scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Chromosome-level genome assembly and annotation of the sharpsnout seabream (*Diplodus puntazzo*)

Cinta Pegueroles[1,2,3,9] ✉, Carles Galià-Camps[1,2,4,9] ✉, Marta Pascual[1,2], Marta Bassitta[5], Didac González[1], Carola Greve[6,7], Enrique Macpherson[4], Núria Raventós[4], Tilman Schell[6,7], Héctor Torrado[8] & Carlos Carreras[1,2] ✉

Diplodus puntazzo is a demersal fish inhabiting the Mediterranean Sea and the eastern Atlantic and plays an important ecological role in coastal areas. Here, we present the first nuclear genome assembly and annotation of this species and genus. We used a combination of PacBio CLR long reads, Illumina short reads and chromatin capture reads (Omni-C) to generate a chromosome-level assembly. The nuclear genome assembly has a total span of 788 Mb, containing 24 chromosome-scale scaffolds (98.76% of the total length), coinciding with its known karyotype. By using RNA-Seq data from _D. puntazzo_ and gene models from closely related species, we also generated a high-quality nuclear annotation. We predicted a total of 87,572 transcripts from the nuclear genome, 26,838 coding, and 60,734 non-coding that included lncRNA, snoRNA, and tRNAs. We also assembled and annotated the mitochondrial genome, circularized in 16,642 bp comprising 13 protein-coding genes, 2 rRNA, and 22 tRNA. This high-quality reference genome will enrich the current genomic resources available to the large fish scientific community.

Background & Summary

Diplodus puntazzo (Walbaum, 1792) (Osteichthyes: Sparidae), commonly called the sharpsnout seabream, is a protandrous hermaphrodite fish from the Mediterranean Sea and the eastern Atlantic[1]. It inhabits rocky shore reefs and seagrass meadows down to a depth of 150 metres[2–4]. It has an important ecological role since its wide prey spectrum includes toxic species such as sponges, echinoderms, and coelenterates[5,6]. *D. puntazzo* has been cultured since several decades ago[7], being an important and valuable demersal species in coastal fisheries[8,9]. Although it is a popular food fish and a commercial species, it is also regarded as one of the flagship species for conservation and, in particular, the establishment of Marine Protected Areas (MPAs) in the Mediterranean[10–13].

The sharpsnout seabream grows large in size (up to 60 cm), is moderate long-lived (up to 10 years), but slow-growing with a late onset of maturity (at the age of three). This species has a complex reproductive biology as a protandrous hermaphrodite[14–16] and its reproduction occurs from August to September[17]. Based on its life history characteristics, its pelagic larvae are the main contributor to gene flow, similar to other species of *Diplodus*[18]. The pelagic larvae are rarely captured in plankton nets but remain for 16–29 days in the planktonic stage[19]. Unfortunately, data on its dispersal potential is scarce, similar to the current information status of many other commercially overfished marine species[20]. Settlement occurs at very shallow depths (0–2 m) on sandy-rocky bottoms sheltered from winds[21,22]. The first settlers of *D. puntazzo* usually arrive at the beginning of

[1]Department of Genetics, Microbiology and Statistics, Faculty of Biology, University of Barcelona, Barcelona, Spain. [2]Institut de Recerca de la Biodiversitat (IRBio), Faculty of Biology, University of Barcelona, Barcelona, Spain. [3]Department of Genetics and Microbiology, Universitat Autònoma de Barcelona, Bellaterra, Barcelona, Spain. [4]Blanes Centre for Advanced Studies, Spanish National Research Council (CEAB-CSIC), Blanes, Spain. [5]Department of Biology, University of Balearic Islands, Palma de Mallorca, Spain. [6]Centre for Translational Biodiversity Genomics (LOEWE-TBG), Frankfurt am Main, Germany. [7]Senckenberg Research Institute, Frankfurt am Main, Germany. [8]Marine Laboratory, University of Guam, Guam, USA. [9]These authors contributed equally: Cinta Pegueroles, Carles Galià-Camps. ✉e-mail: cinta.pegueroles@uab.cat; cgaliacamps@gmail.com; carreras@ub.edu

October or November[23], suffering a density-dependent mortality during the first weeks after settlement[24]. New settlers of *D. puntazzo* are gregarious, form monospecific shoals, and are practically devoid of pigmentation except for the same faint, dark vertical bands that are characteristic of adults[25]. Due to its popularity as a the target species for commercial and recreational fishermen, the sharpsnout seabream population has declined dramatically the last decades with alarming reductions reported throughout its range in the Mediterranean Sea[2].

The karyotype of *D. puntazzo* is composed of 20 acrocentric and four meta/submetacentric chromosome pairs[26]. Previous genetic studies in *D. puntazzo* reported significant genetic differentiation between the Atlantic and the Mediterranean as well as within the Mediterranean[27]. Furthermore, a recent study using 85,031 SNPs found common signals across localities in the western Mediterranean for potential selection resulting in higher survival for individuals hatching later, growing more slowly, and experiencing lower temperatures during their planktonic phase[18]. These authors identified several loci with significant associations with phenotypic and environmental factors. Interestingly, one of the loci showed parallel frequency changes in non-synonymous mutations in the three populations studied and was identified as the *il20rb* gene, whose function is involved in the immune system. However, the lack of an annotated reference genome for *D. puntazzo* prevented the genomic nature of most of the candidate loci for juvenile survival, since the number of mapped fragments clearly reduces with the phylogenetic distance to the closest reference genome[28], and the enzyme used in reduced representation technique impacts on the genomic targeted regions[29]. These results highlight the importance of obtaining the complete reference genome of *D. puntazzo*.

Here we provide the annotated genome of the demersal fish *D. puntazzo*, which has been sequenced within the framework of the Catalan Initiative for the Earth Biogenome Project (CBP)[30]. This is the first chromosome-level reference for the genus *Diplodus*. Due to its high quality in both assembly and annotation, it will facilitate future investigations into the evolution, biology, and conservation of *Diplodus puntazzo*, and will also contribute to enriching genomic resources for fish and biodiversity in general.

Methods

Sampling and sequencing of *Diplodus puntazzo*. A single juvenile individual of *D. puntazzo* was collected in Cala Sant Francesc, Blanes, Spain, in April 2021. It was immediately euthanized in liquid nitrogen and preserved at −80 °C. Once in the laboratory, ~25 mm^2 of muscle were excised for DNA extraction, and the same amount of eye, mouth, and muscle were prepared for RNA extractions. For DNA extraction we followed a protocol based on that of Sambrook and Russell (2001)[31], whereas for RNA extractions we used TRIzol reagent (Invitrogen) according to the manufacturer's instructions. The quality and concentration of the extractions were assessed using the TapeStation 2200 (Agilent Technologies) and the Qubit Fluorometer with the appropriate Qubit dsDNA/RNA BR Reagents Assay Kit (Thermo Fisher Scientific, Waltham, MA). We first sequenced a cytochrome oxidase I (*COI*) fragment using fish-specific primers[32] to confirm the taxonomic identity of our specimen using molecular tools. Subsequently, we constructed a SMRTbell library following the instructions of the SMRTbell Express Prep kit v2.0 with Low DNA Input Protocol (Pacific Biosciences, Menlo Park, CA). We performed one SMRT cell sequencing run in Continuous Long Read (CLR) mode on a Sequel System IIe with the Sequel II Sequencing Kit 2.0. Additionally, we sent a DNA extract of the same specimen to Novogene (UK) for Illumina Short Reads (SR) Whole Genome Sequencing (WGS). They prepared a genomic library (insert size: 350 bp) and sequenced 150 bp paired-end reads on an Illumina NovaSeq 6000 platform (San Diego, CA), aiming for 50 Gigabases (Gb) output. The proximity ligation library was constructed using the Dovetail® Omni-C® Kit (Dovetails Genomics). 230 mg of muscle tissue from the same individual was processed according to the Omni-C Proximity Ligation Assay protocol version 1.2. The library was sequenced on the NovaSeq 6000 platform at Novogene (Cambridge, UK) using a 150 paired-end sequencing strategy with an insert size of 350 bp, aiming for 60 Gb. Finally, RNA extractions from the eye, mouth, and muscle were sent to Novogene (UK) for Illumina paired-end 150 bp RNAseq of a cDNA library (insert size: 350 bp) with an expected output of 10 Gb per tissue. The sequencing yielded 146.9 Gb for PacBio CLR, 56.3 Gb for Illumina WGS and 92.5 Gb for Omni-C. The quality of the sequencing output was checked, and data was filtered (see below).

Nuclear genome assembly. To generate a reference genome at the chromosome level we used a combination of long and short genomic reads, as well as contact reads (Fig. 1). We transformed PacBio CLR subreads raw sequence files from bam to fastq format using Bedtools v.2.31.0 and checked their quality using Nanoplot v.1.28.1. For short reads, including WGS and Omni-C data, we trimmed and filtered the raw data files using Trimmomatic v.0.39, and checked their quality with fastQC v.0.11.9. We estimated the genome size, heterozygosity, and repeated content of *D. puntazzo* using polished Illumina WGS reads with GenomeScope v.2.0[33], using as input a histogram of k(21)-mer frequencies previously obtained with jellyfish v.2.2.3[34]. We used Flye v.2.8 to assemble the filtered Pacbio CLR predefining the genome size with the value obtained with GenomeScope 2 and with three iterations of self-polishing, followed by three polishing rounds with the WGS data using Pilon v.1.23. Finally, the assembly draft was deduplicated using purge_dups.py v.1.2.6 (Fig. 1). For genome scaffolding we used Omni-C data and uploaded the polished assembly to the program Juicer v.1.6, defining 'none' as the restriction enzyme, since Omni-C libraries use an endonuclease with random cleavage sites. The function run-asm-pipeline.sh of the program 3D-DNA v.201008 was called to obtain the contact map, which was manually curated using the interface of Juicebox v.1.5. The final chromosome-level assembly was recovered by running the script juicebox_assembly_converter.py of the same program (Fig. 1).

Our first assembly had a size of ~799 Mb, in line with the GenomeScope profile (genome size = 785.5 Mb; Heterozygosity = 0.55%; Repeated content = 19.2%, Supplementary Figure 1), and the genome size of the sibling species *Diplodus sargus* (788 Mb, Supplementary Table 1), although it was highly fragmented, with 1,990 contigs. After polishing, removing repeats, and scaffolding the contigs with Omni-C data, we obtained a final genome assembly of ~788 Mb, 1,275 contigs, and an N50 and L50 of 34,7 Mb and 15, respectively

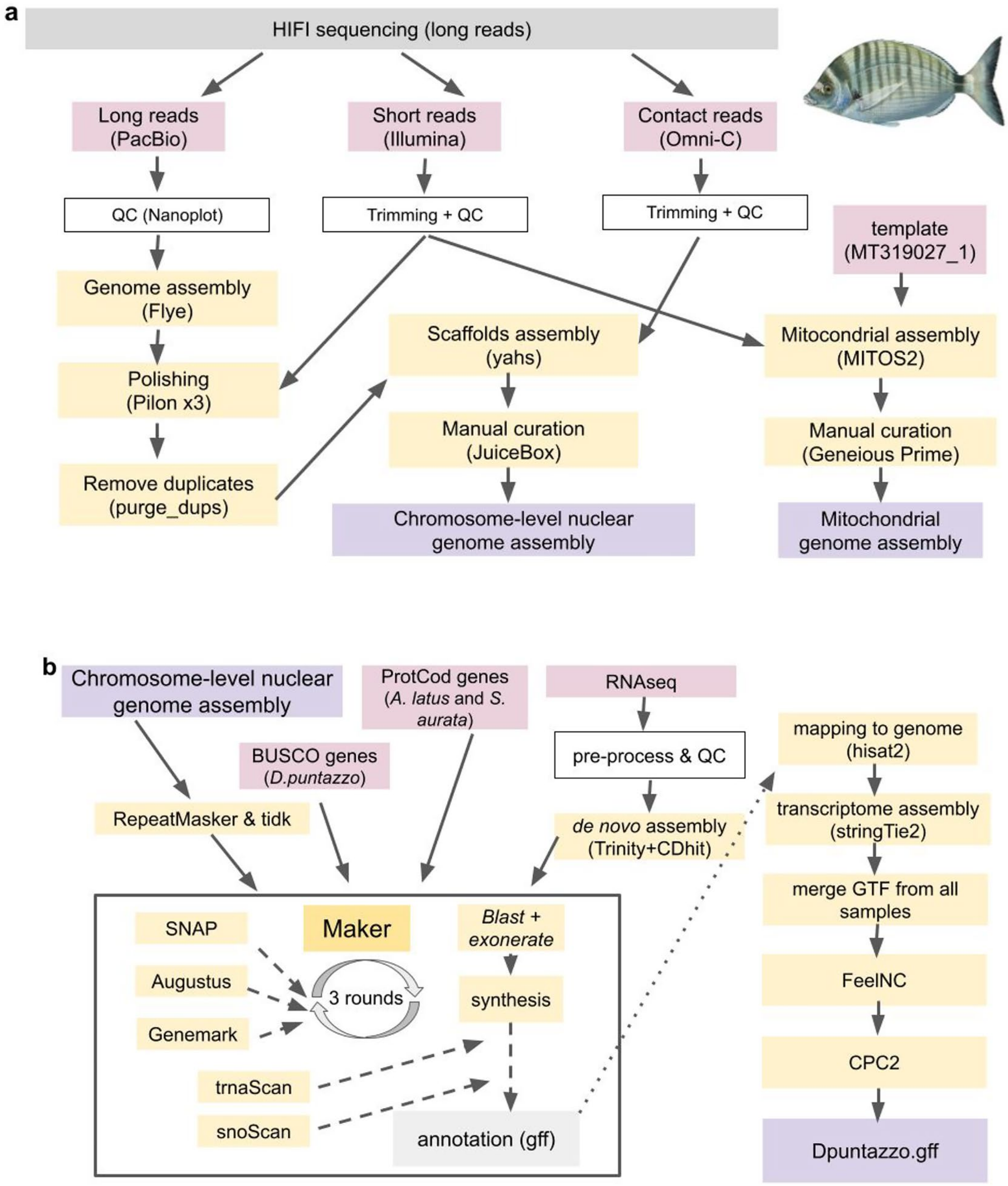
a
HIFI sequencing (long reads)
Long reads (PacBio)
Short reads (Illumina)
Contact reads (Omni-C)
QC (Nanoplot)
Trimming + QC
Trimming + QC
template (MT319027_1)
Genome assembly (Flye)
Polishing (Pilon x3)
Remove duplicates (purge_dups)
Scaffolds assembly (yahs)
Manual curation (JuiceBox)
Chromosome-level nuclear genome assembly
Mitocondrial assembly (MITOS2)
Manual curation (Geneious Prime)
Mitochondrial genome assembly
b
Chromosome-level nuclear genome assembly
ProtCod genes (A. latus and S. aurata)
RNAseq
BUSCO genes (D.puntazzo)
RepeatMasker & tidk
pre-process & QC
de novo assembly (Trinity+CDhit)
mapping to genome (hisat2)
transcriptome assembly (stringTie2)
merge GTF from all samples
FeelNC
CPC2
Dpuntazzo.gff
SNAP
Augustus
Genemark
Maker
3 rounds
Blast + exonerate
synthesis
trnaScan
snoScan
annotation (gff)

Fig. 1 Workflow for (**a**) genome assembly and (**b**) annotation. Pink designates the input data: sequencing data generated in the current research and available in ENA (PRJEB49350[49]), the mitochondrial genome use as template (MT319027_1) and available gene models for other species (protein coding and BUSCO genes), yellow designates the processes and/or the software used, and violet the output files generated.

(Supplementary Figure 2), and comprised only 7.99 uncertain positions (Ns) every 100 kb. Our results were similar to the chromosome-level assemblies of *Sparus aurata*[35,36] (GCA_900880675.2) and *Acanthopagrus latus*[37,38] (GCA_904848185.1) in terms of N50 and L50, and better than those of the congeneric species *D. sargus*[39,40] (GCA_903131615.1) which is at the scaffold level. The twenty-four macro scaffolds were identified as chromosomes and numbered according to their assembled length (Supplementary Figure 2, Supplementary Table 2). The remaining microscaffolds, accounting for 1.23% of the total assembly, provided significant BLAST hits against fish species (Supplementary Table 3), ruling out possible contamination. The mean GC content across the genome is 41.9%, with a tendency of higher GC content in the smallest chromosomes (Supplementary Figure 3). This tendency is also observed in the other two fish species with assemblies at the chromosome level (Supplementary Figure 4).

Repeated elements analysis. Obtaining the most complete genome annotation is necessary to unravel fundamental functional aspects of the biology of the species. Thus, we conducted the annotation of repetitive elements, telomeric regions, and genes (Fig. 1). To identify repeated regions in the reference genome assembly, we downloaded transposable elements (TE) for the closest species available in Repbase 86 (*Fugu*

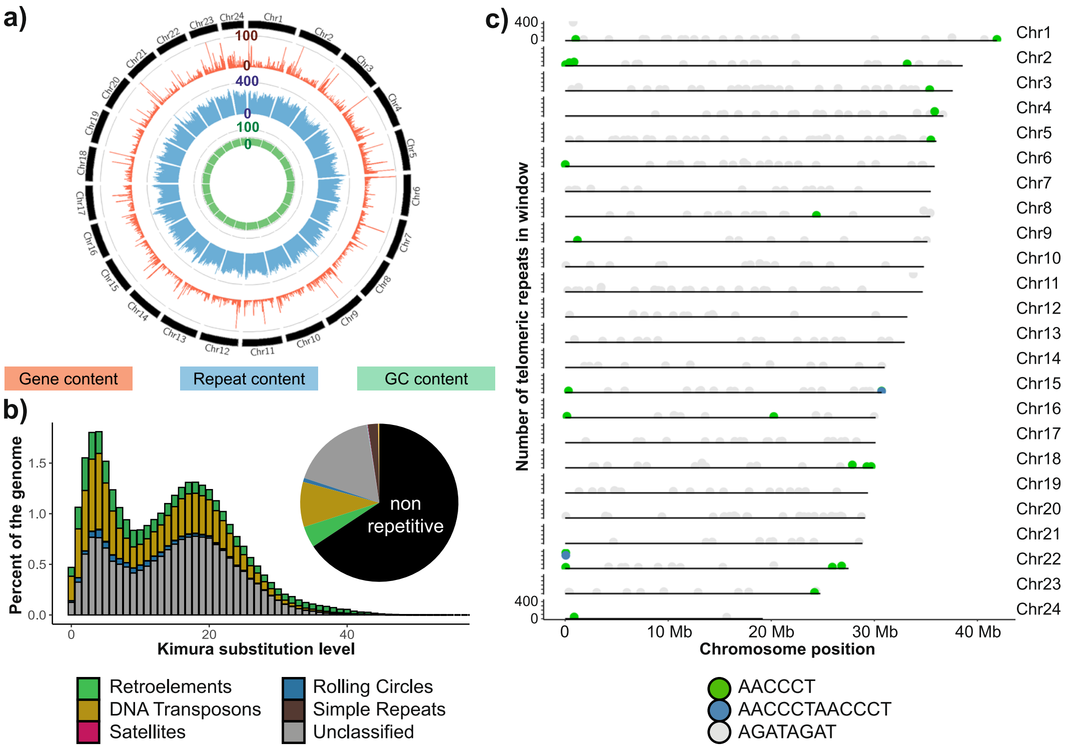

Fig. 2 Nuclear genome assembly of *D. puntazzo*. (**a**) Plot showing the relative length of the 24 chromosomes, and the GC% content, repetitive elements content, and gene content across each chromosome span, (**b**) genome coverage of repetitive elements clustered according to their Kimura substitution level and (**c**) plot showing the location of putative telomeric regions and the repeats involved.

rubripes and *Danio rerio*). In addition, to aid in the TE annotation of *D. puntazzo* we downloaded the reference genome of *Sparus aurata* and *Acanthopagrus latus*, which are closely related species (divergence time being 38 Myr and 25.7 Myr, respectively, according to https://timetree.org/) but lacking TEs annotation. We generated de novo predictive TE models for *D. puntazzo*, *S. aurata* and *A. latus* using RepeatModeler v.2.0. Subsequently, we combined the models obtained for the three species with those obtained from Repbase. We used this model file including the TEs of all five species to soft- and hard-mask the chromosome-level assembly using RepeatMasker v.4.1.2. To plot TE families abundance and Kimura2 substitution levels profile we used the script RepeatLandscape.pl, which is included in the RepeatMasker program. Regarding transposable elements (TE), we found that they cover 34.32% of the genome assembly (Fig. 2, Supplementary Table 4). The most abundant TE family were DNA transposons (9.23%), followed by retroelements (4.35%) and simple repeats (2.07%). Interestingly, most of the repeats (17.38%) corresponded to unclassified repetitive elements, since these regions could not be assigned to any specific TE family. The Kimura2 profile identified a bimodal distribution, in which the most recent event (the one with lower K2 mutations) was characterised by the expansion of DNA transposons and retroelements in the two peaks. This bimodal pattern is also typical in other teleost fish such as cichlids, also involving the same families[41].

Telomere and centromere analysis. We looked for repetitive motifs corresponding to telomeric or centromeric regions with tidk v.0.2.31[42], by screening the three most common telomeric repeats in fishes reported in the Telomeric Repeat Database (AGATAGAT, AACCCT, and AACCCTAACCCT, https://github.com/tolkit/a-telomeric-repeat-database) across the whole genome using 10,000 bp windows with the tidk function *screen* (–screen). We kept those windows with more than 25 motif counts and graphically visualised their distribution and abundance across the assembly with ggplot2.

The three most common telomeric repeats scanned in the final genome assembly, account for 86% of the telomeric repeats described in Actinopteri (Supplementary Table 5). In *D. puntazzo*, the motifs AACCCT and AACCCTAACCCT were found in large numbers in the distal regions of the chromosomes, and thus identified as potential telomeres (Fig. 2). On the contrary the motif AGATAGAT was not indicative of telomeric regions since it was distributed across all chromosomes. This telomeric motif is also found in amphibians, and therefore is not fish specific. We identified three chromosomes having telomeric repeats in both tips, whereas most chromosomes lack one or both chromosome ends. In chromosomes 8 and 16, we found a region rich in AACCCT motif situated in the middle of the chromosome, which could correspond to potential centromeres. However, we cannot dismiss the possibility that some regions identified as telomeres but found in locations other than just the ends of chromosomes (such as chromosome 2), could result from misassembled regions. Overall, these results are in line with those of previous karyotype research, which identified most chromosomes as acrocentric and the remaining as meta/submetacentric[43]. However, the number of meta/submetacentric chromosomes was four, suggesting that two of the centromeres might be not identified in the present study.

Nuclear genome annotation.

The annotation of genes in the sharpsnout seabream genome was performed using the MAKER software[44] (Fig. 1). To generate a complete annotation, we combined RNAseq from *D. puntazzo*, its own BUSCO gene sequences, and amino acid sequences from *S. aurata* and *A. latus* previously downloaded from UniProt. In addition to our sequenced RNAseq data (eye, mouth, and muscle), we downloaded the RNAseq data of *D. puntazzo* from NCBI BioProject database (PRJNA241484[45], corresponding to gonad and brain tissue from four adult females and four adult males[46]), yielding 33.9 Gigabytes (GB) and 43 GB of data for the collected individual and public data respectively, adding a total of 76.9 GB of raw data (Fig. 2, Supplementary Table 6). RNAseq data was filtered with Trimmomatic v.0.39 and assembled into transcripts with Trinity v.2.11. To reduce redundancy in the de novo generated transcripts we used CD-HIT v.4.8 with 99% similarity threshold. We used both downloaded and the *de novo* generated genes to obtain homology-based annotations on the hard-masked genome using blast and exonerate, as implemented in MAKER v.2.31.10. In addition, we conducted 'ab-initio' gene predictions with three different software: AUGUSTUS v.3.4.0, GeneMark-EP v.4.71 and SNAP v.2013-11-29. To refine the genome annotation, we performed a total of three rounds of protein modelling (Fig. 1). For the first modelling round, we used BUSCO genes to generate a gene model with SNAP and RNAseq for AUGUSTUS, implemented in MAKER. For the second and third modelling rounds we used as training set the genes obtained by the most updated annotation draft. Additionally, in the last annotation round, we also used tRNAscan-SE v.2.0 and snoscan v.0.9.1 to annotate tRNA and small non-coding RNA (snoRNA). Genes and transcripts annotated with MAKER were renamed using maker_map_ids and map_gff_ids scripts from MAKER 2.31.10 as DPUNG[0-9]{6} and DPUNT[0-9]{6}-T[0-9] respectively. For the long non-coding RNAs (lncRNAs) annotation, we mapped the filtered RNAseq reads to the reference genome using HISAT2 v2.2.1121 and we obtained individual BAM files for each sample. Later, we generated a transcriptome assembly for each BAM file using the MAKER gene models as a reference annotation, with StringTie v2.1.4122. Subsequently we merged the individual GTF files obtained for each sample into a single GTF file using the merge option from StringTie. Then we used the FeelNC software and the single GTF file to identify the candidate lncRNAs. First, we filtered out transcripts shorter than 200 bp, mono-exonic, and overlapping protein-coding genes. Afterwards, we discarded transcripts with coding potential. To compute the coding potential, we used the shuffle approach, which in brief takes a set of mRNAs and shuffles them while preserving 7-mer frequencies. We selected the transcripts identified as non-coding and used them as input to estimate the coding potential using CPC2 (http://cpc2.gao-lab.org/, accessed in June 2023). LncRNA genes were named as MSTRG.[0-9]{5} and transcripts as MSTRG.[0-9]{6}.[0-9]. We generated a final GTF file by merging MAKER and lncRNAs annotations.

Overall, we annotated a total of 20,040 protein-coding genes and 26,838 transcripts (1.34 transcript per gene on average, with a maximum of 18 transcripts per gene, ± 0.95 SD) (Supplementary Table 1). Protein-coding transcripts contain 10.85 exons on average, a mean length of 2,795.53 bp and 49.01% of GC content (Supplementary Table 7). In addition, we annotated 57,874 non-coding genes, including 5,867 lncRNA (8,727 transcripts), 49,780 snoRNA, and 2,227 tRNAs (Supplementary Table 1). For lncRNAs, we annotated 1.49 transcript per gene on average, with a maximum of 26 transcripts per gene (±1.47 SD). The number of exons per lncRNA transcript was 2.85 on average, their mean length of 1,269.40 ncl and 45.01% of GC content on average. In addition, we identified 36,635 and 33,518 three-prime and five-prime UTR regions respectively. The GC%, repeated element content, and transcript content were visually represented across all chromosomes with the software Circos[47]

Mitochondrial genome assembly and annotation.

The mitochondrial genome was obtained using Novoplasty 2.2. To do so, we provided a complete mitochondrial genome (obtained from NCBI, reference: MT319027_1) and a COI sequence (obtained from NCBI, reference: KJ012350_1) from *D. puntazzo* as inputs, together with filtered Illumina reads. The circularised mitochondrial genome was subsequently annotated using MITOS2 software (http://mitos2.bioinf.uni-leipzig.de/index.py, accessed in 10th October 2024). Annotations were manually curated using Geneious Prime 2021.0.3.

The annotation of the mitogenome indicated a similar length (16,642 bp) and the same number of genes (13 genes, 2 rRNA, and 22 tRNA) as previous studies on the same species[48]. Our new mitogenome assembly is complete and contributes to enriching the available sequence diversity (Fig. 2).

Data Records

All the sequencing data generated in this study are available at the ENA European Nucleotide Archive https://identifiers.org/ena.embl: PRJEB49350[49], including Pacbio long reads (ERR13946750), Illumina short reads (ERR7747938 and ERR7747939), OmniC data (ERR7747940 and ERR7747941) and RNAseq paired-end reads from mouth (forward: ERR13946757 and reverse: ERR13946758), from muscle (forward: ERR13946753 and reverse: ERR13946755) and from eye (forward: ERR13946751 and reverse: ERR13946752). The final genome assembly can also be found at ENA under the accession number ERZ21879019 and NCBI (GCA_963678695.2[50]). A summary table of the length and GC content of the scaffolds generated for *D. puntazzo* is available at FigShare[51]. Nuclear and mitochondrial genome annotations are available at Figshare[51] and at https://github.com/EvolutionaryGenetics-UB-CEAB/Diplodus_puntazzo_genome/.

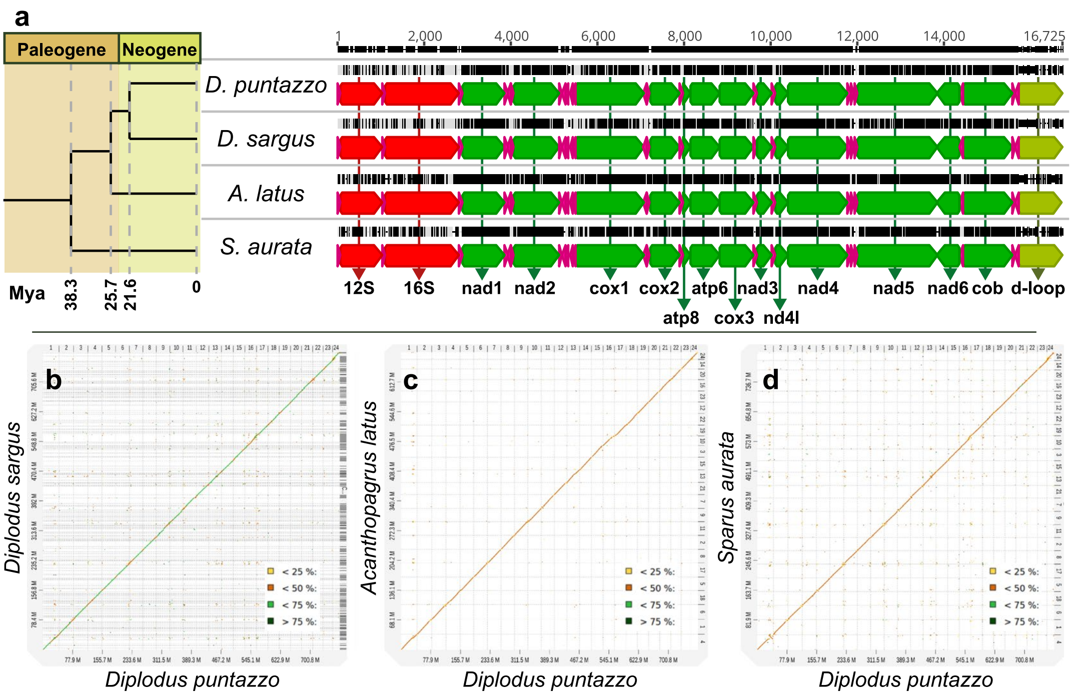

Fig. 3 Syntenic relationships between *D. puntazzo*, *D. sargus*, *A. latus* and *S. aurata*. (**a**) Divergence time among the four fish species and alignment of their mitochondrial genomes showing their annotation. Dot plots show syntenic similarity between the genome assemblies of *D. puntazzo* and *D. sargus* (**b**), *D. puntazzo* and *A. latus* (**c**), and *D. puntazzo* and *S. aurata* (**d**).

Technical Validation

To evaluate the quality of the final chromosome-level assembly we used different approaches. We estimated the genome assembly coverage by back-mapping initial CLR and WGS to the assembly. To do so, we used the pipeline backmap.pl v.0.4, which implements Minimap2 v.2.24 and BWA-mem v.0.7.17, for CLR and WGS respectively, and visualised the results with Qualimap v.2.2.1 and multiQC v.1.8. We obtained contiguity statistics using QUAST, and genome completeness was corroborated by detecting universal single-copy orthologs of Metazoan (metazoa_odb10) and Actinopterygii (Actinopterygii_odb10) using BUSCO v.5.2.2. Finally, we checked for DNA contaminations using BLAST v.2.12 on the non-redundant (nr) NCBI database and graphically represented the results using BlobTools v.2.0. Syntenic analyses between the *D. puntazzo* genome and the genome assemblies of *Diplodus vulgaris*, *A. latus* and *S. aurata* were obtained using D-genies software, selecting the option Minimap2 v2.28 to align genomes (https://dgenies.toulouse.inra.fr/run, accessed in October 2024).

In the final genome assembly, we identified 98.5% of Actinopterygii and 98.3% of Core Vertebrate Genes BUSCO genes (Supplementary Table 8). Overall, 86.05% of the PacBio HiFi reads mapped back to the reference genome, providing a mean coverage of 124.33X, while 99.9% of the Illumina reads mapped against the new reference genome, corresponding to a mean coverage of 67.69X (Supplementary Table 9). Backmapping of Illumina WGS reads resulted in 3,825,128 polymorphic sites (including both SNPs and indels), corresponding to an observed intraspecific heterozygosity of 0.5%. The conserved syntenies detected with the other three closely-related species, *D sargus*, *A. latus* and *S. aurata*, supports the quality of the *D. puntazzo* genome assembly (Fig. 3).

Regarding the genome annotation, the number of protein coding genes ranged from 20,000 to 25,000, with *D. puntazzo* and *S. aurata* having the lower and higher values respectively. However, *D. puntazzo* has the highest number of lncRNAs annotated and the highest mean number of transcripts per gene (Supplementary Table 1). As observed in many species, the gene length, number of exons and GC content of protein-coding genes is higher than in lncRNAs (Supplementary Figure 5[52–54]). The transcriptome has a lower BUSCO completeness than the total assembly (70.42% and 98.41% respectively, Supplementary Table 8), suggesting that a small fraction of genes is still missing in the annotation but present in the assembly. This could be likely improved in future studies by using additional RNAseq data from other tissues.

Code availability

No specific code was used in this study. Data processing and analysis were performed using available software properly cited and the workflow followed is detailed in the methods section.

Received: 6 December 2024; Accepted: 25 March 2025;
Published online: 04 April 2025

References

1. Quéro, J. C. Check-List of the Fishes of the Eastern Tropical Atlantic: Clofeta (1990).
2. Coll, J. *et al*. Using no-take marine reserves as a tool for evaluating rocky-reef fish resources in the western Mediterranean. *ICES J. Mar. Sci.* **70**, 578–590 (2013).
3. Macpherson, E., Gordoa, A. & García-Rubies, A. Biomass size spectra in littoral fishes in protected and unprotected areas in the NW Mediterranean. *Estuar. Coast. Shelf Sci.* **55**, 777–788 (2002).
4. García-Rubies, A. & Zabala i Limousin, M. Effects of total fishing prohibition on the rocky fish assemblages of Medes Islands marine reserve (NW Mediterranean). *Scientia Marina* **54**, 317–328 (1990).
5. Albouy, C., Mouillot, D., Rocklin, D., Culioli, J. M. & Le Loc'h, F. Simulation of the combined effects of artisanal and recreational fisheries on a Mediterranean MPA ecosystem using a trophic model. *Mar. Ecol. Prog. Ser.* **412**, 207–221 (2010).
6. Sala, E. & Ballesteros, E. Partitioning of space and food resources by three fish of the genus Diplodus (Sparidae) in a Mediterranean rocky infralittoral ecosystem. *Mar. Ecol. Prog. Ser.* **152**, 273–283 (1997).
7. Greco, S. *et al*. Controlled spawning and larval development in the sharpsnout seabream (Diplodus puntazzo Sparidae). in *Production, environment and quality. Bordeaux Aquaculture'92* (eds. Barnabè, G. & Kestement, P.) 127–135 (European Aquaculture Society, Special Pubblication n° 18, 1992).
8. García-Charton, J. A. *et al*. Multi-scale spatial heterogeneity, habitat structure, and the effect of marine reserves on Western Mediterranean rocky reef fish assemblages. *Mar. Biol.* **144**, 161–182 (2004).
9. Coll, M. *et al*. The Mediterranean Sea under siege: spatial overlap between marine biodiversity, cumulative threats and marine reserves. *Glob. Ecol. Biogeogr.* **21**, 465–480 (2012).
10. Guidetti, P. & Sala, E. Community-wide effects of marine reserves in the Mediterranean Sea. *Mar. Ecol. Prog. Ser.* **335**, 43–56 (2007).
11. Di Franco, A., Bussotti, S., Navone, A., Panzalis, P. & Guidetti, P. Evaluating effects of total and partial restrictions to fishing on Mediterranean rocky-reef fish assemblages. *Mar. Ecol. Prog. Ser.* **387**, 275–285 (2009).
12. Appolloni, L. *et al*. Does full protection count for the maintenance of β-diversity patterns in marine communities? Evidence from Mediterranean fish assemblages. *Aquat. Conserv.* **27**, 828–838 (2017).
13. Astruch, P. *et al*. Assessment of the conservation status of coastal detrital sandy bottoms in the Mediterranean Sea: an ecosystem-based approach in the framework of the ACDSEA project. in *Proceedings of the 3rd symposium on the conservation of coralligenous and other calcareous bio-constructions, Antalya, Turkey, 15-16 January 2019* (eds. Langar, H. & Ouerghi, A.) 23–29 (RAC/SPA publications, Tunis, 2019).
14. Mouine, N., Francour, P., Ktari, M. H. & Chakroun-Marzouk, N. Reproductive biology of four Diplodus species Diplodus vulgaris, D. annularis, D. sargus and D. puntazzo (Sparidae) in the Gulf of Tunis (central Mediterranean). *J. Mar. Biol. Assoc. U. K.* **92**, 623–631 (2012).
15. Domínguez-Seoane, R., Pajuelo, J. G., Lorenzo, J. M. & Ramos, A. G. Age and growth of the sharpsnout seabream Diplodus puntazzo (Cetti, 1777) inhabiting the Canarian archipelago, estimated by reading otoliths and by backcalculation. *Fish. Res.* **81**, 142–148 (2006).
16. Pajuelo, J. G., Lorenzo, J. M. & Domínguez-Seoane, R. Gonadal development and spawning cycle in the digynic hermaphrodite sharpsnout seabream Diplodus puntazzo (Sparidae) off the Canary Islands, northwest of Africa. *J. Appl. Ichthyol.* **24**, 68–76 (2007).
17. Raventos, N., Torrado, H., Arthur, R., Alcoverro, T. & Macpherson, E. Temperature reduces fish dispersal as larvae grow faster to their settlement size. *J. Anim. Ecol.* **90**, 1419–1432 (2021).
18. Torrado, H. *et al*. Genomic basis for early-life mortality in sharpsnout seabream. *Sci. Rep.* **12**, 17265 (2022).
19. Raventós & Macpherson Planktonic larval duration and settlement marks on the otoliths of Mediterranean littoral fishes. *Mar. Biol.* **138**, 1115–1120 (2001).
20. Torrado, H. *et al*. Impact of individual early life traits in larval dispersal: A multispecies approach using backtracking models. *Prog. Oceanogr.* **192**, 102518 (2021).
21. Cheminee, A., Francour, P. & Harmelin-Vivien, M. Assessment of Diplodus spp. (Sparidae) nursery grounds along the rocky shore of Marseilles (France, NW Mediterranean). *Sci. Mar.* **75**, 181–188 (2011).
22. García-Rubies, A. & Macpherson, E. Substrate use and temporal pattern of recruitment in juvenile fishes of the Mediterranean littoral. *Mar. Biol.* **124**, 35–42 (1995).
23. Vigliola, L. *et al*. Spatial and temporal patterns of settlement among sparid fishes of the genus Diplodus in the northwestern Mediterranean. *Mar. Ecol. Prog. Ser.* **168**, 45–56 (1998).
24. Macpherson, E. *et al*. Mortality of juvenile fishes of the genus Diplodus in protected and unprotected areas in the western Mediterranean Sea. *Marine Ecology Progress Series* **160**, 135–147 (1997).
25. Macpherson, E. Ontogenetic shifts in habitat use and aggregation in juvenile sparid fishes. *J. Exp. Mar. Bio. Ecol.* **220**, 127–150 (1998).
26. Vitturi, R., Libertini, A., Mazzola, A., Colomba, M. S. & Sara, G. Characterization of mitotic chromosomes of four species of the genus Diplodus: karyotypes and chromosomal nucleolar organizer region phenotypes. *J. Fish Biol.* **49**, 1128–1137 (2014).
27. Bargelloni, L. *et al*. The Atlantic-Mediterranean transition: discordant genetic patterns in two seabream species, Diplodus puntazzo (Cetti) and Diplodus sargus (L.). *Mol. Phylogenet. Evol.* **36**, 523–535 (2005).
28. López, A., Carreras, C., Pascual, M. & Pegueroles, C. Evaluating restriction enzyme selection for reduced representation sequencing in conservation genomics. *Mol. Ecol. Resour*. (2023).
29. Galià-Camps, C., Pegueroles, C., Turon, X., Carreras, C. & Pascual, M. Genome composition and GC content influence loci distribution in reduced representation genomic studies. *BMC Genomics* **25**, 410 (2024).
30. Corominas, M. *et al*. The Catalan initiative for the Earth BioGenome Project: contributing local data to global biodiversity genomics. *NAR Genom Bioinform* **6**, lqae075 (2024).
31. Protocol 1: DNA Isolation from Mammalian Tissue Molecular Cloning: A Laboratory Manual. (Cold Spring Harbor Laboratory Press, New York, NY, 2001).
32. Ward, R. D., Zemlak, T. S., Innes, B. H., Last, P. R. & Hebert, P. D. N. DNA barcoding Australia's fish species. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **360**, 1847–1857 (2005).
33. Ranallo-Benavidez, T. R., Jaron, K. S. & Schatz, M. C. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat. Commun.* **11**, 1432 (2020).
34. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770 (2011).
35. Bertolini, F. *et al*. Whole genome sequencing provides information on the genomic architecture and diversity of cultivated gilthead seabream (Sparus aurata) broodstock nuclei. *Genes (Basel)* **14**, 839 (2023).
36. Wellcome Sanger Institute. Genome of Sparus aurata. *GenBank* https://identifiers.org/insdc.gca:GCA_900880675.2 (2019).
37. Zhu, K.-C. *et al*. A chromosome-level genome assembly of the yellowfin seabream (Acanthopagrus latus; Hottuyn, 1782) provides insights into its osmoregulation and sex reversal. *Genomics* **113**, 1617–1627 (2021).
38. Wellcome Sanger Institute. Genome of Acanthopagrus latus. *GenBank* https://identifiers.org/insdc.gca:GCA_904848185.1 (2020).

39. Fietz, K. *et al*. New genomic resources for three exploited Mediterranean fishes. *Genomics* **112**, 4297–4303 (2020).
40. Centre d'Ecologie Functionelle & Evolutive. Genome of Diplodus sargus. *GenBank* https://identifiers.org/insdc.gca:GCA_903131615.1 (2020).
41. Shao, F., Han, M. & Peng, Z. Evolution and diversity of transposable elements in fish genomes. *Sci Rep* **9**, 15399 (2019).
42. Brown, M. G., De la Rosa, P. M. & Blaxter, M. *A Telomere Identification Toolkit*. https://doi.org/10.5281/zenodo.10091385.
43. Vitturi, R., Libertini, A., Mazzola, A., Colomba, M. S. & Sara, G. Characterization of mitotic chromosomes of four species of the genus Diplodus: karyotypes and chromosomal nucleolar organizer region phenotypes. *J. Fish Biol.* **49**, 1128–1137 (1996).
44. Holt, C. & Yandell, M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* **12**, 491 (2011).
45. Hellenic Center of Marine Reasearch. RNA data of Diplodus puntazzo. *GenBank* https://identifiers.org/bioproject:PRJNA241484 (2014).
46. Manousaki, T. *et al*. The sex-specific transcriptome of the hermaphrodite sparid sharpsnout seabream (Diplodus puntazzo). *BMC Genomics* **15**, 655 (2014).
47. Krzywinski, M. *et al*. Circos: an information aesthetic for comparative genomics. *Genome Res.* **19**, 1639–1645 (2009).
48. Ceruso, M. *et al*. The complete mitochondrial genome of the sharpsnout seabream (Perciformes: Sparidae). *Mitochondrial DNA B Resour* **5**, 2379–2381 (2020).
49. *ENA European Nucleotide Archive*. https://identifiers.org/ena.embl:PRJEB49350 (2021).
50. Universitat de Barcelona. Genome of Diplodus puntazzo. *GenBank* https://identifiers.org/insdc.gca:GCA_963678695.2 (2024).
51. Pegueroles, C. *et al*. *Figshare* https://doi.org/10.6084/m9.figshare.28466042 (2025).
52. Galià-Camps, C. *et al*. Chromosome-level genome assembly and annotation of the black sea urchin Arbacia lixula (Linnaeus, 1758). *DNA Res* **31**, (2024).
53. Pegueroles, C. *et al*. Transcriptomic analyses reveal groups of co-expressed, syntenic lncRNAs in four species of the genus Caenorhabditis. *RNA Biol* **16**, 320–329 (2019).
54. Haerty, W. & Ponting, C. P. Mutations within lncRNAs are effectively selected against in fruitfly but not in human. *Genome Biol* **14**, R49 (2013).

Acknowledgements

We would like to thank The Catalan Initiative for the Earth Biogenome Project (CBP) for coordination and promotion efforts of this genome project. We thank Damian Baranski of the LOEWE Translational Biodiversity Genomics (TBG) project based in the Senckenberg Research Institute and Museum, Frankfurt, for support with lab work. We thank the Genome Technology Center (RGTC) at Radboudumc for the use of the Sequencing Core Facility (Nijmegen, The Netherlands), which provided the PacBio SMRT sequencing service on the Sequel IIe platform. This research was funded by the Institut d'Estudis Catalans under the Catalan Initiative for the Earth Biogenome Project (PRO2020-S02-Carreras), the project MarGeCh (PID2020-118550RB, funded by MCIN/AEI/10.13039/501100011033) from the Spanish Government and BlueDNA (PID2023-146307OB) funded by MCIN/AEI/10.13039/501100011033 and ERDF/EU. The authors CP, CGC, MP and CC are members of the research group SGR2021-01271, EM and NR are members of the research group SGR2021-00405, both SGR are funded by the Generalitat de Catalunya.

Author contributions

C.C. and M.P. designed the study and acquired funds. E.M. and N.R. sampled the sequenced individual. C.G. established the experimental protocols. Bioinformatic analyses were carried out by C.P., C.G.C., M.B., D.G., T.S. and C.C. All the authors contributed to the final version of the paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-025-04902-3.

Correspondence and requests for materials should be addressed to C.P., C.G.-C. or C.C.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.