BMC Genomics

RESEARCH

Open Access

De novo genome hybrid assembly and annotation of the endangered and euryhaline fish *Aphanius iberus* (Valenciennes, 1846) with identification of genes potentially involved in salinity adaptation

Alfonso López-Solano[1*], Ignacio Doadrio[1], Tessa Lynn Nester[1] and Silvia Perea[1,2]

Abstract

Background The sequencing of non-model species has increased exponentially in recent years, largely due to the advent of novel sequencing technologies. In this study, we construct the Reference Genome of the Spanish toothcarp (*Aphanius iberus* (Valenciennes, 1846)), a renowned euryhaline fish species. This species is native to the marshes along the Mediterranean coast of Spain and has been threatened with extinction as a result of habitat modification caused by urbanization, agriculture, and its popularity among aquarium hobbyists since the mid-twentieth century. It is also one of the first Reference Genome for Euro-Asian species within the globally distributed order Cyprinodontiformes. Additionally, this effort aims to enhance our comprehension of the species' evolutionary ecology and history, particularly its remarkable adaptations that enable it to thrive in diverse and constantly changing inland aquatic environments.

Results A hybrid assembly approach was employed, integrating PacBio long-read sequencing with Illumina short-read data. In addition to the assembly, an extensive functional annotation of the genome is provided by using AUGUSTUS, and two different approaches (InterProScan and Sma3s). The genome size (1.15 Gb) is consistent with that of the most closely related species, and its quality and completeness, as assessed with various methods, exceeded the suggested minimum thresholds, thus confirming the robustness of the assembly. When conducting an orthology analysis, it was observed that nearly all genes were grouped in orthogroups that included genes of genetically similar species. GO Term annotation revealed, among others, categories related with salinity regulation processes (ion transport, transmembrane transport, membrane related terms or calcium ion binding).

Conclusions The integration of genomic data with predicted genes presents future research opportunities across multiple disciplines, such as physiology, reproduction, disease, and opens up new avenues for future studies in comparative genomic studies. Of particular interest is the investigation of genes potentially associated with salinity adaptation, as identified in this study. Overall, this study contributes to the growing database of Reference Genomes,

*Correspondence:
Alfonso López-Solano
alfonso.lopez@mncn.csic.es
Full list of author information is available at the end of the article

provides valuable information that enhances the knowledge within the order Cyprinodontiformes, and aids in improving the conservation status of threatened species by facilitating a better understanding of their behavior in nature and optimizing resource allocation towards their preservation.

Keywords Reference genome, *Aphanius iberus*, Cyprinodontiformes, De novo hybrid assembly, Annotation

Background

Recent advances in genomics have created an unprecedented opportunity to resolve long-standing questions regarding the evolutionary processes of organisms. These questions were previously difficult to understand without comprehensive genomic analyses. Specifically, it is particularly interesting the study of unusual adaptations in species which thrive in dynamic and ever-changing habitats, such as the genetic basis of fish adaptations, such as responses to hypoxia and air exposure, and fishes facing fluctuating salinity levels in brackish waters [1–3]. In light of the ongoing global warming that has resulted in significant and rapid environmental transformations, this kind of investigation has become increasingly relevant. Aquatic ecosystems, especially those in inland environments, are particularly vulnerable to the adverse effects of climate change and face immense threats, largely due to anthropogenic pressures that have significantly altered and transformed them [4]. Climate change in the Mediterranean region is increasing the frequency and intensity of catastrophic and unpredictable events, such as cut-off low storms (DANAs) and droughts.

Fish populations are especially susceptible to the consequences of climate change, including rising average temperatures, alterations in biological oxygen demand, and greater fluctuations in salinity. These challenges are particularly significant for euryhaline fish species, renowned for their remarkable adaptability to extensive salinity and temperature ranges. For instance, these species possess intricate physiological mechanisms that enable them to efficiently regulate their osmotic balance in response to fluctuations in water salinity [5–12]. Nonetheless, alterations in salinity patterns triggered by climate change disrupt the delicate equilibrium that is essential for the well-being of euryhaline fishes [13]. This underscores the need for a more comprehensive understanding of the functioning of various adaptation mechanisms that different species have evolved to survive in these highly distinctive environments, as well as the enhancement of our understanding of their stress resistance [1, 3]. Additionally, numerous fish species that inhabit coastal lagoons, salty rivers, and other coastal water bodies are experiencing drastic population declines, posing a severe threat to their survival [14–17].

The Spanish toothcarp, *Aphanius iberus* (Valenciennes, 1846), is one of those singular species that has evolved unique traits enabling it to thrive in highly variable, dynamic, and demanding habitats along the eastern coast of Spain, including groundwater springs (locally known as ullals), coastal lagoons, river mouths, and even salt marshes [18] (Fig. 1). The species is classified as "Endangered" (IUCN category: EN; National and European legislation—Habitats Directive of the Council of Europe, Act 1992; National Catalog of Threatened Species, Act 2011) due to various factors, including pollution, habitat destruction, unregulated management by aquarium enthusiasts, and the presence of invasive species, which are displacing it from its natural environment to more hostile habitats [18–20]. Intensive management programs, such as habitat restoration and stocking, are in place to protect the species [21, 22]. Despite its conservation status and remarkable adaptability, the Spanish toothcarp has not been recognized as a model species for genetic studies. Nevertheless, numerous investigations have been conducted to explore its genetic structure and diversity using traditional methods such as allozymes, mitochondrial genes and microsatellites [19, 23–26], alongside contemporary Next-generation sequencing methods [27, 28]. Analyses of the species genetic structure based on different markers have been conducted in particular within the context of conservation [29–31]. Nevertheless, evolutionary questions regarding its remarkable adaptability to the variable and changing environmental conditions in which the species naturally inhabits remain unanswered. One potential avenue for addressing this knowledge gap is through the analysis of loci under selection in a broad genome context to understand how populations response to environmental changes in the ongoing global warming. In the meantime, analogous genetic studies have been conducted in other fish species with the objective of deepening our understanding of their biology and expanding our knowledge of comparative genomics, adaptations genomics, and functional gene variation [28, 32–35], among others. Subsequently, there has been a notable increase in the number of publications related to Reference Genomes. The total number of Reference Genomes available on GenBank in the order Cyprinodontiformes has grown exponentially. The findings of these studies can be applied to *A. iberus*, a species with an exceptional capacity for adaptation to fluctuating

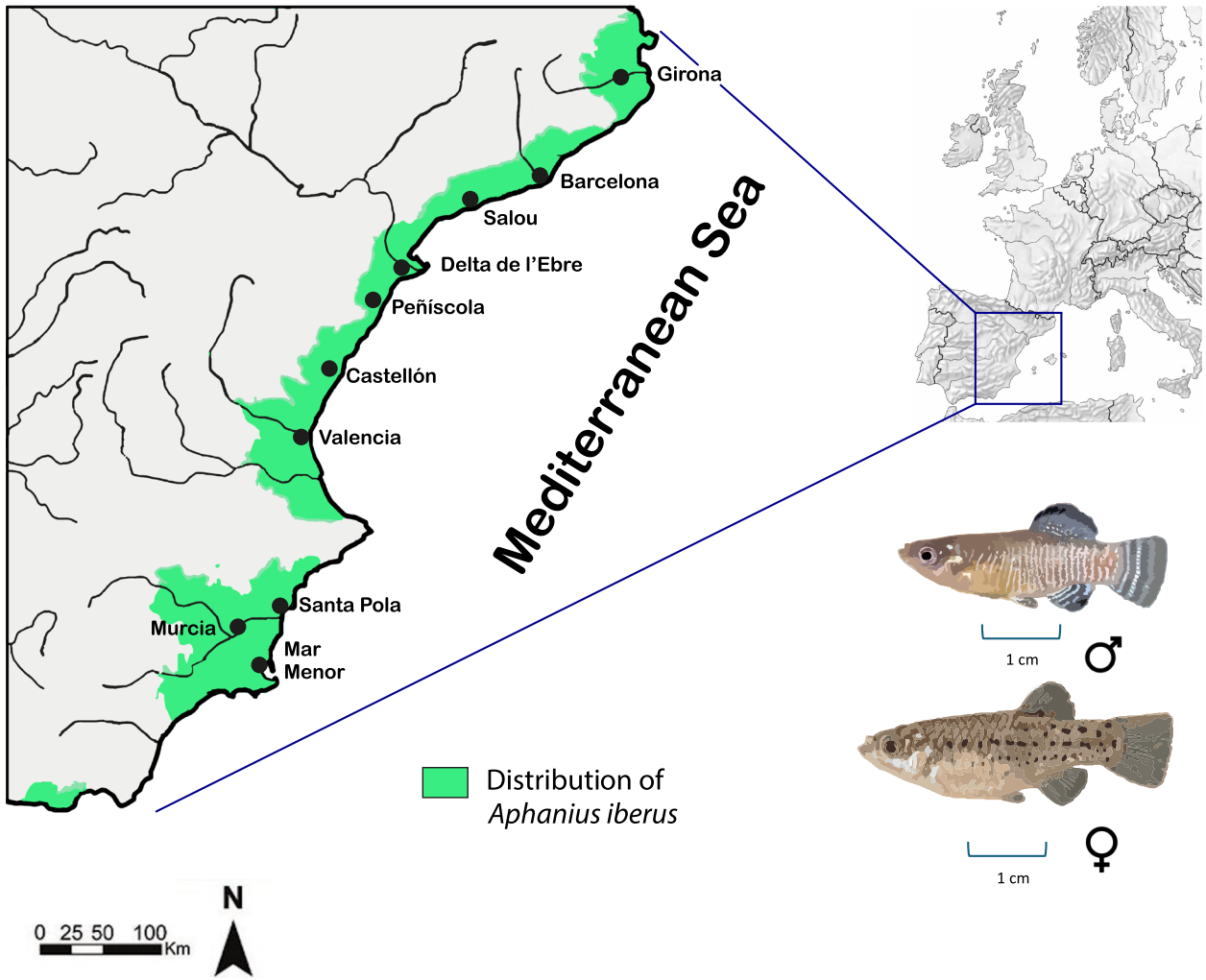

Fig. 1 Current distribution of the Spanish toothcarp (*Aphanius iberus*) along the eastern coast of the Iberian Peninsula. Bottom right are a male (top) and a female (bottom) individuals of *A. iberus*

saline environments, and other limiting environmental conditions, in a wider evolutionary context within the Cyprinodontiformes order underscoring the necessity for a Reference Genome of the species.

In the present study, a high-quality genome assembly of *A. iberus* was generated for the first time using a de novo hybrid assembly strategy that combined both high-coverage Pacific Biosciences (PacBio) long-read sequencing with precise Illumina short-read data. The strategy of combining two or more sequencing techniques has significantly increased the availability of Reference Genomes, primarily driven by the substantial advancements in the accuracy and cost-effectiveness of genome sequencing [36, 37]. Furthermore, gene prediction was performed using AUGUSTUS (3.2.3) [38] and, in order to improve robustness, two different approaches, InterProScan v5.50–84.0 [39] and Sma3s v2 [40], were utilized for functional annotation of the genes. Additionally, various analyses, including orthologous comparisons and the construction of a phylogenetic tree, were performed to compare the newly sequenced genome with previously published and well-annotated genomes of closely related genera. The fully sequenced and annotated genome of the Spanish toothcarp, the first Euro-Asian species of the order Cyprinodontiformes to be released, provides an invaluable genetic resource for investigating the mechanisms of evolutionary adaptation in these species, which may be linked to the dynamic history of the coastline in this region, as well as facilitating future studies in ecology, phylogenetics or evolution.

Methods

DNA extraction

DNA was extracted from muscle tissue from three samples collected at the Centro de Conservación de Especies Dulceacuícolas which belongs to the Government of Valencia (Spain) ("Piscifactoría de El Palmar"), Valencia, Spain, serving as a genetic refuge for various populations of the Spanish toothcarp. The three specimens come from the Albuixech population, which represents the most widely distributed genetic lineage of the Spanish toothcarp [19, 24, 29]. The fishes were euthanized using 0.1% tricaine methanesulfonate (MS222) following the standard internationally approved protocols by qualified personal at the mentioned center in Valencia and posteriorly sent to the National Museum of Natural Sciences in Madrid. DNA isolation was performed using the MagAttract HMW DNA isolation kit (Qiagen) and the final elution step was carried out with a volume of 100 µL. DNA quantification was performed using the Qubit High Sensitivity dsDNA Assay (Thermo Fisher Scientific).

Library preparation and genome assembly

We opted for sequencing the data from PacBio and Illumina DNA sequencing in order to obtain trustworthy assembly and annotation. Prior to library preparation, the sample was further purified and size-selected to keep the largest fragments.

PacBio and Illumina library preparation and sequencing

For library preparation, the SMRTbell Express Template Prep Kit 2.0 (PacBio) was utilized in accordance with the manufacturer's instructions. Subsequently, sequencing was performed on a Sequel II sequencer (PacBio) with a SMRT Cell 8M, using the Long-reads mode. The Illumina DNA Prep kit was used for preparing the Illumina library in strict accordance with the manufacturer's guidelines. The Agilent 2100 Bioanalyzer was used to verify the library's fragment size distribution and concentration using the Agilent HS DNA Kit. Then, the library was sequenced on a portion of a NovaSeq PE150 flow cell with a total output target of 50 Gb.

Raw data pre-processing

The *de novo* genome sequencing using the Illumina NovaSeq platform generated a total of 436,121,806 paired-end reads (R1+R2). The raw fastq files underwent quality assessment using the software FastQC v0.11.5 [41]. In addition, 7,307,982 subreads were obtained by *de novo* genome sequencing using the PacBio Sequel II platform. The PacBio reads were quality-checked using SequelTools software [42], giving the longest subread of 50.93Mbp with a mean read length of 9,349bp and a N50 value of 13,224bp (see Supplementary Table 1).

De novo hybrid assembly

Two distinct *de novo* approaches were utilized to assemble the genome of *A. iberus.* Firstly, the short and long genomic sequencing reads were assembled de novo into "mega-reads" using the software MaSuRCA v3.4.2 [43], which incorporates the advantages of combining de Bruijn graph and Overlap-Layout-Consensus assembly approaches. QuorUM software [44] was used following the instructions, to error-correct the short reads, which were then extended into "super-reads" [45] and aligned to the long reads. Consistent alignments of "super-reads" were merged into "mega-reads," and Flye v2.5 [46], implemented in MaSuRCA, was used to assemble the "mega-reads". For the second approach, it was used the software HASLR [47]. This software uses a new data structure called a backbone graph in addition to the de Bruijn graph and the SPOA algorithm.

After generating the assemblies, both MaSuRCA and HASLR were polished using POLCA [48], which involved aligning the raw Illumina reads to the assembly using the BWA mem algorithm [49], and calling short variants using FreeBayes [50]. Subsequently, the quality of the assemblies, both before and after polishing, was assessed using QUAST 5.0.2 [51]. The number of scaffolds obtained were 3,026 for MaSuRCA polished and 8,313 for HASLR polished. The total length of the genome assembly was 1,198,861,738bp and 1,131,748,719bp respectively for each method (Supplementary Table 2). The Scaffold N50 and L50 were 1,678,775bp and 201 for MaSuRCA polished and 284,695bp and 1,180 for HASLR polished. The GC content was 39.17% and 39.12% respectively after each polishing process.

The software KMC ver. 3.1.1 [52] was employed to enumerate the frequency of *k-mers* in the corrected reads, with the *k-mer* size parameter set to k=21. This analysis helps to detect sequencing errors, contamination, or repetitive sequences, and aids in determining whether the genome assembly process was successful. The resulting *k-mer* profile was generated using GenomeScope 2.0 [53] and is depicted in Supplementary Figure 1.

Quality control of the hybrid assembly

The genomic coverage of each region was determined by mapping short sequencing reads to the genome assembly produced by MaSuRCA, using BWA v0.7.15 [54], and the mapping statistics were calculated with SAMtools 1.3.1 [55].

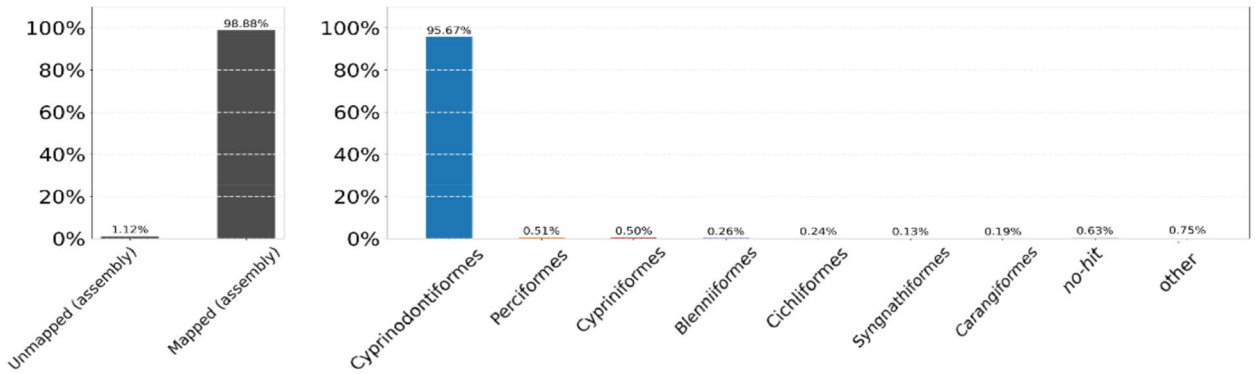

Fig. 2 ReadCovPlot of the *A. iberus* assembly, displaying on the left the proportion of mapped and unmapped reads against the assembly, and on the right the percentage of mapped reads assigned to different taxa at the rank of 'order'

To identify potential contaminant sequences in the assembly, BlobTools v1.1.1 was used to create a scatter plot and a bar chart (see Figure 2 and Supplementary Figure 2). The scatter plot represents assembly contigs/scaffolds as dots colored according to their taxonomic affiliation, based on sequence similarity search results. The bar chart illustrates the sequences mapped and unmapped against the assembly and mapped sequences assigned to different taxa.

The completeness and quality of the genome assembly were assessed using the "genome" mode of BUSCO V5.5 (Benchmarking Universal Single-Copy Orthologs) V5.beta.1 [56]. To assess the results of *A. iberus* and to make comparisons between species, BUSCO analysis was conducted on three closely related species: *Cyprinodon variegatus* (NCBI GenBank: GCA_000732505.1), *Poecilia formosa* (NCBI GenBank: GCA_000485575.1) and *Xiphophorus maculatus* (NCBI GenBank: GCA_002775205.2). To predict eukaryotic genes, the Metaeuk pipeline [57] was utilized with default parameters and the lineage-specific cyprinodontiformes_odb10 database (last update 2021-02-19) [56].

To conduct a comprehensive genome-wide comparison, we utilized *Poecilia reticulata,* a closely related species with a chromosome-level well-characterized genome. The complete Reference Genome sequence for *P. reticulata* was obtained from NCBI GenBank (GCF_000633615.1) and aligned against the newly assembled genome of *A iberus* using minimap2 [58]. This alignment served as input for D-GENIES [59] in order to compare similarity between both genomes.

Repetitive elements

De novo identification of repetitive elements in the assembly was performed using RepeatModeler v2.0.1 [60]. The assembly was then masked using RepeatMasker v4.1.2-p1 [61] and the Repbase-20170127 library of known repeats [62]. The number and length of masked repeats, classified by repeat class, are reported in Supplementary Table 3.

Genome annotation

The first step in genome annotation in a given genomic sequence is to predict all gene structures. Gene prediction was conducted using AUGUSTUS v.3.2.3 [38] which defines probability distributions for the different sections of genomic sequences (i.e. exons, introns, intergenic regions) based on a generalized Hidden Markov Model. *Danio rerio* was the reference species used as a setting. The result obtained was a gff (General Feature Format) which was employed as an input for the utility gffread v0.12.6 [63]. This utility extracts the sequence of all transfrags, which are transcripts or fragments that result from the assembly process, generating a FASTA file with all the predicted sequences. To identify candidate coding regions within these predicted mRNA sequences, TransDecoder v5.5.0 [64] was used. This software is particularly useful in the analysis of incomplete genomes or in the identification of new genes in non-model or understudied species.

To gain the maximum possible information about the biological function of the predicted genes, two different approaches were performed to functionally annotate them:

The precited protein-coding genes were initially functionally annotated using the software InterProScan v5.50–84.0 [39]. The annotation was performed using general-content databases, including: the conserved domain database (CDD) [65], the Coils database [66], the Gene3D database [67], the HAMAP database [68], the MobiDBLite database [69], the protein analysis through evolutionary relationships (PANTHER) classification system [70], the protein families database (Pfam) [71], the protein information resource

and superfamily (PIRSF) classification system [72], the protein motif fingerprints (PRINTS) database [73], the protein domains, families and functional sites (ProSitePatterns and ProSiteProfiles) databases [74], the structure function linkage (SFLD) database [75], the simple modular architecture research tools (SMART) [76], the SUPERFAMILY database [77], and the TIGRFAM database [78].

The second annotation was performed using the software Sma3s v2 [40], and the complete manually annotated and reviewed Swiss-Prot database from UniProtKB [79]. Sma3s reports a summary with different categories and the number of sequences belonging to each functional category. Sequence annotations also contain the most probable gene name and the most probable description (including putative EC enzyme codes).

Gene Ontology (GO) Terms were generated for both annotation methods, and genes were classified into three categories, GO Function, which includes the genes in general categories of Molecular Function; GO Process, which includes genes in different categories of Biological Processes; and GO Component, which includes genes in different categories of Cellular Components. Many genes were classified in more than one category due to its multifunctionality.

The information obtained from the performed annotations is compiled in a table accessible through DIGITAL.CSIC: "Annotation_Aphanius_iberus.xslx" (http://hdl.handle.net/10261/365271). The table includes 73,242 predicted genes, with each gene's location within a Scaffold, its length, strand orientation (+ or -), mRNA sequence, and all annotation features conducted with InterProScan and Sma3s, as described previously.

Phylogenomics and comparative genomics

A set of fourteen species from various freshwater fish genera representative of different orders, for which Reference Genome data were available on NCBI GenBank, were selected to conduct different analyses, with a particular focus on the Cyprinodontiformes genera. Afterwards, their complete protein sequences were downloaded from the database (Table 1). The most closely related species was *Cyprinodon variegatus*, from the same order as *A. iberus*. Four other species also belonged to the same order and five more to the same superorder. Predominantly, most species were phylogenetically closely related, however, some distantly related ones were also included in the analysis due to their extensive use as model species in several studies [80–87].

The downloaded protein sequences were aligned using MAFFT/7.475-with-extension [88], where the newly sequenced data from *A. iberus* were also included. Subsequently, TrimAl was employed to evaluate and remove poorly aligned regions [89]. The resulting alignment based on the longest isoform was used as input for the identification of orthologous genes in the studied species. We employed OrthoFinder 2.5.4 [90], to cluster homologous genes from all 14 species through sequence similarity [90], among other functionalities.

A phylogenetic analysis was conducted with all 14 fish species using the orthogroups clustered by Orthofinder. The phylogenetic tree was inferred based on the single-copy orthologous gene sequences implemented in the IQ-tree software [91], with the LG+F+I+G4

Table 1 Genome size comparison to some related species to *A. iberus*

Species	Order	Family	Genome size (Mb)	Accesion number
Aphanius iberus (Valenciennes, 1846)	Cyprinodontiformes	Aphaniidae	1,199	GCA_028564705.1
Cyprinodon variegatus (Lacepède, 1803)	Cyprinodontiformes	Cyprinodontidae	1,035	GCA_000732505.1
Xiphophorus maculatus (Günther, 1866)	Cyprinodontiformes	Poeciliidae	704.3	GCA_002775205.2
Gambusia affinis (Baird and Girard, 1853)	Cyprinodontiformes	Poeciliidae	680.1	GCA_019740435.1
Poecilia formosa (Girard, 1859)	Cyprinodontiformes	Poeciliidae	748.9	GCA_000485575.1
Fundulus heteroclitus (Linnaeus, 1766)	Cyprinodontiformes	Fundulidae	1,203	GCA_011125445.2
Oryzias latipes (Temminck and Schlegel, 1846)	Beloniformes	Adrianichthyidae	734	GCA_002234675.1
Gasterosteus aculeatus (Linnaeus, 1758)	Gasterosteiformes	Gasterosteidae	471.9	GCA_016920845.1
Oreochromis niloticus (Linnaeus, 1758)	Cichliformes	Cichlidae	1,006	GCA_001858045.3
Takifugu rubripes (Temminck and Schlegel, 1850)	Tetraodontiformes	Tetraodontidae	384.1	GCA_901000725.2
Danio rerio (Hamilton-Buchanan, 1822)	Cypriniformes	Danionidae	1,373	GCA_000002035.4
Cyprinus carpio (Linnaeus, 1758)	Cypriniformes	Cyprinidae	1,680	GCA_018340385.1
Astyanax mexicanus (De Filippi, 1853)	Characiformes	Characidae	1,373	GCA_023375975.1
Gadus morhua (Linnaeus, 1758)	Gadiformes	Gadidae	669.9	GCA_902167405.1

substitution model for proteins, as estimated by ModelFinder implemented in IQ-tree [92, 93]. The amino acid sequences were concatenated and treated as a single partition, which is the default setting of ModelFinder. The tree was rooted using a midpoint root approximation and branch support was evaluated with 1000 bootstrap replicates based on the ultrafast algorithm [94].

To assess more accurately the overlap in orthologous genes between *A. iberus* and the rest of species, a new analysis using Orthofinder was performed. This analysis specifically incorporated only two cyprinodontiform closely related species of *A. iberus* (Valenciennes, 1846): Sheepshead minnow *Cyprinodon variegatus* (Lacepède, 1803), and the Amazon molly *Poecilia formosa* (Girard, 1859) along with one species of a closely related genus from its sister order Beloniformes: the Japanese medaka *Oryzias latipes* (Temminck y Schlegel, 1846).

Results

High molecular weight (HMW) DNA was extracted from muscle tissue from three individual samples collected at the Centro de Conservación de Especies Dulceacuícolas, belonging to the Government of Valencia ("Piscifactoría de El Palmar"), Valencia, Spain. We pursued a de novo hybrid assembly strategy employing long-read PacBio and short-read Illumina DNA sequencing, along with two distinct assembly approaches in order to obtain a trustworthy assembly and annotation (See Material & Methods section). The de novo genome sequencing in the PacBio Sequel II platform yielded a total of 7.3 million subreads (mean read length 9.6 kbp; N50 13,224). The de novo genome sequencing in the Illumina NovaSeq platform rendered a total of close to 436 million paired-end reads. After assembly, the genome size was estimated to be 1.15 Gb at 95 × coverage (See Supplementary Table 1 for details in assembly quality). Detailed information, including the total length and assembly statistics of this hybrid genome, can be found in Table 2 and Supplementary Table 3.

Table 2 General statistics of the hybrid assembly

Assembly global statistics	
Total sequence length (bp)	1,198,861,116 bp
Total ungapped length (bp)	1,198,858,230 bp
Gaps between scaffolds	0
Number of scaffolds	3,025
Scaffold N50	1,678,775 bp
Scaffold L50	201
Contig N50	1,625,559 bp
Contig L50	205
GC percent	39%

Different analyses indicated the high accuracy and robustness of the genome assembly and annotation. The plot profile illustrating the observed *k-mer* frequency distribution, based on a *k-mer* size of 21 (the recommended size) in the corrected reads, is presented in Supplementary Fig. 1. The genomic coverage, determined by mapping short sequencing reads to the genome assembly produced by MaSuRCA, resulted in a mapping rate of 98.8% of the read sequences (see Fig. 2 and Supplementary Fig. 1). Furthermore, when SAMtools was employed, the percentage increased to 99.67% of reads mapped back to the assembly. Additionally, Fig. 2 illustrates the percentage of mapped sequences assigned to different taxa, with 95.67% of the sequences belonging to the order Cyprinodontiformes, which indicates a robust affinity of the assembled genomic data with its corresponding taxonomic group (Fig. 2 on the right).

In a comparison of *A. iberus* with three other closely related species in the order Cyprinodontiformes (*Cyprinodon variegatus, Poecilia formosa, and Xiphophorus maculatus*; Table 1), a total of 15,213 BUSCO data sets (n) were analyzed. The results indicated that *A. iberus* exhibited 14,067 complete orthologue genes (C: 92.5%) versus 13,838, 14,455 and 14,423 respectively in the other three species. Of these, 13,975 were single-copy orthologue genes (S: 99.3%), while 92 were duplicated (D: 0.7%) versus 13,712 and 92, 14,289 and 166 and 14,360 and 63 in that order for the other three species. Furthermore, only 0.5% of the remaining orthologue genes were fragmented (F, 73), while approximately 7% were missing (M, 1,073) versus 260 (F) and 1,115 (M) for *C. variegatus*, 91 (F) and 667 (M) for *P. formosa* and 70 (F) and 720 (M) for *X. maculatus* (Fig. 3).

To conduct a comprehensive genome-wide comparison, a closely related species with an extensively characterized genome, including chromosome-level sequencing, was selected: *Poecilia reticulata* (NCBI GenBank: GCF_000633615.1), that was aligned against the newly assembled genome of *A. iberus* and used as an input for D-GENIES [59]. The results based on a 0.75 identity level suggest that these two genomes exhibited high whole-genome scale similarity (Fig. 4).

Approximately half of the genome (49.10%) was composed of repetitive elements. Among these, DNA transposons were the most abundant, comprising for 16.37% of the genome. Tc1-IS630-Pogo was the most predominant DNA transposon (7.22%) (Table 3, Supplementary Table 3). Retroelements accounted for 14.11% of the genome, with L2/CR1/Rex being the most abundant

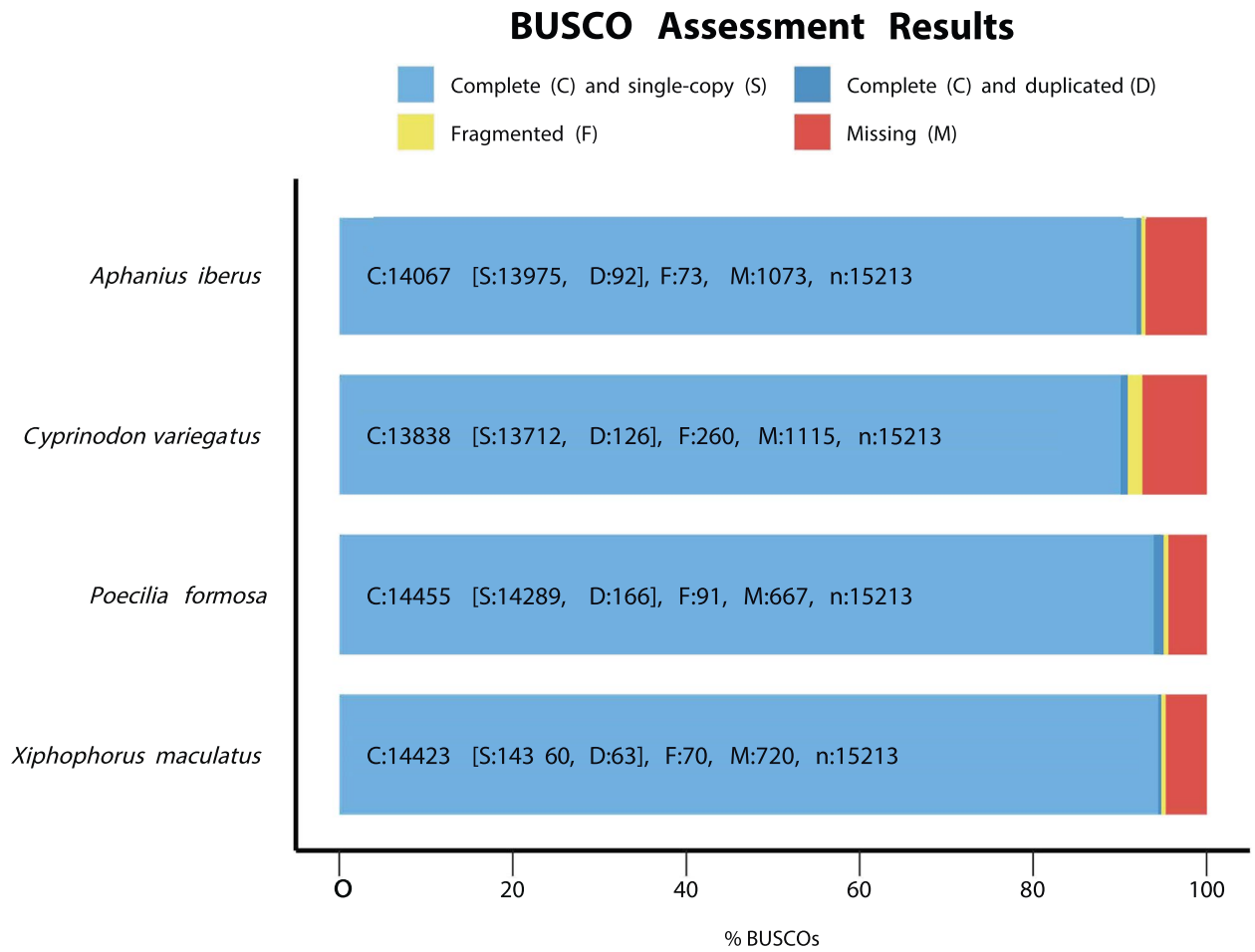

Fig. 3 Summary of the BUSCO analysis results for the *A. iberus* assembly and the lineage dataset cyprinodontiformes_odb10 (created 2021–02–19) in comparison with three closely related species

retroelement within the LINE class, representing a 7.62% of the total abundance. A total of 18.63% of the Repetitive Elements remained unclassified.

Subsequently, gene prediction yielded a total of 73,242 genes (Table 3). Of the total number of genes, 57.41% (42,045 out of 73,242) of them were functionally annotated using either of the two methods (http://hdl.handle.net/10261/365271 to see the whole annotation). Out of the 19,960 genes annotated with Sma3s, at least 2,300 have been associated with salinity.

A total of 1,399, 1,629, and 455 Gene Ontology (GO) Terms were identified for Molecular Function, Biological Processes, and Cellular Component, respectively. The most abundant GO Terms related to Molecular Function encompassed Nucleic acid binding, DNA binding and Protein binding. With regard to Biological Processes, the most prevalent GO Terms were DNA integration, Transposition, DNA-mediated and Regulation of transcription, DNA-templated. In the case of Cellular Component, the most common GO Terms were Internal component of membrane, Membrane and Nucleus (Fig. 5 represents GO Terms for each of the three GO categories, with the most abundant Terms for each category appearing with their percentage). The analysis of GO Terms suggested that the majority of annotated genes are primarily associated with DNA processes of duplication, transcription to RNA, and protein synthesis, as well as the movement of these molecules across membranes.

The Orthofinder analysis clustered genes from all the 14 species (Table 1) into 25,144 orthogroups, with approximately 97,4% of the total genes assigned to at least one orthogroup. A maximum likelihood phylogenetic tree was constructed with IQ-tree using the database of 2,367,048 amino acids obtained from the Orthofinder analyses (Fig. 6). The tree clustered the order Cyprinodontiformes with the Japanese medaka (*Oryzias latipes*) from the order Beloniformes and the Nile tilapia (*Oreochromis niloticus*) from the order Cichliformes, as the most closely related groups.

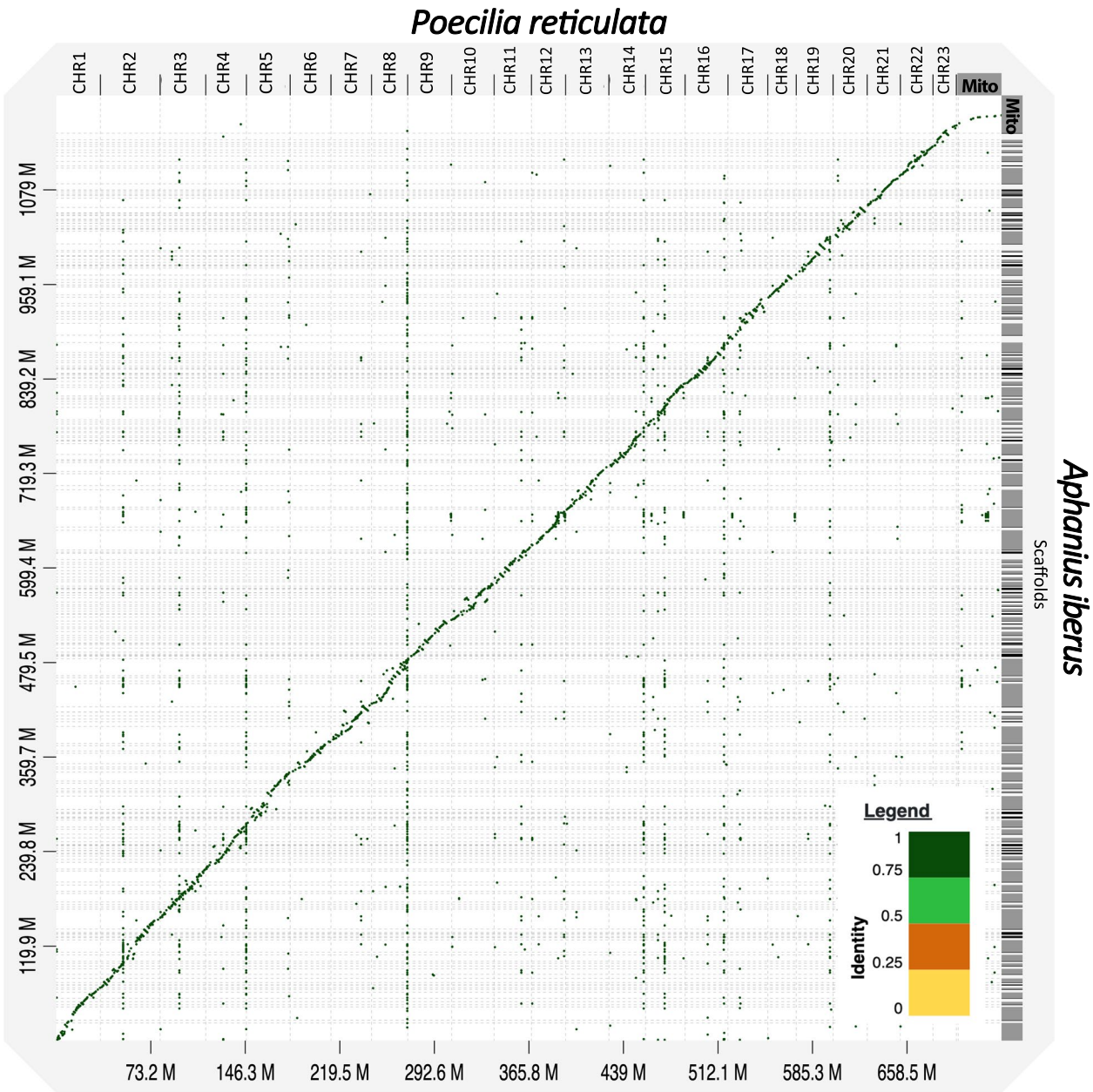

Fig. 4 Dot-plot produced using D-Genies for the comparison of *Poecilia reticulata* chromosome-level assembly (horizontal axis) versus *Aphanius iberus* Scaffold-level assembly (vertical axis). Dotted gridlines represent scaffold/chromosome boundaries. Only sequence similarity higher than 0.75 (darker green dots) are represent in the graphics

The Orthofinder analysis that incorporated only the species, *A. iberus*, *C. variegatus*, *P. formosa* and *O. latipes* showed that of the total number of orthogroups (22,147), approximately 67% (14,864 orthogroups) were shared across all four species. Only about 1% of the orthogroups were found to be species-specific, with 1,213 orthogroups assigned exclusively to *A. iberus,* 83 to *C. variegatus*, 218 to *P. formosa*, and 198 to *O. latipes. A. iberus* shared exclusively 509 orthogroups with *P. formosa*, 401 with *C. variegatus* and 187 with *O. latipes*. Additionally, 1,300 orthogroups were identified as exclusive to the order Cyprinodontiformes (Fig. 7).

Discussion

The de novo hybrid assembly of the genome of the endangered Spanish toothcarp, *A. iberus*, presented in our study, is of paramount importance as it was the first released Reference Genome of a Euro-Asian species

Table 3 Repetitive elements statistics in percentage and the annotation statistics of the newly sequenced genome of *A. iberus*

Repetitive elements	
Total (%)	**49.10**
Retroelements (%)	14.11
SINEs (%)	0.41
LINEs (%)	11.26
LTR elements (%)	2.45
DNA transposons (%)	16.37
Unclassified (%)	18.63
Annotation	
Predicted genes	73,242
Annotated genes	42,045
Mean [median] gene length (bp)	13,532.7 bp [8,018 bp]
Mean [median] exon length (bp)	195.9 bp [128 bp]
Mean [median] intron length (bp)	1,735.3 bp [541 bp]
Mean [median] exons per gene	7.9 [5]
Mean [median] introns per gene	6.9 [4]

within the order Cyprinodontiformes. The only other such genome released is that of the Valencia toothcarp, *Valencia hispanica* (Fig. 1, Fig. 8). This research provides significant insights into the biological understanding and conservation of threatened species such as euryhaline toothcarps, which have a high potential to adapt to different habitat salinity conditions. This is particularly relevant in light of the ongoing impact of salinization variation and seawater intrusion on the Mediterranean coastal wetlands [95, 96].

While genome sequencing has become a crucial resource in fish genomics research, complete genome sequences remain scarce and unevenly distributed across genera within different orders of fishes as well as throughout their global distribution. This is exemplified by the order Cyprinodontiformes, in which the majority of studies and Reference Genomes have been developed for American species [97–100] (Fig. 8). Out of the 79 Reference Genomes available in GenBank for the order Cyprinodontiformes, which includes more than 1900 species [101], only five families account for 70 of them: (Poeciliidae (30), Rivulidae (21) Nothobranchiidae (4), Goodeidae (9) and Cyprinodontidae (6)). The remaining nine Reference Genomes are distributed sparsely across all other six families. There are four families that currently lack fully sequenced genomes, highlighting the need for more diverse genomic resources (Fig. 8).

The genome assembly was of high quality, as evidenced by a comparative analysis of the genome size, the quality and completeness of the sequencing, and the GC content (39%), when compared with related species. The size of the de novo sequenced genome of *A. iberus* (1.15 Gb) is likely to be similar to that of closely related species, such us *Valencia hispanica*: 1,231.84 Mb (GCA_963556495.1), *Fundulus heteroclitus*: 1,203 Mb (GCA_011125445.2), *Cyprinodon brontotheroides*: 1,163 Mb (GCA_018398635.1), *Girardinichthys multiradiatus:* 1,150 Mb (GCA_021462225.2), or *Anableps anableps*: 867.6 Mb (GCA_014839685.1). The quality and completeness of the genome, assessed using various methods, exceed the minimum thresholds (90%) proposed by a recent study for evaluating genome sequenced data quality [102]. The BUSCO analysis revealed that 92.5% are complete gene copies (Fig. 3). Furthermore, BWA and SAMtools mapped back to the assembly 99.67% of short sequencing reads, while BlobTools v1.1.1 mapped 98.88% (Fig. 2). Additionally, the variation in GC content between the *A. iberus* genome and the genomes of the other species analyzed revealed no sequencing-based GC preferences, indicating the high quality of the genome assembly. In previous studies, it has been proposed that assemblies with a minimum N50 value between 200 kb and 1 Mb should be employed to identify big synteny blocks with an error rate below 5% [103]. The subread Scaffold N50 value of the de novo sequencing genome of *A. iberus* was 1,600 Mb, which is similar to the previous sequenced genomes of other species in the same order, such as *Poecilia formosa* (N50: 1,574 Mb, GCA_000485575.1), *Nothobranchius kuhntae* (N50: 1,178 Mb, GCA_006942095.1), *Aphyosemion austral* (N50: 1,435 Mb, GCA_006937985.1), or *Callopanchax toddi* (N50: 1,656 Mb, GCA_006937965.1). The comparison between the genomes of *Poecilia reticulata* and *A. iberus* revealed a significant degree of sequence similarity, despite differences in their assembly levels (scaffolds *vs.* chromosomes). Collectively, these analyses affirmed the precision, robustness, and reliability of the genome assembly.

High percentage of the *A. iberus* genome was constituted by repetitive elements, a typical pattern observed in eukaryotes which can suggest a strong selective pressure [33, 104, 105]. The Tc1 DNA transposon was observed to be present in the second most common group (Tc1-IS630-Pogo). Previously, Tc1 had been identified as one of the most prevalent Repetitive Element in freshwater bony fish when phylogenetic considerations are not taken into account. Besides, the freshwater environment was observed to be a more favorable environment for the proliferation of the Tc1 transposon [33].

To ensure the reliability of the results and to evaluate the quality of the assembly based on sequence homology (orthogroups), orthology comparisons were conducted with closely related species, including well-studied models (Figs. 6 and 7). The results demonstrated a significant

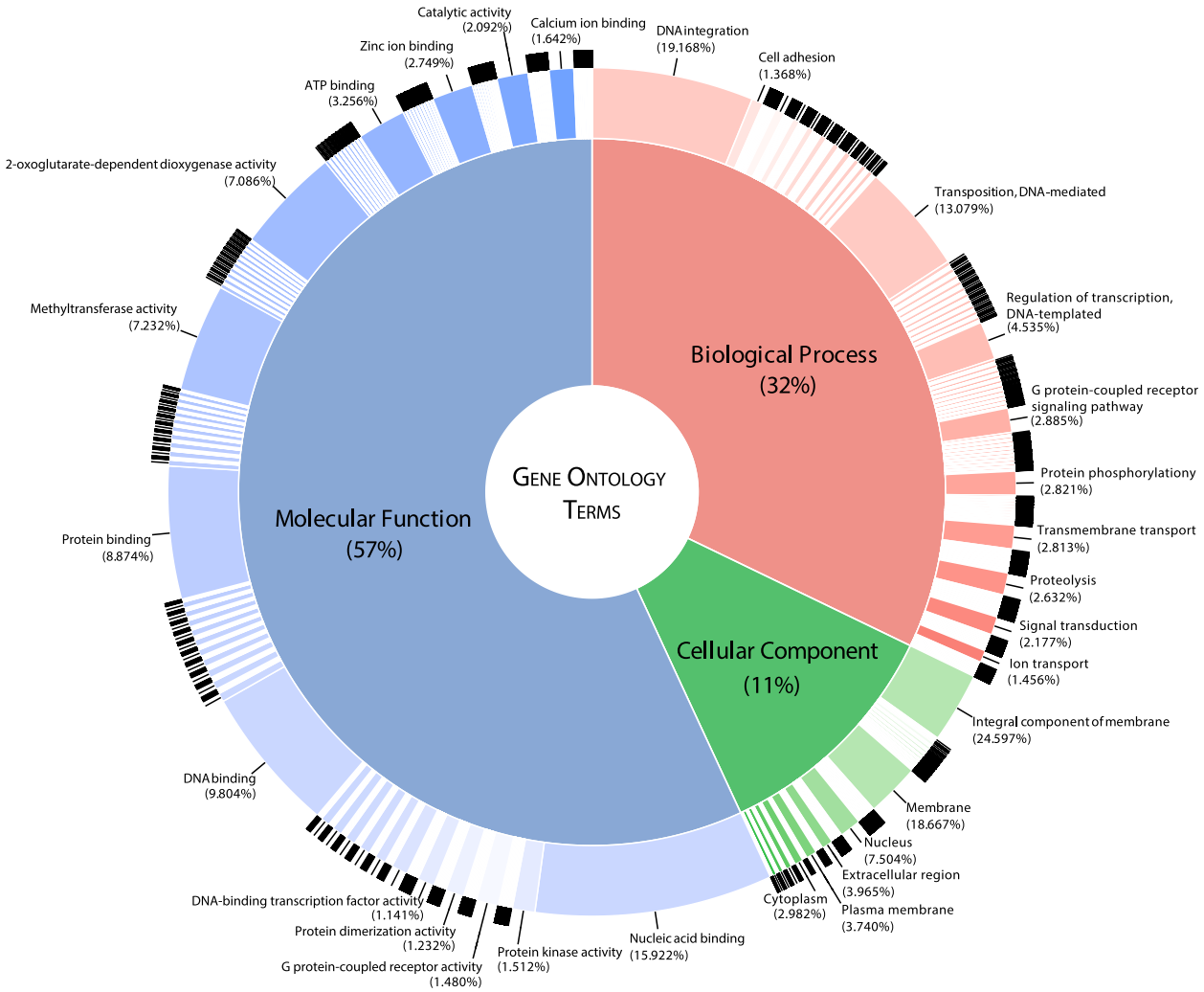

Fig. 5 Gene Ontology Terms Treemap. Top GO Molecular Functions in blue, top Biological Processes in red and top Cellular Components in green. The higher % for the different GO terms in each category are represented in the Treemap. Nucleid acid binding and DNA binding are the most abundant GO terms in Molecular Function; DNA integration in Biological Processes and Integral component of membrane in the corresponding Cellular Component

degree of overlap in transcripts across all assemblies, particularly when a more restricted analysis involving four closely related species was performed. The results of the analysis demonstrated a high degree of congruence with phylogenetic relationships among the species, with a minimal percentage of unassigned genes throughout the analysis. Additionally, the number of orthogroups that were exclusive or shared among these four species was found to be similar, highlighting the conserved nature of certain gene groups (Fig. 7). The phylogenetic tree constructed with IQ-tree revealed that *A. iberus*, along with the other members of the order Cyprinodontiformes and closely related genera, was also positioned in accordance with previous studies [106–108] (Fig. 6).

Our study provides further insights by additionally providing the complete genome annotation for this Reference Genome. This information can be used in future studies to identify potential candidate genes involved in the processes of adaptation and resilience to several external abiotic stressors, such as salinity, temperature and hypoxia. In particular, and given the euryhaline condition of *A. iberus*, our annotation has identified over 2,300 genes with a function related to salinity out of the 19,960 genes annotated with Sma3s. particularly hyperosmotic and hypotonic salinity responses, osmosensory signaling, ion transport, ion transmembrane transport, bicellular tight junctions, and hormone systems such as, among others,

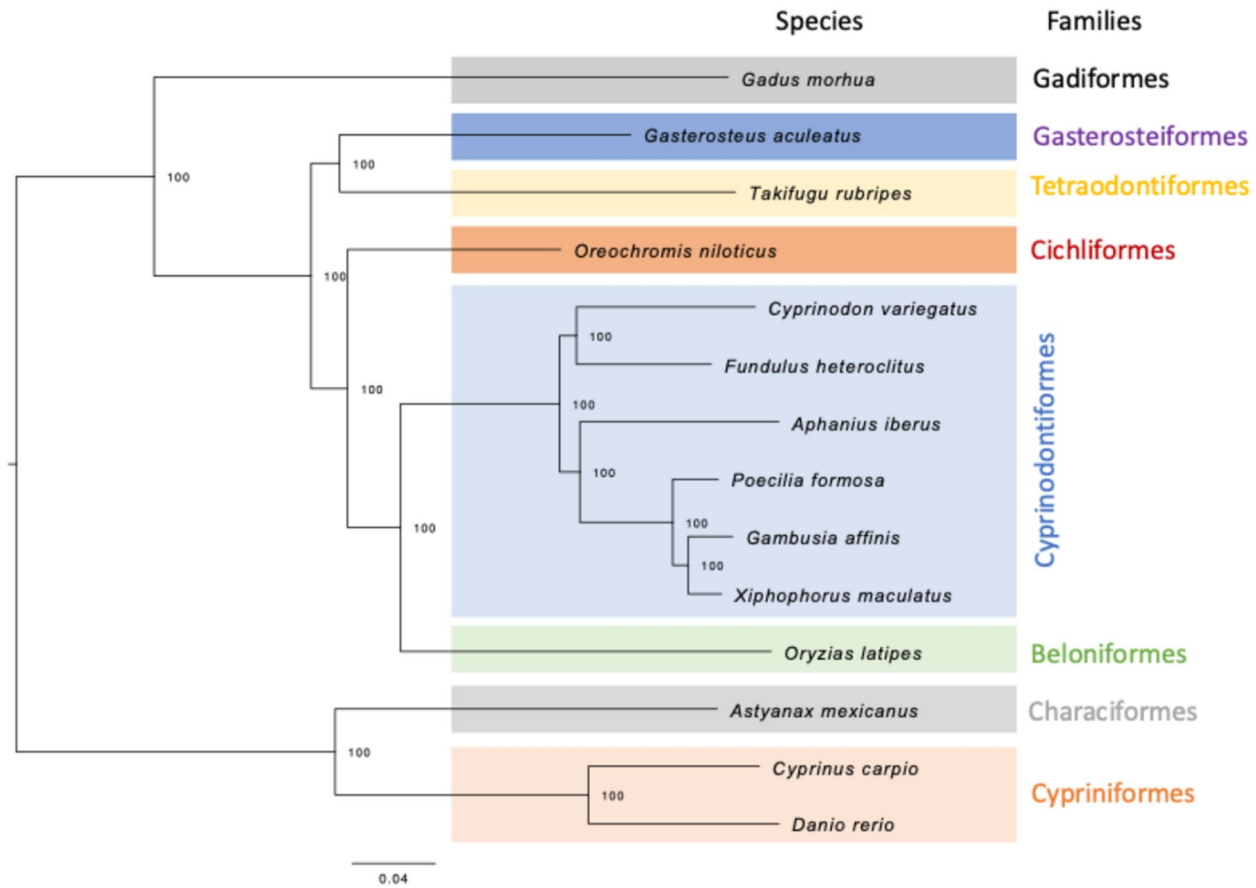

Fig. 6 Phylogenetic tree inferred by Orthofinder and IQ-tree. The numbers at nodes represent bootstrap values

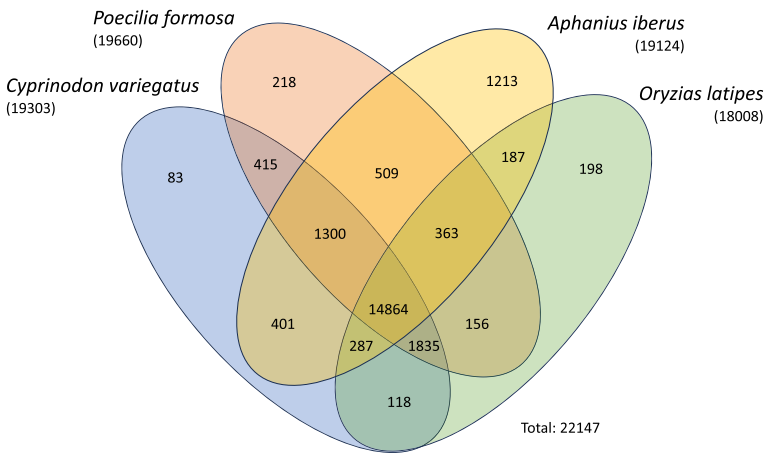

Fig. 7 Orthogroups of genes exclusive to and shared among the genome transcripts of the four analyzed fish species: *Aphanius iberus*, *Poecilia formosa*, *Cyprinodon variegatus* and *Oryzias latipes*

the renin–angiotensin–aldosterone system. These findings align with similar results in other research, implying that osmoregulation genes are reasonably stable across studies [3]. Nevertheless, more comprehensive comparisons are necessary due to variations in gene annotation methods across datasets. Therefore, the

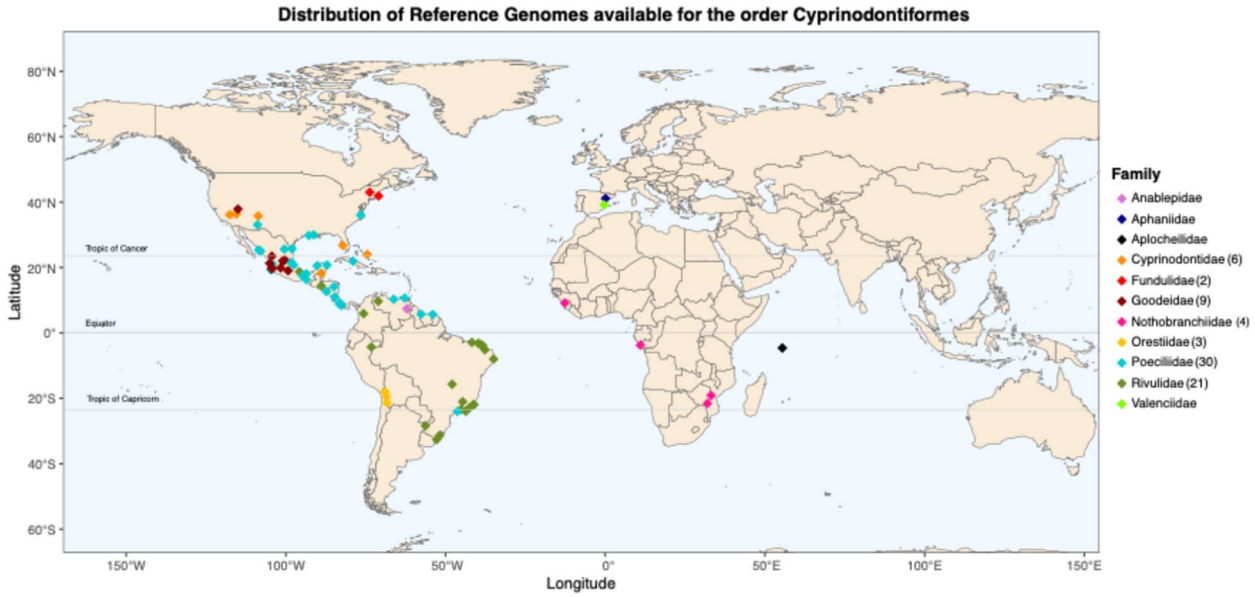

Fig. 8 Geographic distribution of Reference Genomes data available in GenBank, color and shape-coded by family. Numbers in brackets on the legend indicate the number of species belonging to each family when they are higher than 1 species per family

potential for conducting genetic studies on this species, such as its tolerance to varying salinity levels or gene expression responses, among others topics, remains highly intriguing. For instance, it is known that genes involved in osmoregulation play a crucial role for the survival of different Aphaniidae species [109]. These genes mediate differential gene expression in response to varying environmental conditions, enabling the maintenance of internal homeostasis at different salinity levels [109, 110]. A variety of proteins with diverse functions are implicated in these processes, including ion transporters, water channels, barrier proteins, signaling enzymes, and structural components. Regarding claudins and occludins, which are major transmembrane tetraspan proteins of tight junctions, have been described to play an important role in regulating paracellular permeability and ion and molecule equilibrium in several organisms [111, 112]. These two protein families are involved in the gill permeability changes during the process of acclimatization to fluctuating salinity conditions in fishes [113, 114].

Major Intrinsic Proteins (MIPs), including aquaporins and aquaglyceroporins, are essential for osmoregulation. Aquaporins function as water channels vital for hydric and osmotic regulation across cellular membranes, helping maintain homeostasis under stress conditions such as drought and salinity changes [115–120]. Additionally, proteins like the sodium–potassium pump (Na^+/K^+ ATPase), NKCC2, and NBCe1 facilitate ion exchange processes, while various proteins with specific functions, such as cathepsins, immunoglobulins, actins, connexins, and GTPases, are expressed in different tissues [121–123].

Our gene annotation has identified several hormone systems crucial for salinity adaptation, including the renin-angiotensin system, which regulates blood pressure and osmoregulation in teleost fishes, favoring cardiovascular homeostasis and renal sodium and water reabsorption [124–128]. Euryhaline species particularly benefit from these osmoregulatory behaviors. Another important hormone linked to osmotic regulation is arginine vasotocin, which is associated with aquaporin function [129–132].

Conclusions

The hybrid assembly presented in this study represents a significant step forward in our understanding of the biology of *A. iberus*, providing a well-sequenced and annotated Reference Genome that enhances our knowledge of the globally distributed order Cyprinodontiformes, which is currently predominantly limited to species from America and Africa. The application of more recent sequencing technologies, such as Omni-C, Chicago, or Hi-C could further enhance the assembly, achieving chromosomal-level resolution and addressing additional questions in the future.

These findings will contribute to the expanding database of Reference Genomes and will provide valuable information that can facilitate future studies not only in

the species but in the whole order. Moreover, the integration of genomic data with predicted genes offers a wide range of research opportunities across various disciplines, including physiology, reproduction, disease, and comparative genomic studies.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-025-11327-0.

Additional file 1.

Acknowledgements
We would like to express our gratitude to the Spanish Ministry of Science and Innovation and the State Agency of Investigation for their financial support, as well as to Pilar Risueño for her assistance in providing fish for the study. Furthermore, we express our gratitude to the Universidad Complutense de Madrid for their invaluable assistance in the development of the doctoral program for Alfonso López-Solano and Tessa Lynn Nester. In conclusion, we would like to express our gratitude to CESGA (Centro de Supercomputación de Galicia) for its indispensable contribution to the data analysis and the results of this study.

Authors' contributions
ALS contributed to the conceptualization of the project and played a pivotal role in the data curation and bioinformatics analysis. Additionally, the original draft was prepared by ALS, and participated on the reviews and edits. ID led the conceptualization, participated in the reviewing and editing and made a significant contribution to overall supervision. ID also acquired funding. TN contributed to the conceptualization, editing, and reviewing of the methodologies and provided input on the English vocabulary. SP also lead the conceptualization and supervision and participated in the data curation of the bioinformatic analysis.

Funding
Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature. This research study was funded by the Spanish Ministry of Science and Innovation and the State Agency of Investigation (MCIN/AEI/https://doi.org/10.13039/501100011033) as a part of Project "Conservation biology of endangered endemic cyprinodontiform fishes (APHANIUS)" (PID2019-103936GB-C21 and PID2019-103936GB-C22).

Data availability
The Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank under the accession number JAPXFQ01. The Submitted GenBank assembly described in this paper is version GCA_028564705.1. BioProject: PRJNA913687. BioSample: SAMN32303939. Annotation of genes are available at http://hdl.handle.net/10261/365271.

Declarations

Ethics approval and consent to participate
The fish were provided euthanized by Pilar Risueño from the Centro de Conservación de Especies Dulceacuícolas ("Piscifactoría de El Palmar"), with the requisite permission granted by the Government of Valencia (Spain). The study protocol methods used were approved by the Animal Experimentation Ethics Committee of the National Museum of Natural Sciences (CEEA-MNCN; Spanish Animal Experimentation Research Centre num. ES280790000189) in strict accordance with current Spanish law (RD53/2013) transposed from European Union regulation (art 2, 5f, in 2010/63/UE).

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]Museo Nacional de Ciencias Naturales, C/ José Gutiérrez Abascal, 2, 28006 Madrid, Spain. [2]Tragsatec. Grupo Tragsa, C/ Julián Camarillo 6B, Madrid 28037, Spain.

Received: 7 August 2024 Accepted: 5 February 2025
Published online: 12 February 2025

References

1. Ao J, Mu Y, Xiang LX, Fan D, Feng M, Zhang S, et al. Genome sequencing of the perciform fish *Larimichthys crocea* provides insights into molecular and genetic mechanisms of stress adaptation. PLoS Genet. 2015;11(4):e1005118. https://doi.org/10.1371/journal.pgen.1005118.
2. Figueras A, Robledo D, Corvelo A, Hermida M, Pereiro P, Rubiolo JA, et al. Whole genome sequencing of turbot (*Scophthalmus maximus*; Pleuronectiformes): a fish adapted to demersal life. DNA Res. 2016;23(3):181–92.
3. Takehana Y, Zahm M, Cabau C, Klopp C, Roques C, Bouchez O, et al. Genome sequence of the euryhaline *Javafish medaka*, Oryzias javanicus: A small aquarium fish model for studies on adaptation to salinity. G3 (Bethesda). 2020;10(3):907–15.
4. Crain CM, Halpern BS, Beck MW, Kappel CV. Understanding and managing human threats to the coastal marine environment. Ann N Y Acad Sci. 2009Apr;1162:39–62. https://doi.org/10.1111/j.1749-6632.2009.04496.x.
5. Evans DH, Piermarini PM, Choe KP. The multifunctional fish gill: dominant site of gas exchange, osmoregulation, acid-base regulation, and excretion of nitrogenous waste. Physiol Rev. 2005;85(1):97–177.
6. Evans DH. A brief history of the study of fish osmoregulation: the central role of the Mt. Desert Island biological laboratory Front Physiol. 2010;1:13. https://doi.org/10.3389/fphys.2010.00013.
7. Grosell M. The role of the gastrointestinal tract in salt and water balance. Fish Physiology. Academic Press; 2010. p. 135–164. https://doi.org/10.1016/S1546-5098(10)03004-9.
8. Yamanoue Y, Miya M, Doi H, Mabuchi K, Sakai H, Nishida M. Multiple invasions into freshwater by pufferfishes (Teleostei: Tetraodontidae): A mitogenomic perspective. PLoS One. 2011;6(2):e17410.
9. Gonzalez RJ. The physiology of hyper-salinity tolerance in teleost fish: a review. J Comp Physiol B. 2012;182:321–9.
10. Xia JH, Li HL, Zhang Y, Meng ZN, Lin HR. Identifying selectively important amino acid positions associated with alternative habitat environments in fish mitochondrial genomes. Mitochondrial DNA A DNA Mapp Seq Anal. 2017;28(5–6):809–19.
11. Jiang DL, Gu XH, Li BJ, Zhu ZX, Qin H, Meng ZN, et al. Identifying a long QTL cluster across chrLG18 associated with salt tolerance in tilapia using GWAS and QTL-seq. Mar Biotechnol. 2019;21(2):250–61.
12. Takvam M, Wood CM, Kryvi H, Nilsen TO. Ion transporters and osmoregulation in the kidney of teleost fishes as a function of salinity. Front Physiol. 2021;12: 664588.
13. Agarwal D, Shanmugam SA, Kathirvelpandian A, Eswaran S, Rather MA, Rakkannan G. Unraveling the impact of climate change on fish physiology: a focus on temperature and salinity dynamics. J Appl Ichthyol. 2024;2024(1):5782274.
14. Frankham R, Briscoe DA, Ballou JD. Introduction to conservation genetics. Cambridge: Cambridge University Press; 2002.
15. Reed DH, Frankham R. Correlation between fitness and genetic diversity. Conserv Biol. 2003;17(1):230–7.
16. Allendorf FW, Luikart GH, Aitken SN. Conservation and the genetics of populations. John Wiley & Sons; 2012.
17. Tickner D, Opperman JJ, Abell R, Acreman M, Arthington AH, Bunn SE, et al. Bending the curve of global freshwater biodiversity loss: An emergency recovery plan. Bioscience. 2020;70(4):330–42.
18. Doadrio I. Atlas y Libro rojo de los Peces continentales de España. Madrid: MNCN-CSIC, Dirección General de Conservación de la Naturaleza; 2001. 364 p.

19. Doadrio I, Perdices A, Machordom A. Allozymic variation of the endangered killifish *Aphanius iberus* and its application to conservation. Environ Biol Fishes. 1996;45:259–71.
20. Oliva-Paterna FJ, Mar T, Fernández-Delgado C. Threatened fishes of the world: Aphanius iberus (Cuvier & Valenciennes, 1846)(Cyprinodontidae). Environ Biol Fish. 2006;75(3):307–9.
21. Caiola N, Ibañez C. Restoration of Coastal Ecological Processes versus Fish Conservation: To Be or Not to Be… Biol Life Sci Forum. 2022;13(1):51; https://doi.org/10.3390/blsf2022013051.
22. Masó G, García-Berthou E, Merciai R, Latorre D, Vila-Gispert A. Effects of captive-breeding conditions on metabolic and performance traits in an endangered, endemic cyprinidontiform fish. Curr Zoo. 2024:zoae018.
23. García-Marín JL, Vila A, Pla C. Genetic variation in the Iberian toothcarp, *Aphanius iberus* (Cuvier & Valenciennes). J Fish Biol. 1990;37:233–4.
24. Perdices A, Carmona JA, Fernández-Delgado C, Doadrio I. Nuclear and mitochondrial data reveal high genetic divergence among Atlantic and Mediterranean populations of the Iberian killifish *Aphanius iberus* (Teleostei: Cyprinodontidae). Heredity. 2001;87(3):314–24.
25. Schönhuth S, Luikart G, Doadrio I. Effects of a founder event and supplementary introductions on genetic variation in a captive breeding population of the endangered Spanish killifish. J Fish Biol. 2003;63(6):1538–51.
26. Araguas RM, Roldán MI, García-Marín JL, Pla C. Management of gene diversity in the endemic killifish *Aphanius iberus*: revising Operational Conservation Units. Ecol Freshw Fish. 2007;16(2):257–66.
27. Nester TL, López-Solano A, Perea S, Doadrio I. Genomic population structure and diversity of the Endangered *Aphanius iberus*: strategies for killifish conservation. Conserv Genet. 2024. https://doi.org/10.1007/s10592-024-01665-z.
28. González EG, Pedraza-Lara C, Doadrio I. Genetic diversity and population history of the endangered killifish *Aphanius baeticus*. J Hered. 2014;105(5):597–610.
29. González EG, Cunha C, Ghanavi HR, Oliva-Paterna FJ, Torralva M, Doadrio I. Phylogeography and population genetic analyses in the Iberian toothcarp (*Aphanius iberus* Valenciennes, 1846) at different time scales. J Hered. 2018;109(3):253–63.
30. Pappalardo AM, González EG, Tigano C, Doadrio I, Ferrito V. Comparative pattern of genetic structure in two Mediterranean killifishes *Aphanius fasciatus* and *Aphanius iberus* inferred from both mitochondrial and nuclear data. J Fish Biol. 2015;87(1):69–87.
31. López-Solano A, Nester TL, Perea S, Doadrio I. Complete mitochondrial genome of the Spanish toothcarp, *Aphanius iberus* (Valenciennes. (Actinopterygii, Aphaniidae) and its phylogenetic position within the Cyprinodontiformes order. Mol Biol Rep. 1846;2023:1–10.
32. Gutiérrez V, Rego N, Naya H, García G. First complete mitochondrial genome of the South American annual fish *Austrolebias charrua* (Cyprinodontiformes: Rivulidae): peculiar features among cyprinodontiforms mitogenomes. BMC Genomics. 2015;16(1):1–15.
33. Yuan Z, Liu S, Zhou T, Tian C, Bao L, Dunham R, et al. Comparative genome analysis of 52 fish species suggests differential associations of repetitive elements with their living aquatic environments. BMC Genomics. 2018;19(1):1–10.
34. Di Genova A, Nardocci G, Maldonado-Agurto R, Hodar C, Valdivieso C, Morales P, et al. Genome sequencing and transcriptomic analysis of the Andean killifish *Orestias ascotanensis* reveals adaptation to high-altitude aquatic life. Genomics. 2022;114(1):305–15.
35. Yusuf LH, Lemus YS, Thorpe P, Garcia CM, Ritchie MG. Genomic signatures associated with transitions to viviparity in Cyprinodontiformes. bioRxiv. 2022; https://doi.org/10.1101/2022.05.05.490692.
36. Pareek CS, Smoczynski R, Tretyn A. Sequencing technologies and genome sequencing. J Appl Genet. 2011;52(4):413–35.
37. Giani AM, Gallo GR, Gianfranceschi L, Formenti G. Long walk to genomics: History and current approaches to genome sequencing and assembly. Comput Struct Biotechnol J. 2020;18:9–19.
38. Stanke M, Steinkamp R, Waack S, Morgenstern B. AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic Acids Res. 2006;34(Suppl 2):W435–9.
39. Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, et al. InterProScan 5: Genome-scale protein function classification. Bioinformatics. 2014;30(9):1236–40.
40. Muñoz-Mérida A, Viguera E, Claros MG, Trelles O, Pérez-Pulido AJ. Sma3s: A three-step modular annotator for large sequence datasets. DNA Res. 2014;21(3):341–53.
41. Andrews S. FastQC: A quality control tool for high throughput sequence data. 2010.
42. Hufnagel DE, Hufford MB, Seetharam AS. SequelTools: A suite of tools for working with pacbio sequel raw sequence data. BMC Bioinformatics. 2020;21(1):1–11.
43. Zimin AV, Puiu D, Luo MC, Zhu T, Koren S, Marçais G, et al. Hybrid assembly of the large and highly repetitive genome of *Aegilops tauschii*, a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm. Genome Res. 2017;27(5):787–92.
44. Marçais G, Yorke JA, Zimin A. QuorUM: An error corrector for Illumina reads. PLoS One. 2015;10(6):e0130821.
45. Zimin AV, Marçais G, Puiu D, Roberts M, Salzberg SL, Yorke JA. The MaSuRCA genome assembler. Bioinformatics. 2013;29(21):2669–77.
46. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol. 2019;37(5):540–6.
47. Haghshenas E, Asghari H, Stoye J, Chauve C, Hach F. HASLR: Fast hybrid assembly of long reads. iScience. 2020;23(1):101389.
48. Zimin AV, Salzberg SL. The genome polishing tool POLCA makes fast and accurate corrections in genome assemblies. PLoS Comput Biol. 2020;16(6):e1007981.
49. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics. 2010;26(5):589–95. https://doi.org/10.1093/bioinformatics/btp698.
50. Garrison E, Marth G. FreeBayes source repository. 2012.
51. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: Quality assessment tool for genome assemblies. Bioinformatics. 2013;29:1072–5. https://doi.org/10.1093/bioinformatics/btt086.
52. Kokot M, Długosz M, Deorowicz S. KMC 3: Counting and manipulating k-mer statistics. Bioinformatics. 2017;33(17):2759–61.
53. Vurture GW, Sedlazeck FJ, Nattestad M, Underwood CJ, Fang H, Gurtowski J, et al. GenomeScope: Fast reference-free genome profiling from short reads. Bioinformatics. 2017;33(14):2202–4.
54. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv preprint. 2013. Available from: https://arxiv.org/abs/1303.3997.
55. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25(16):2078–9. https://doi.org/10.1093/bioinformatics/btp352.
56. Seppey M, Manni M, Zdobnov EM. BUSCO: Assessing genome assembly and annotation completeness. In: Kollmar M, editor. Gene Prediction: Methods in Molecular Biology, vol. 1962. Humana; 2019. p. 227–45.
57. Fadiji AE, Babalola OO. Metagenomics methods for the study of plant-associated microbial communities: A review. J Microbiol Methods. 2020;170:105860. https://doi.org/10.1016/j.mimet.2020.105860.
58. Li H. Minimap2: Pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34(18):3094–100. https://doi.org/10.1093/bioinformatics/bty191.
59. Cabanettes F, Klopp C. D-GENIES: dot plot large genomes in an interactive, efficient and simple way. PeerJ. 2018;6:e4958.
60. Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, et al. RepeatModeler2 for automated genomic discovery of transposable element families. Proc Natl Acad Sci U S A. 2020;117:9451–7. https://doi.org/10.1073/pnas.1921046117.
61. Smit A, Hubley R, Green P. RepeatMasker open-4.0. 2015. Available from: http://www.repeatmasker.org
62. Bao W, Kojima KK, Kohany O. Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob DNA. 2015;6:11.
63. Pertea G, Pertea M. GFF utilities: GffRead and gffCompare. F1000Res. 2020;9:704.
64. Bryant DM, Johnson K, DiTommaso T, Tickle T, Couger MB, Payzin-Dogru D, et al. TranDecoder v5.5.0. Zenodo. 2019; https://doi.org/10.5281/zenodo.3388956.
65. Lu S, Wang J, Chitsaz F, Derbyshire MK, Geer RC, Gonzales NR, et al. CDD/SPARCLE: The conserved domain database in 2020. Nucleic Acids Res. 2020;48(D1):D265–8.
66. Lupas A, Van Dyke M, Stock J. Predicting coiled coils from protein sequences. Science. 1991;252(5009):1162–4.

67. Lewis TE, Sillitoe I, Dawson N, Lam SD, Clarke T, Lee D, et al. Gene3D: Extensive prediction of globular domains in proteins. Nucleic Acids Res. 2018;46(D1):D435–9.
68. Pedruzzi I, Rivoire C, Auchincloss AH, Coudert E, Keller G, De Castro E, et al. HAMAP in 2015: Updates to the protein family classification and annotation system. Nucleic Acids Res. 2015;43(D1):D1064–70.
69. Necci M, Piovesan D, Dosztányi Z, Tosatto SC. MobiDB-lite: Fast and highly specific consensus prediction of intrinsic disorder in proteins. Bioinformatics. 2017;33(10):1402–4.
70. Mi H, Muruganujan A, Thomas PD. PANTHER in 2013: Modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. Nucleic Acids Res. 2012;41(D1):D377–86.
71. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, et al. Pfam: The protein families database. Nucleic Acids Res. 2014;42:D222–30.
72. Wu CH, Nikolskaya A, Huang H, Yeh LS, Natale DA, Vinayaka CR, et al. PIRSF: Family classification system at the protein information resource. Nucleic Acids Res. 2004;32(suppl_1):D112-4.
73. Attwood T, Beck M. PRINTS–a protein motif fingerprint database. Protein Eng Des Sel. 1994;7:841–8.
74. Sigrist C, Castro E, Cerutti L, Cuche BA, Hulo N, Bridge A, et al. New and continuing developments at PROSITE. Nucleic Acids Res. 2012;41(D1):D344-7.
75. Akiva E, Brown S, Almonacid DE, Barber AE 2nd, Custer AF, Hicks MA, et al. The structure–function linkage database. Nucleic Acids Res. 2014;42:D521–30.
76. Letunic I, Bork P. 20 years of the SMART protein domain annotation resource. Nucleic Acids Res. 2018;46(D1):D493–6.
77. Gough J, Karplus K, Hughey R, Chothia C. Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. J Mol Biol. 2001;313:903–19.
78. Haft DH, Selengut JD, Richter RA, Harkins D, Basu MK, Beck E. TIGRFAMs and genome properties in 2013. Nucleic Acids Res. 2012;41(D1):D387–95.
79. Consortium U. The universal protein resource (uniprot). Nucleic Acids Res. 2007;36:D190–5.
80. Barbazuk WB, Korf I, Kadavi C, Heyen J, Tate S, Wun E, et al. The syntenic relationship of the zebrafish and human genomes. Genome Res. 2000;10(9):1351–8.
81. Iwamatsu T. Stages of normal development in the medaka *Oryzias latipes*. Mech Dev. 2004;121(7–8):605–18.
82. Jaillon O, Aury JM, Brunet F, Petit JL, Stange-Thomann N, Mauceli E, et al. Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. Nature. 2004;431(7011):946–57.
83. Kasahara M, Naruse K, Sasaki S, Nakatani Y, Qu W, Ahsan B, et al. The medaka draft genome and insights into vertebrate genome evolution. Nature. 2007;447(7145):714–9.
84. Löhr H, Hammerschmidt M. Zebrafish in endocrine systems: Recent advances and implications for human disease. Annu Rev Physiol. 2011;73:183–211.
85. Kinkel MD, Prince VE. On the diabetic menu: Zebrafish as a model for pancreas development and function. BioEssays. 2009;31(2):139–52.
86. Jones FC, Grabherr MG, Chan YF, Russell P, Mauceli E, Johnson J, et al. The genomic basis of adaptive evolution in threespine sticklebacks. Nature. 2012;484(7392):55–61.
87. Howe K, Clark MD, Torroja CF, Torrance J, Berthelot C, Muffato M, et al. The zebrafish reference genome sequence and its relationship to the human genome. Nature. 2013;496(7446):498–503.
88. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Mol Biol Evol. 2013;30(4):772–80.
89. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009;25(15):1972–3. https://doi.org/10.1093/bioinformatics/btp348.
90. Emms DM, Kelly S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol. 2015;16(1):1–14.
91. Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 2015;32(1):268–74.
92. Le SQ, Gascuel O. An improved general amino acid replacement matrix. Mol Biol Evol. 2008;25(7):1307–20.
93. Kalyaanamoorthy S, Minh B, Wong T, et al. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat Methods. 2017;14:587–9. https://doi.org/10.1038/nmeth.4285.
94. Minh BQ, Nguyen MA, von Haeseler A. Ultrafast approximation for phylogenetic bootstrap. Mol Biol Evol. 2013;30(5):1188–95.
95. Rodriguez-Santalla I, Navarro N. Main threats in Mediterranean coastal wetlands. The Ebro Delta case. J Mar Sci Eng. 2021;9(11):1190.
96. Lorrain-Soligon L, Robin F, Bertin X, Jankovic M, Rousseau P, Lelong P, et al. Long-term trends of salinity in coastal wetlands: Effects of climate, extreme weather events, and sea water level. Environ Res. 2023;237:116937. https://doi.org/10.1016/j.envres.2023.11693.
97. Parenti LR. A phylogenetic and biogeographic analysis of the cyprinodontiform fishes (Teleostei, Atherinomorpha). Bull Am Mus Nat Hist. 1981;168(1):335–557.
98. Parker A, Kornfield I. Molecular perspective on evolution and zoogeography of cyprinodontid killifishes (Teleostei; Atherinomorpha). Copeia. 1995;1995(1):8–21.
99. Hrbek T, Meyer A. Closing of the Tethys Sea and the phylogeny of Eurasian killifishes (Cyprinodontiformes: Cyprinodontidae). J Evol Biol. 2003;16(1):17–36.
100. Helmstetter AJ, Papadopulos AS, Igea J, Van Dooren TJ, Leroi AM, Savolainen V. Viviparity stimulates diversification in an order of fish. Nat Commun. 2016;7(1):11271.
101. Fricke R, Eschmeyer WN, van der Laan R, editors. ESCHMEYER'S CATALOG OF FISHES: GENERA, SPECIES, REFERENCES. 2024. Available from: http://researcharchive.calacademy.org/research/ichthyology/catalog/fishcatmain.asp. Accessed June 30, 2024.
102. Rhie A, McCarthy SA, Fedrigo O, Damas J, Formenti G, Koren S, et al. Towards complete and error-free genome assemblies of all vertebrate species. Nature. 2021;592(7856):737–46.
103. Liu D, Hunt M, Tsai IJ. Inferring synteny between genome assemblies: A systematic evaluation. BMC Bioinformatics. 2018;19(1):1–13. https://doi.org/10.1186/s12859-018-2026-4.
104. Charlesworth B, Sniegowski P, Stephan W. The evolutionary dynamics of repetitive DNA in eukaryotes. Nature. 1994;371:215–20.
105. Lee SI, Kim NS. Transposable elements and genome size variations in plants. Genomics Inform. 2014;12(3):87–97.
106. Pohl M, Milvertz FC, Meyer A, Vences M. Multigene phylogeny of cyprinodontiform fishes suggests continental radiations and a rogue taxon position of *Pantanodon*. Vertebr Zool. 2015;65(1):37–44.
107. Betancur-R R, Wiley EO, Arratia G, Acero A, Bailly N, et al. Phylogenetic classification of bony fishes. BMC Evol Biol. 2017;17(1):1–40. https://doi.org/10.1186/s12862-017-0958-3.
108. Esmaeili HR, Teimori A, Zarei F, Sayyadzadeh G. DNA barcoding and species delimitation of the Old World tooth-carps, family Aphaniidae Hoedeman, 1949 (Teleostei: Cyprinodontiformes). PLoS ONE. 2020;15(4):e0231717. https://doi.org/10.1371/journal.pone.0231717.
109. Zarei A, Motamendi M, Soltanian S, Teimori A. A contribution to the understanding of osmoregulation in two toothcarps occupying different osmotic niches. Iran J Sci Technol Trans Sci. 2021;45(1):127–34.
110. Bonzi LC, Monroe AA, Lehmann R, et al. The time course of molecular acclimation to seawater in a euryhaline fish. Sci Rep. 2021;11:18127. https://doi.org/10.1038/s41598-021-97295-3.
111. Chasiotis H, Kelly SP. Occludin and hydromineral balance in *Xenopus laevis*. J Exp Biol. 2009;212(2):287–96.
112. Krause G, Winkler L, Mueller SL, Haseloff RF, Piontek J, Blasig IE. Structure and function of claudins. Biochim Biophys Acta Biomembr. 2008;1778(3):631–45.
113. Zhou B, Qi D, Liu S, Qi H, Wang Y, Zhao K, et al. Physiological, morphological and transcriptomic responses of Tibetan naked carps (*Gymnocypris przewalskii*) to salinity variations. Comp Biochem Physiol Part D Genomics Proteomics. 2022;42:100982.
114. Chen X, Liu S, Ding Q, Teame T, Yang Y, Ran C, et al. Research advances in the structure, function, and regulation of the gill barrier in teleost fish. Water Biol Secur. 2023;2(2):100139; https://doi.org/10.1016/j.watbs.2023.100139.
115. Abascal F, Irisarri I, Zardoya R. Diversity and evolution of membrane intrinsic proteins. Biochim Biophys Acta. 2017;1840:1468–81.

116. Verma RK, Gupta AB, Sankararamakrishnan R. Major intrinsic protein superfamily: Channels with unique structural features and diverse selectivity filters. In: Methods in Enzymology. Vol. 557. Academic Press; 2015. p. 485–520.
117. Zardoya R. Phylogeny and evolution of the major intrinsic protein family. Biol Cell. 2005;97(6):397–414.
118. Kruse E, Uehlein N, Kaldenhoff R. The aquaporins. Genome Biol. 2006;7(2):206. https://doi.org/10.1186/gb-2006-7-2-206.
119. Ruhr IM, Wood CM, Schauer KL, Wang Y, Mager EM, Stanton B, et al. Is aquaporin-3 involved in water-permeability changes in the killifish during hypoxia and normoxic recovery, in freshwater or seawater? J Exp Zool A Ecol Integr Physiol. 2020;333(7):511–25.
120. Yepes-Molina L, Bárzana G, Carvajal M. Controversial regulation of gene expression and protein transduction of aquaporins under drought and salinity stress. Plants. 2020;9(12):1662.
121. Tort L. Stress and immune modulation in fish. Dev Comp Immunol. 2011;35(12):1366–75.
122. Dominguez R, Holmes KC. Actin structure and function. Annu Rev Biophys. 2011;40(1):169–86.
123. Wei CJ, Xu X, Lo CW. Connexins and cell signaling in development and disease. Annu Rev Cell Dev Biol. 2004;20(1):811–38.
124. Nishimura H. Renin-angiotensin system in vertebrates: Phylogenetic view of structure and function. Anat Sci Int. 2017;92(2):215–47.
125. McCormick SD, Regish A, O'Dea MF, Shrimpton M. Are we missing a mineralocorticoid in teleost fish? Effects of cortisol, deoxycorticosterone and aldosterone on osmoregulation, gill Na+, K+-ATPase activity and isoform mRNA levels in Atlantic salmon. Gen Comp Endocrinol. 2008;157(1):35–40. https://doi.org/10.1016/j.ygcen.2008.03.024.
126. Bader M, Ganten D. Update on tissue renin–angiotensin systems. J Mol Med (Berl). 2008;86:615–21. https://doi.org/10.1007/s00109-008-0336-0.
127. Sakamoto T. Hormonal regulation of body fluid in teleost fishes. Bull Soc Sea Water Sci Jpn. 2015;69(4):244–6.
128. Brown JA, Hazon N. The Renin-Angiotensin Systems of Fish and their Roles in Osmoregulation. In: Baldisserotto B, editor. Fish Osmoregulation. CRC Press; Boca Raton. 2019; https://doi.org/10.1201/9780429063909.
129. McCormick SD, Farrell AP, Brauner CJ, editors. Fish physiology: Euryhaline fishes. San Diego: Academic Press; 2013.
130. Wenxiao C, Aijun MA, Zhihui H, Xin'an W, Zhibin S, Zhifeng L, et al. Transcriptomic analysis reveals putative osmoregulation mechanisms in the kidney of euryhaline turbot *Scophthalmus maximus* responded to hypo-saline seawater. J Oceanol Limnol. 2020;38(2):467–79.
131. An KW, Kim NN, Choi CY. Cloning and expression of aquaporin 1 and arginine vasotocin receptor mRNA from the black porgy, *Acanthopagrus schlegeli*: effect of freshwater acclimation. Fish Physiol Biochem. 2008;34:185–94. https://doi.org/10.1007/s10695-007-9175-0.
132. Breves JP. Hormonal regulation of aquaporins in fishes. In: Litwack G, editor. Aquaporin regulation. Vitamins and Hormones. 2020;112:265–287; https://doi.org/10.1016/bs.vh.2019.10.002.

Publisher's Note