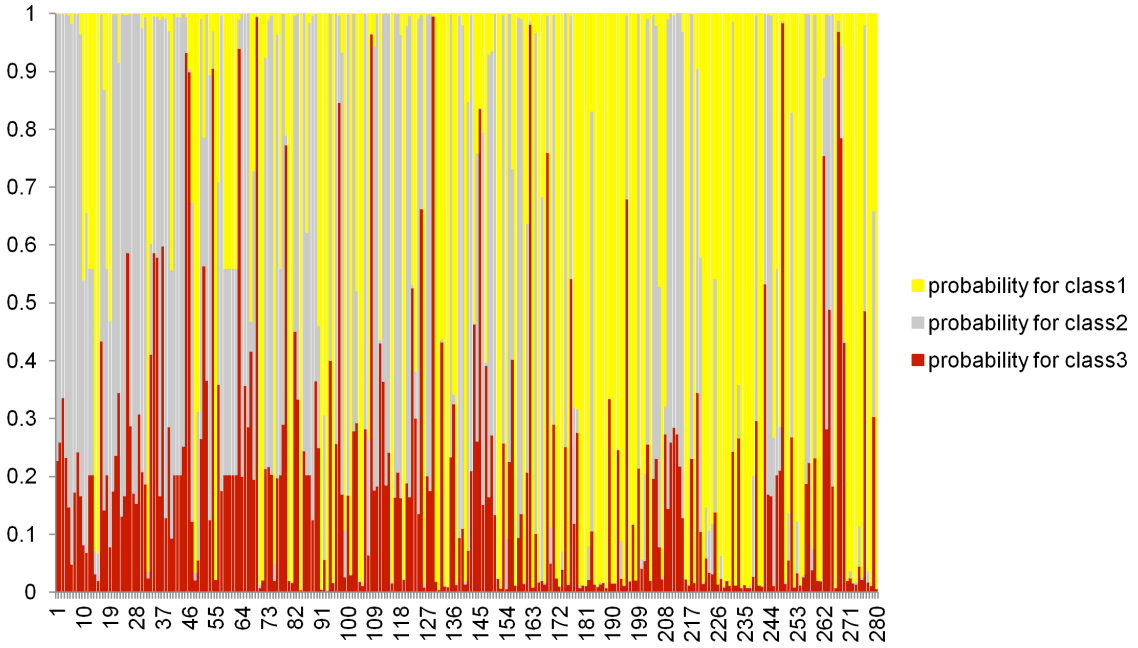

probability for class1
probability for class2
probability for class3
PPARgamma_N
38-100
Glucocorticoid receptor-like(DNA-binding domain
128-210
Nuclear receptor ligand-binding domain
230-525
Strdhormoner
190-200;337-374;426-441;483-500

(D)



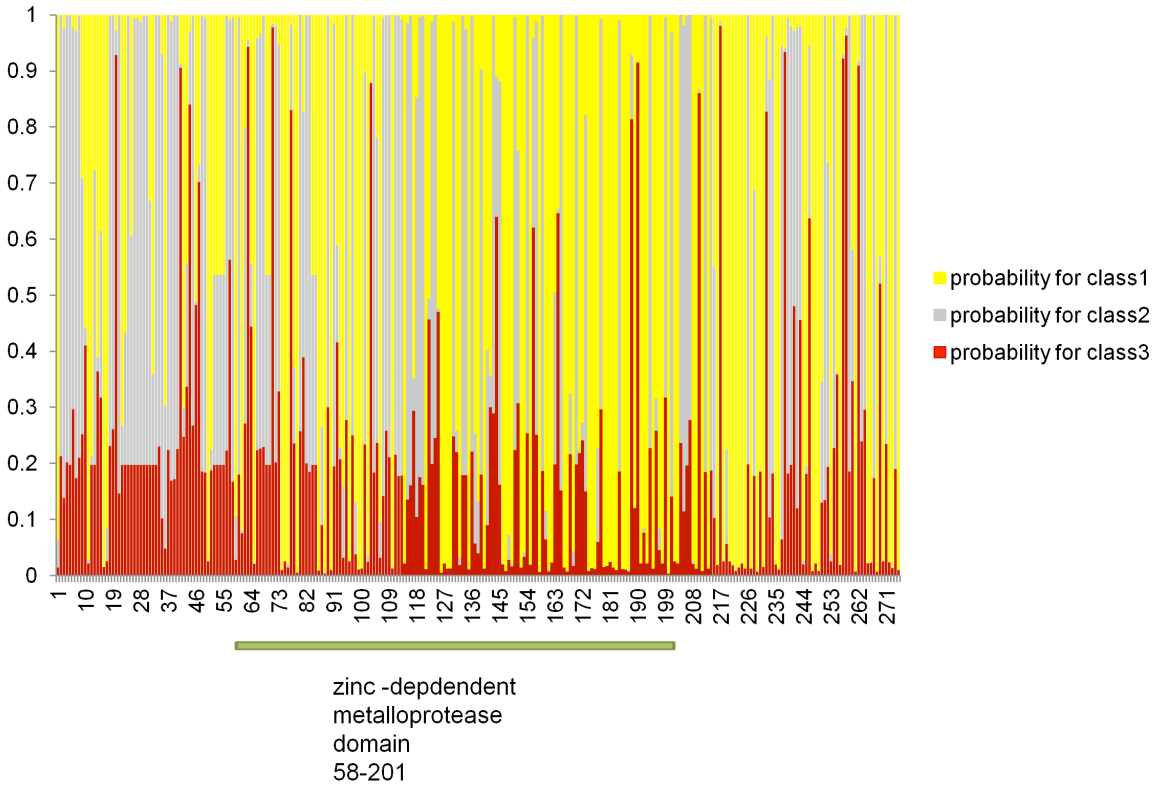
No motif found in InterProScan

(E)

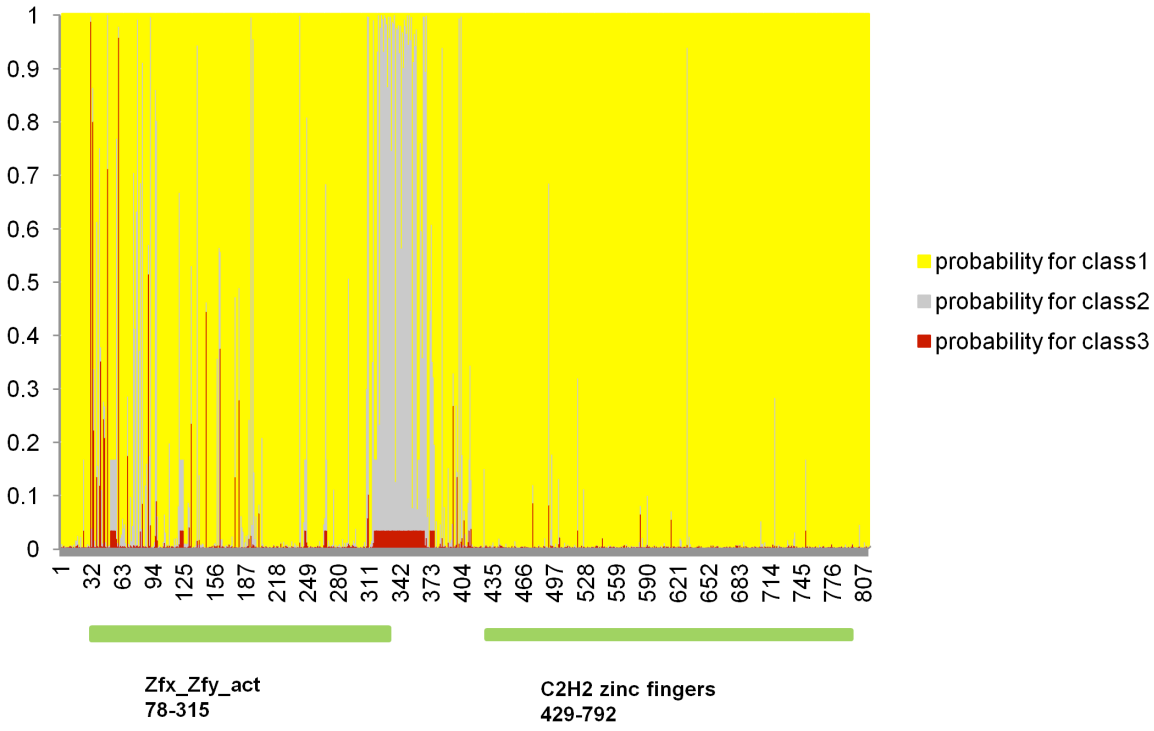


No motif found in InterPro Scan

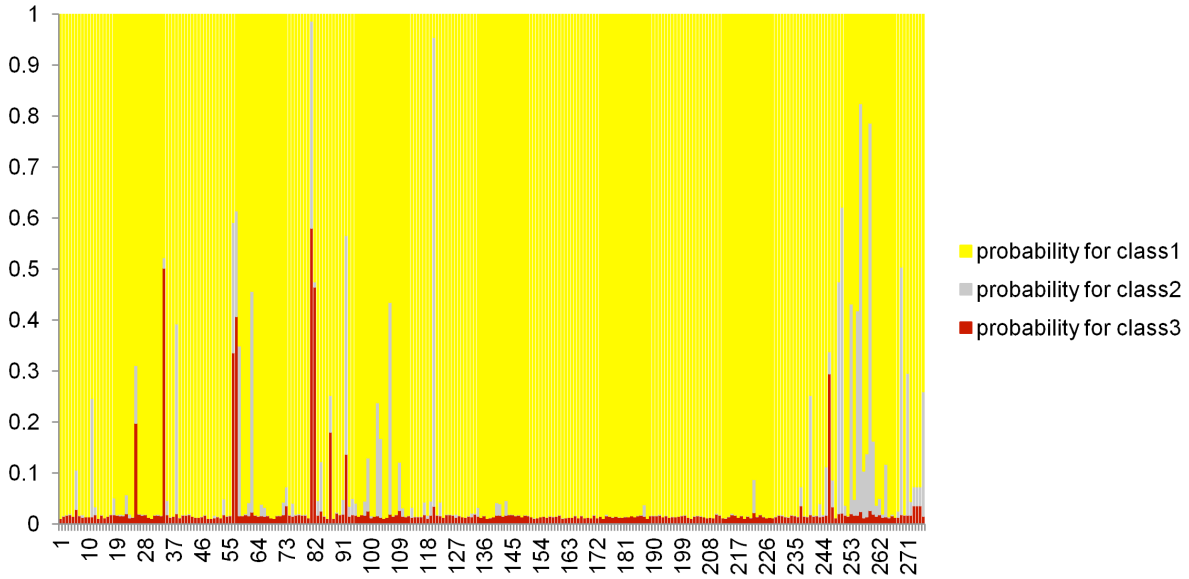
(F)


probability for class1
probability for class2
probability for class3
zinc -depdendent
metalloprotease
domain
58-201

(G)

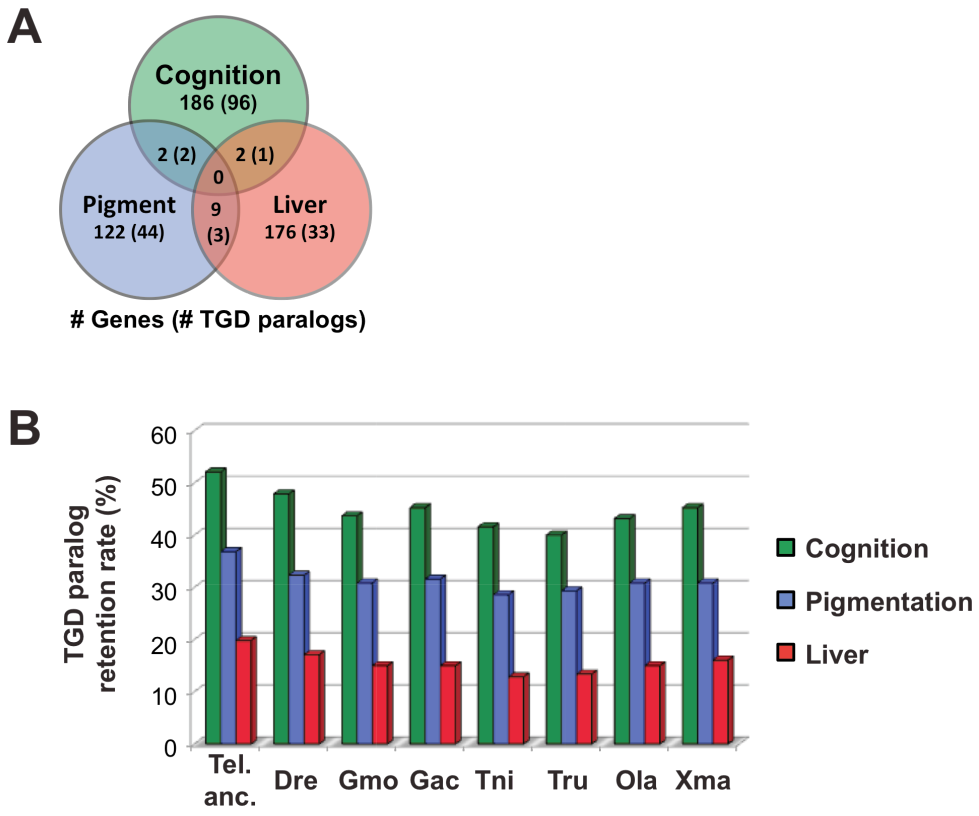

probability for class1
probability for class2
probability for class3
Zfx_Zfy_act
78-315
C2H2 zinc fingers
429-792

(H)


1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
1
10
19
28
37
46
55
64
73
82
91
100
109
118
127
136
145
154
163
172
181
190
199
208
217
226
235
244
253
262
271
probability for class1
probability for class2
probability for class3

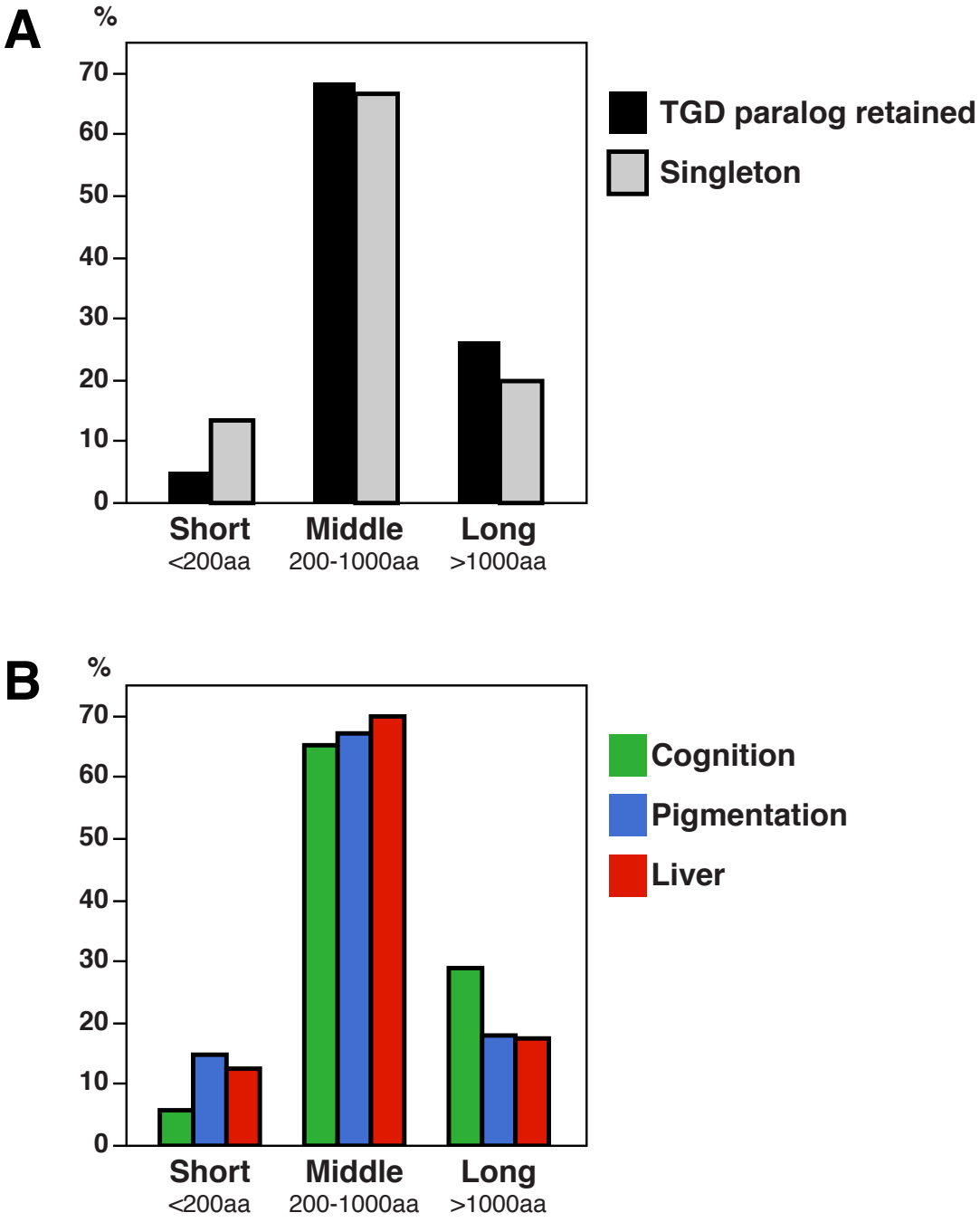
Supplementary Figure 6: Analysis of paralog retention rate after the teleost genome duplication (TGD)

(**A**) Venn diagram of three partially overlapping gene classes (cognition, pigmentation, liver) analyzed for TGD paralog retention. Numbers indicate genes analyzed and (in brackets) number of genes still present in two copies in at least one of the seven teleost genomes. (**B**) TGD paralog retention rate for the three gene classes for zebrafish (*Danio rerio*, Dre), Atlantic cod (*Gadus morhua*, Gmo), stickleback (*Gasterosteus aculeatus*, Gac), two pufferfishes (*Tetraodon nigroviridis*, Tni; *Takifugu rubripes*, Tru), medaka (*Oryzias latipes*, Ola), platyfish (*Xiphophorus maculatus*, Xma), and their inferred teleost ancestor (tel. anc.).



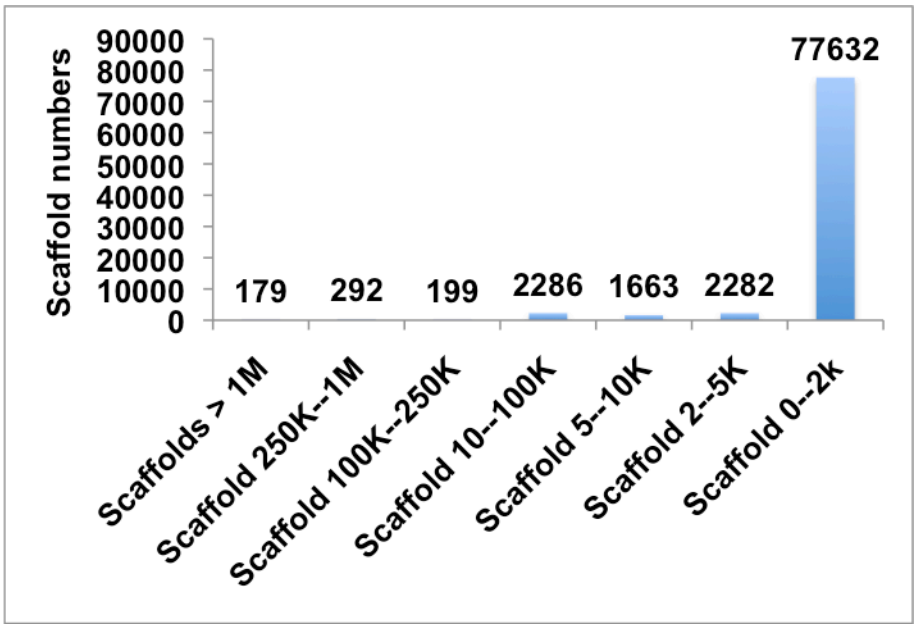
Supplementary Figure 7: Protein length analysis

(A) Retained TGD paralogs vs. singletons. (B) Functional categories.



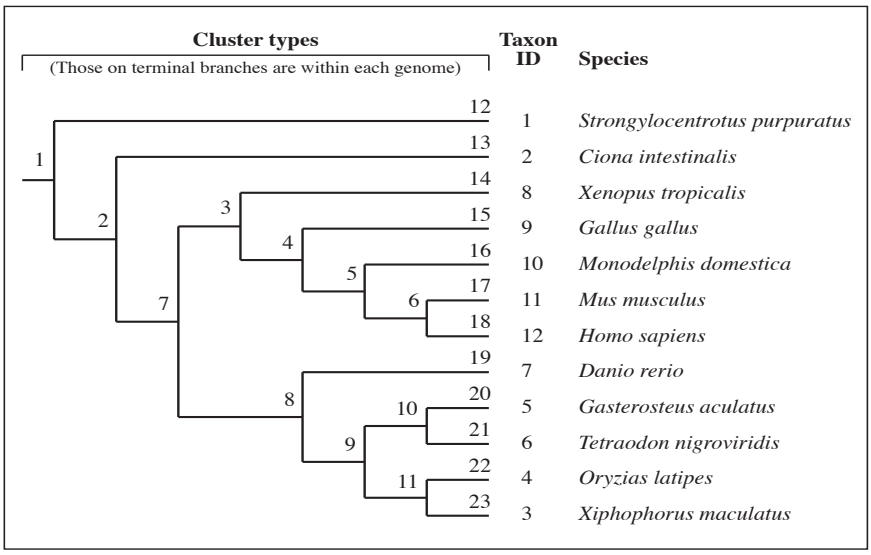
Supplementary Figure 8: Whole genome assembled scaffold distribution by base length.

Each scaffold represents individual contigs joined by long distance paired sequences.



Supplementary Figure 9: Species tree used for the PHRINGE analysis.

Numerals in parentheses are taxon IDs. Other numerals are cluster levels. Cluster levels 12-23 are internal to each genome.



SUPPLEMENTARY REFERENCES

1. Huang, X., Wang, J., Aluru, S., Yang, S.P. & Hillier, L. PCAP: a whole-genome assembly program. *Genome research* **13**, 2164-70 (2003).
2. Lamatsch, D.K., Steinlein, C., Schmid, M. & Schartl, M. Noninvasive determination of genome size and ploidy level in fishes by flow cytometry: detection of triploid *Poecilia formosa*. *Cytometry* **39**, 91-5 (2000).
3. Tiersch, T.R., Chandler, R.W., Kallman, K.D. & Wachtel, S.S. Estimation of nuclear DNA content by flow cytometry in fishes of the genus Xiphophorus. *Comparative biochemistry and physiology. B, Comparative biochemistry* **94**, 465-8 (1989).
4. Garcia, T.I. *et al.* Effects of short read quality and quantity on a de novo vertebrate transcriptome assembly. *Comp Biochem Physiol C Toxicol Pharmacol* **155**, 95-101 (2012).
5. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology* **10**, R25 (2009).
6. Zerbino, D.R. & Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* **18**, 821-9 (2008).
7. Trapnell, C., Pachter, L. & Salzberg, S.L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105-11 (2009).
8. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology* **28**, 511-5 (2010).
9. Stanke, M. & Waack, S. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* **19 Suppl 2**, ii215-25 (2003).
10. Kent, W.J. BLAT - the BLAST-like alignment tool. *Genome Research* **12**, 656-64 (2002).
11. Schulz, M.H., Zerbino, D.R., Vingron, M. & Birney, E. Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics* **28**, 1086-92 (2012).
12. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. Basic local alignment search tool. *Journal of Molecular Biology* **215**, 403-10 (1990).
13. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
14. Wu, T.D. & Watanabe, C.K. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**, 1859-75 (2005).
15. Wicker, T. *et al.* The repetitive landscape of the chicken genome. *Genome Research* **15**, 126-36 (2005).
16. Aparicio, S. *et al.* Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. *Science* **297**, 1301-10 (2002).
17. Catchen, J.M., Conery, J.S. & Postlethwait, J.H. Automated identification of conserved synteny after whole-genome duplication. *Genome Research* **19**, 1497-505 (2009).
18. Walter, R.B. *et al.* A microsatellite genetic linkage map for Xiphophorus. *Genetics* **168**, 363-72 (2004).
19. Wu, T.D. & Nacu, S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* **26**, 873-81 (2010).
20. Kent, W.J. BLAT--the BLAST-like alignment tool. *Genome Research* **12**, 656-64 (2002).

21. Drummond, A.J. *et al.* Geneious v5.5. *Available from http://www.geneious.com* (2010).
22. Yang, Z., Wong, W.S. & Nielsen, R. Bayes empirical bayes inference of amino acid sites under positive selection. *Molecular Biology and Evolution* **22**, 1107-18 (2005).
23. Braasch, I. & Postlethwait, J.H.I.e. Polyploidy in fish and the teleost genome duplication. in *Polyploidy and Genome Evolution* (eds. Soltis, P.S. & Soltis, D.E.) in press (Springer, 2012).
24. Papp, B., Pal, C. & Hurst, L.D. Dosage sensitivity and the evolution of gene families in yeast. *Nature* **424**, 194-7 (2003).
25. Makino, T. & McLysaght, A. Ohnologs in the human genome are dosage balanced and frequently associated with disease. *PNAS U.S.A.* **107**, 9270-4 (2010).
26. Brown, K.H. *et al.* Extensive genetic diversity and substructuring among zebrafish strains revealed through copy number variant analysis. *PNAS U.S.A.* **109**, 529-34 (2012).
27. Sato, Y., Hashiguchi, Y. & Nishida, M. Temporal pattern of loss/persistence of duplicate genes involved in signal transduction and metabolic pathways after teleost-specific genome duplication. *BMC Evolutionary Biology* **9**, 127 (2009).
28. Braasch, I., Brunet, F., Volff, J.N. & Schartl, M. Pigmentation pathway evolution after whole-genome duplication in fish. *Genome Biology and Evolution*, 479-493 (2009).
29. Volff, J.N., Bouneau, L., Ozouf-Costaz, C. & Fischer, C. Diversity of retrotransposable elements in compact pufferfish genomes. *Trends in genetics : TIG* **19**, 674-8 (2003).
30. Shirak, A. *et al.* Identification of repetitive elements in the genome of Oreochromis niloticus: tilapia repeat masker. *Marine Biotechnology* **12**, 121-5 (2010).
31. Consortium, I.C.G.S. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* **432**, 695-716 (2004).
32. Ivics, Z., Izsvák, Z. & Hackett, P.B. Repeated sequence elements in zebrafish and their use in molecular genetic studies. *The Zebrafish Science Monitor* **3**(1995).
33. Waterston, R.H. *et al.* Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520-62 (2002).
34. Davidson, W.S. *et al.* Sequencing the genome of the Atlantic salmon (Salmo salar). *Genome Biology* **11**, 403 (2010).
35. Carroll, D., Knutzon, D.S. & Garrett, J.E. Transposable elements in Xenopus species. in *Mobile DNA* (eds. **Howe, M.** & **Berg, D.**) 567-574 (1989).
36. Consortium, I.H.G.S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921 (2001).
37. Shibata, Y. *et al.* Identification and cDNA cloning of alveolin, an extracellular metalloproteinase, which induces chorion hardening of medaka (Oryzias latipes) eggs upon fertilization. *J Biol Chem* **275**, 8349-54 (2000).
38. Romano, M., Rosanova, P., Anteo, C. & Limatola, E. Vertebrate yolk proteins: a review. *Mol Reprod Dev* **69**, 109-16 (2004).
39. Beck, F., Erler, T., Russell, A. & James, R. Expression of Cdx-2 in the mouse embryo and placenta: possible role in patterning of the extra-embryonic membranes. *Dev Dyn* **204**, 219-27 (1995).

40. Kudo, N., Yasumasu, S., Iuchi, I. & Tanokura, M. Crystallization and preliminary X-ray analysis of HCE-1, a hatching enzyme of medaka fish, Oryzias latipes. *Acta Crystallogr D Biol Crystallogr* **60**, 725-6 (2004).
41. Rinkenberger, J.L., Cross, J.C. & Werb, Z. Molecular genetics of implantation in the mouse. *Dev Genet* **21**, 6-20 (1997).
42. Hanaoka, R. *et al.* Zebrafish gcmb is required for pharyngeal cartilage formation. *Mech Dev* **121**, 1235-47 (2004).
43. Paulesu, L. *et al.* Evidence of H beta 58, a gene involved in mammalian placental development, in the three-toed skink, Chalcides chalcides (Squamata: Scincidae), a viviparous placentotrophic reptile. *Placenta* **22**, 735-41 (2001).
44. Cross, J.C. *et al.* Genes, development and evolution of the placenta. *Placenta* **24**, 123-30 (2003).
45. Guillemot, F., Nagy, A., Auerbach, A., Rossant, J. & Joyner, A.L. Essential role of Mash-2 in extraembryonic development. *Nature* **371**, 333-6 (1994).
46. Hou, Z., Romero, R., Uddin, M., Than, N.G. & Wildman, D.E. Adaptive history of single copy genes highly expressed in the term human placenta. *Genomics* **93**, 33-41 (2009).
47. Barak, Y. *et al.* PPAR gamma is required for placental, cardiac, and adipose tissue development. *Mol Cell* **4**, 585-95 (1999).
48. Koh, D., Inohaya, K., Imai, Y. & Kudo, A. The novel medaka transglutaminase gene is expressed in developing yolk veins. *Gene Expr Patterns* **4**, 263-6 (2004).
49. Zhang, Z.B. *et al.* [Gene cloning, sequence analysis and tissue expression of estrogen-related receptor alpha (Erralpha) in Japanese medaka and its transcriptional responses after differential EDCs exposure]. *Huan Jing Ke Xue* **29**, 3153-8 (2008).
50. Hyllner, S.J., Westerlund, L., Olsson, P.E. & Schopen, A. Cloning of rainbow trout egg envelope proteins: members of a unique group of structural proteins. *Biol Reprod* **64**, 805-11 (2001).
51. Kanamori, A., Naruse, K., Mitani, H., Shima, A. & Hori, H. Genomic organization of ZP domain containing egg envelope genes in medaka (Oryzias latipes). *Gene* **305**, 35-45 (2003).
52. Fu, X. *et al.* Estimating accuracy of RNA-Seq and microarrays with proteomics. *BMC Genomics* **10**, 161 (2009).
53. Galtier, N., Gouy, M. & Gautier, C. SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. *Comput Appl Biosci* **12**, 543-548 (1996).