Title:

Chromosome assembly for the meagre, *Argyrosomus regius*,
reveals species adaptations and sciaenid sex-related locus
evolution

Running Title:

Meagre genome assembly and analysis

Authors and Affiliations:

Vasileios Papadogiannis¹, Tereza Manousaki¹, Orestis Nousias¹,², Alexandros Tsakogiannis¹,
Jon B. Kristoffersen¹, Constantinos C. Mylonas¹, Constantinos Batargias³, Dimitrios Chatziplis⁴
& Costas S. Tsigenopoulos¹  *

¹ Institute of Marine Biology Biotechnology and Aquaculture (IMBBC), Hellenic Centre for Marine
Research (HCMR), Heraklion, Crete, Greece
² Department of Biology, University of Crete, Greece
³ Department of Animal Production, Fisheries and Aquaculture, University of Patras, Messolonghi,
Greece
⁴ Department of Agriculture, International Hellenic University (IHU), Thessaloniki, Greece

* Corresponding Author
Abstract

The meagre, *Argyrosomus regius*, has recently become a species of increasing economic interest for the Mediterranean aquaculture and there is ongoing work to boost production efficiency through selective breeding. Access to the complete genomic sequence will provide an essential resource for studying quantitative trait-associated loci and exploring the genetic diversity of different wild populations and aquaculture stocks in more detail. Here, we present the first complete genome for *A. regius*, produced through a combination of long and short read technologies and an efficient in-house developed pipeline for assembly and polishing. Scaffolding using previous linkage map data allowed us to reconstruct a chromosome level assembly with high completeness, complemented with gene annotation and repeat masking. We use this new resource to study the evolution of the meagre genome and other Sciaenids, via a comparative analysis of 25 high-quality teleost genomes. Combining a rigorous investigation of gene duplications with base-wise conservation analysis, we identify candidate loci related to immune, fat metabolism and growth adaptations in the meagre. Following phylogenomic reconstruction, we show highly conserved synteny within Sciaenidae. In contrast, we report rapidly evolving syntenic rearrangements and gene copy changes in the sex-related *dmrt1* neighbourhood in meagre and other members of the family. These novel genomic datasets and findings will add important new tools for aquaculture studies and greatly facilitate husbandry and breeding work in the species.

Keywords

Meagre, Genome, Sciaenidae, Adaptation, Evolution, Sex
Introduction

The meagre, *Argyrosomus regius*, is a teleost fish in the family Sciaenidae, commonly known as croakers or drums because of the characteristic croaking sounds they produce. Sciaenids include some commercially important species with sequenced genomes, such as the related Japanese meagre (*Argyrosomus japonicus*)(Zhao et al., 2021), the large yellow croaker (*Larimichthys crocea*) (Ao et al., 2015), the red drum (*Sciaenops ocellatus*) (T. Xu, Li, Chu, & Zheng, 2021) and the spinyhead croaker (*Collichthys lucidus*) (Cai et al., 2019).

Several favourable characteristics of the species, such as fast growth, large body size and low-fat content have attracted interest in meagre aquaculture in recent years, promoting efforts to improve hatchery techniques and culture efficiency (Mylonas, Mitrizakis, Papadaki, & Sigelaki, 2013). Global meagre production amounted to more than 55,500 tonnes in 2019, with 68% of this volume sourced from aquaculture, while the European Union was the second largest producer in the world after Egypt (www.fao.org & www.apromar.es). Previous work on meagre aquaculture has looked into the application of varying dietary sources for culturing the species (Chatzifotis, Panagiotidou, & Divanach, 2012; Chatzifotis et al., 2010), including vegetable (Ribeiro et al., 2015) and insect-based diets (Guerreiro et al., 2020), the study of the development of the digestive system (Papadakis, Kentouri, Divanach, & Mylonas, 2013), as well as meagre reproduction and spawning (Mylonas, Mitrizakis, Castaldo, et al., 2013; Mylonas, Mitrizakis, Papadaki, et al., 2013; Ramos-Júdez et al., 2019).
The recent increased interest in improving meagre breeding and culture has created the need for genetic resources for the species. In this direction, there has been prior research on the development of genetic markers for the study of growth traits, including the meagre muscle and liver transcriptomes (Manousaki et al., 2018), a microsatellite PCR panel (Vallecillos et al., 2022) a broodstock structure analysis (Orestis Nousias et al., 2021) and a recent meagre ddRAD linkage map (O. Nousias et al., 2022). While such efforts provide valuable tools for meagre aquaculture, they remain constrained in the absence of a complete nuclear genomic sequence and would greatly benefit from the availability of a high-quality assembly for the species.

In this study, we present the first nuclear genome for the meagre. Using a combination of long and short read data and the previously published linkage map, we built a chromosome level assembly with high completeness and provide high quality repeat and gene annotations. Using the new assembly, we carried out phylogenetic and evolutionary analyses, focusing on gene duplications and signatures of accelerated evolution in meagre genes. An investigation of duplicated families and fast evolving loci highlighted immune-system, fat metabolism and cancer-related adaptations, paving the way for understanding the rapid growth and large body size of the species. Finally, exploring of synteny around the sex-related dmrt1 locus in sciaenids revealed rapid changes in the genomic neighbourhood in the meagre, offering a primary candidate for follow up reproductive studies in the species.

### Materials and Methods
Sampling and Sequencing

Genomic material for sequencing was isolated from a female *A. regius* individual, collecting a total of 10ml blood in 1/10 volume of heparin. DNA for long read sequencing on an Oxford Nanopore MinION sequencer was extracted on the same day of the collection of 2 ml blood, using the Qiagen Genomic tip (20 G), following the Oxford Nanopore Technologies protocol for DNA extraction from chicken blood. Assessment of DNA integrity quality was done through electrophoresis on a 0.4% w/v Bio-Rad Megabase agarose gel. DNA purity was estimated by Nanodrop ratios and DNA concentration by Qubit. Library preparation was then carried out following manufacturer instructions, using the sequencing kit SQK-LSK109. Two libraries were sequenced for 96 hours on two R9.4.1 flow cells on the MinION sequencer of IMBBC, HCMR. Raw reads were basecalled with Guppy 4.0.11, using the High Accuracy (hac) configuration. A total of 3,003,301 reads were generated with a quality score above 7, with an N50 of 30,600, totalling 37,962,807,176 bp and 99.94% passed quality control (Supplementary Table 1). For Illumina sequencing, two-day old frozen blood was used, following the same extraction protocol described above, similar to Danis et al. 2021 (Danis et al., 2021). Extracted DNA was fragmented via sonication, followed by PCR free library preparation with the Kapa Hyper Prep DNA kit and paired end (150bp) sequencing on an Illumina Hiseq4000 platform (Norwegian Sequencing Center, NSC). A total of 143,224,530 150bp reads were generated, with 85.81% passing quality control (Supplementary Table 1).
RNA extraction was done from material for tissues shown in table 1, homogenised in TRIzol reagent (Invitrogen), under liquid nitrogen. Whole RNA was extracted from the homogenised material following manufacturer instructions. Library preparation for paired end sequencing (150bp) was done with the Illumina TruSeqTM RNA Sample Preparation Kit v2 and libraries were sequenced on the above Illumina Hiseq4000 platform (NSC). More than 100 million 150bp reads were generated for each tissue and more than 90% of reads passed quality control for any tissue (Supplementary Table 2).

**Assembly and Scaffolding**

A kmer counting/distribution strategy was used for genomic data quality assessment, using jellyfish (v2.3.0) for kmer counting (21 bp length) and genomescope (v1.0) to calculate kmer distribution plots (Vurture et al., 2017).

Assembly construction was carried out with the SnakeCube containerised pipeline (Angelova, Danis, Lagnel, & Tsigenopoulos, n.d.). Briefly, sequencing data quality control and trimming is carried out via Trimmomatic (Bolger, Lohse, & Usadel, 2014) (v0.39) and fastQC (FastQC, 2015) (v0.11.8) for short read data and via Nanoplot (de Coster, D’Hert, Schultz, Cruts, & van Broeckhoven, 2018) (1.29.0) and porechop (v0.2.3) (https://github.com/rrwick/Porechop#license) for long read data. An initial long read based assembly is built using Flye (Kolmogorov, Yuan, Lin, & Pevzner, 2019) (v2.6), followed by long read based polishing through Racon (v1.4.12) (https://github.com/isovic/racon) and
Medaka (v0.9.2) (https://github.com/nanoporetech/medaka). Pilon(Walker et al., 2014) (v1.23 ) is then used for error correction and polishing using the short read data. Assembly quality, contiguity and completeness were assessed via Quast(Gurevich, Saveliev, Vyahhi, & Tesler, 2013) (v5.0.2), BUSCO(Simão, Waterhouse, Ioannidis, Kriventseva, & Zdobnov, 2015) (v5.1.0)) and Merqury(Rhie, Walenz, Koren, & Phillippy, 2020).

ALLMAPS(Tang et al., 2015) was used for the scaffolding of assembly contigs based on the meagre Double digest restriction-site associated DNA (ddRAD) linkage map by Nousias et al. (2022) and the species diploid number of chromosomes (2n=48, Soares et al., 2012(Soares et al., 2012)). Preliminary scaffolding efforts suggested that the previous linkage group (LG) I was an artificial merge of two constituent groups, based on non-overlapping mapping of assembly contigs and the unexpectedly large linkage group size compared to the next largest group. Additionally, poor mapping for LG XXIV and its small content in genetic markers suggested that the latter LG was an artifact. After discarding LG XXIV and splitting LG I, we obtained the final scaffolding results presented in this study.

Repeat Element Annotation

RepeatModeller(Smit, 2015) (v10.1) was used for de novo repeat modelling with the Repbase database(Bao, Kojima, & Kohany, 2015), followed by repeat identification and annotation via RepeatMasker(Smit, 2015) (v 4.1.2-p1) while ltr_finder(Z. Xu & Wang, 2007) (v1.07) was used for additional LTR retrotransposon annotation. Repeats identified
from these tools were merged to a final repeat annotation set via RepeatCraft (Wong & Simakov, 2019).

Transcriptome Assembly and Gene Annotation

Transcriptome data quality control were performed with fastQC and trimming with trimmomatic. Reads were mapped to the genome with STAR (Dobin et al., 2013) (v2.7.9a) and separate transcriptome assemblies were built for each tissue via StringTie (Kovaka et al., 2019) (v2.1.4). These transcriptomes were then merged to a final consensus transcriptome assembly via TACO (Niknafs, Pandian, Iyer, Chinnaiyan, & Iyer, 2017).

Based on the transcriptome data, transcript model prediction was done via Mikado (Venturini, Caim, Kaithakottil, Mapleson, & Swarbreck, 2018) (v2.0), incorporating information from intron-exon junction prediction carried out through portcullis (v1.2.0) (https://github.com/EI-CoreBioinformatics/portcullis), open reading frame identification with TransDecoder (v5.5.0) (https://github.com/TransDecoder/TransDecoder) and homology evidence by searching teleost fish and other vertebrate proteomes (dataset presented in Supplementary Table 3) with DIAMOND (Buchfink, Reuter, & Drost, 2021) (v0.9.14). Mikado produced transcript sets were used for training and de-novo gene model prediction with Augustus (Stanke & Morgenstern, 2005) (v3.4.0). Mikado and Augustus datasets were then merged to a final gene annotation set using PASA (Haas, 2003) (v2.4.1). BUSCO was used to assess the completeness of gene annotation datasets. Functional annotation of gene models was carried out through a combination
of annotations from PANTHER (Mi & Thomas, 2009) (v2.0) and EggNOG (Huerta-Cepas et al., 2019) (v5.0).

Orthogroup Identification and Phylogenetic Analyses

OrthoFinder (Emms & Kelly, 2015, 2019) (v2.5.2) was used for gene family inference and one-to-one ortholog identification for phylogenetic reconstruction. After species tree inference, the reconstructed phylogeny was used to re-run the OrthoFinder analysis with the new species tree, with final orthogroups obtained from this corrected run used for downstream analyses.

Alignments of single copy orthologous proteins were built using MAFFT (Katoh, 2002) (v7.4.80), followed by alignment trimming with trimAl (Capella-Gutierrez, Silla-Martinez, & Gabaldon, 2009) (strict mode) (v1.4.rev15). Trimmed alignments were then concatenated to a final super-alignment (super-matrix), which was used for species phylogenetic tree inference through maximum likelihood analysis with RAxML-ng (A Stamatakis, Ludwig, & Meier, 2003; A. Stamatakis, 2014) (v1.0.2), using the JTT model selected by ModelTest-NG (Darriba et al., 2020) and bootstrap resampling with 1,000 bootstrap replicates.

Synteny Analysis

Macro-synteny analysis in Sciaenidae was based on single copy orthologous loci shared by the meagre and all other species, previously identified from the OrthoFinder
analysis. Whole genome synteny plots for single copy orthologs were plotted using Circos (Krzywinski et al., 2009) (v0.69-8).

Synten in the *dmrt1* neighbourhood was searched using publicly available NCBI data or BLAST (Camacho et al., 2009) (v2.11.0) to investigate if genes of interest are proximal in species without annotated loci in NCBI. For this purpose, we used an *e*-value threshold of $10^{-6}$ and a similarity threshold of 70% to assess if matches to protein queries were co-located in the same genomic scaffold in each species, recording the genomic location of the genes.

**Gene Duplication Identification and Analysis**

To identify potential gene duplication events, we combined two tools that use different approaches; CAFE (v4) (M. v. Han, Thomas, Lugo-Martinez, & Hahn, 2013) calculates rapid gene family expansions and contractions based on gene count tables in each family in each species, while GeneRax (Morel, Kozlov, Stamatakis, & Szöllősi, 2020) (v2.0.2) carries out gene tree reconciliation based on the species phylogeny and infers putative duplication events from the corrected gene trees. For CAFE, we used gene count matrices produced for OrthoFinder orthogroups, with a *p*-value threshold of 0.01, after discarding orthogroups with low representation across our phylogeny (genes present in less than 4 species) or with no representation in the meagre. For GeneRax input, we built alignments (MAFFT), starting trees (IQ-TREE (Nguyen, Schmidt, von Haeseler, & Minh, 2015) v1.6.12) and calculated phylogenetic models (IQ-TREE), using our reconstructed phylogeny and the UndatedDL model (default mode). We then filtered
the output of both tools, keeping the orthogroups suggested as containing duplications by both tools that had two or more duplications in CAFE.

**Multiple Whole-Genome Alignment and Conservation Score Analyses**

Based on our reconstructed phylogeny, multiple whole genome alignments for all 25 genomes included in this study were carried out via CACTUS (Armstrong et al., 2020) (v 2.2.1). Based on a multiple genome alignment for each meagre chromosome sequence and our reconstructed phylogeny, we used phyloFit (Siepel & Haussler, 2004) for phylogenetic model fitting for each chromosomal alignment (using the REV substitution model) and phyloP (Pollard, Hubisz, Rosenbloom, & Siepel, 2010) to calculate base-wise conservation scores (using the SPH method and CONACC mode for p-value computation) for each chromosome. Average phyloP (CONACC) scores were calculated for 50kb windows for the 24 meagre chromosomes and plotted using Circos, while average conservation scores were also calculated for all genes and transcripts.

**Gene Ontology Enrichment**

Ontology enrichment was carried out via gProfiler (Raudvere et al., 2019). Ontology enrichment for fast evolving genes (with phyloP score<0) was carried out using the “ordered list” option, ranking genes by decreasing conservation score, using our meagre functional annotation reference set. To carry out ontology enrichment for
OrthoFinder orthogroups with gene duplications, we built an orthogroup annotation reference set, by leveraging non-redundant ontology annotations from meagre and zebrafish genes belonging to each family. Orthology enrichment for duplication containing orthogroups was then carried out as ordered lists, ranked by number of duplications for each species in the phylogeny. Enriched terms for duplications retrieved for each species from this pipeline were then filtered and ranked by adjusted $p$-values (0.01 threshold). These ranked lists were then compared, retaining only meagre terms that were enriched in a maximum of three other species, to discard terms commonly enriched in duplications across species and shortlist terms that are more likely to be specific to meagre duplications.

**Computational Resources**

All computational work described for genome assembly, gene and repeat element annotation, phylogenomic analyses, multiple whole genome alignment and downstream evolutionary analyses were carried out using the computational resources of the IMBBC HPC facility “Zorbas” of HCMR(Zafeiropoulos et al., 2021).

**Ethics and Permits**

This study does not include special animal treatment, only euthanasia and study of the genetic material. Thus, the work carried out in the scope of the study does not fall
within the HCMR code for animal ethics. Furthermore, we followed all appropriate guidelines for animal care and handling (Guidelines for the treatment of animals in behavioral research and teaching. Anim. Behav. 53, 229–234 (1997)).

Results

Assembly and scaffolding

For the construction of the genomic assembly of *A. regius*, we used a combination of ONT minion long read data (>54X coverage) and Illumina short read data (>30X coverage). An initial assessment of sequencing quality of the selected specimen was performed using the short read data, revealing very low levels of heterozygosity (predicted range 0.241%-0.244% with average of 0.242%,) and sequencing error (orange line in Fig. 1a). Assembly of the genome was carried out using the SnakeCube LSGA pipeline and this initial assembly was then scaffolded using the high-density ddRAD based linkage map for the species by Nousias et al. (2022). This led to merging 205 contigs into 24 scaffolds corresponding to the species chromosome number (Figure 1b). The final chromosome level assembly has 696.267 Mb of sequence in 831 contigs (N50 = 27.87 Mb, L50 = 12), with 92.85% of the total length contained in the 24 chromosomal scaffolds (Table 1). Assembly completeness was calculated at 96.915% when assessed through kmer counting using the short read data via Merqury.
(Fig.1c) and at 98.7% when assessed through BUSCO analysis using the Actinopterygii v10 conserved single copy ortholog database (Fig. 1d).

**Structural and functional annotation**

To identify repeat elements, we used RepeatModeller to obtain species specific repeats models, RepeatMasker and ltr_finder to identify different repeat classes and RepeatCraft to merge all other sets to a final repeat annotation set. The total repeat content of the genome was calculated at 25.86% of the total length, while the genome wide distribution of repeat elements is presented in Figure 2.

Transcriptome data acquired from eight tissues via Illumina sequencing was used to guide gene prediction and annotation. Transcriptome models built using Mikado were provided as a training dataset for Augustus, which was used for de novo gene prediction. The acquired de novo gene models were then merged with the Mikado transcript models through the PASA pipeline. The final gene set comprised a total of 24,589 loci, with 637,465 exons and 49,553 transcripts (Table 1). Gene density distribution across the genome and global GC content over a 50kb sliding window are presented in Figure 2.

**Phylogenomic and Gene Duplication analysis of the meagre genome**
Taking advantage of the high quality and high completeness of the genome, we used the predicted gene models to characterize meagre gene homology to other teleost species. Based on the assigned orthogroups, we obtained 2,104 single copy orthologous loci present across teleosts (present in all 25 species included), which we used to carry out phylogenomic reconstruction for the 25 species in our study. This phylogeny confidently resolved the expected placement of the meagre in Sciaenidae (all branches obtained with 100 bootstrap support), with the European sea bass (Dicentrarchus labrax) placed as an outgroup to the family (Figure 3a). Based on these single copy loci, meagre proteins appear to be evolving comparatively slower compared to other members of the family and most teleosts, based on the short branch length of A. regius in the tree. Using divergence time estimates from TimeTree (Kumar et al., 2022), we also calculated divergence times for the meagre at 22.4 million years ago from other members of the family and Sciaenidae as a group at 79.3 million years ago from the European sea bass.

Combining homology and phylogenomic information, we then characterised gene duplication events in the meagre and compared it to other Sciaenidae and teleosts. Filtering duplication predictions from the two complementary approaches described in materials and methods, we shortlisted 116 orthogroups with duplications in the meagre and 17 orthogroups with duplications in the Sciaenid ancestor. Using the filtered duplication datasets from all species of our phylogeny, we performed comparative ranked (by number of duplications) gene ontology enrichment analysis through gProfiler, to isolate functions associated specifically with meagre duplications. For this purpose, we kept ontology terms significantly enriched in the meagre that are significantly enriched in a maximum of three other species. This led to the shortlisting
of 32 ontology terms, mainly related to immune system regulation, immune-related signalling pathway activation and leukocyte cell-cell adhesion (Fig 3.b, Table 2). A notable expansion in MHC2 genes was also included in the orthogroups associated with these functions.

Accelerated evolution in meagre gene duplications

Complementing gene duplication analyses, we searched for signatures of selection across the meagre genome, after multiple whole genome alignment of the 25 genomes included in the study. Through this analysis, 15,206 transcripts from 8,368 genes were characterised as potentially fast evolving (phyloP CONACC score <0) and 32,379 transcripts from 15,299 genes were characterised as slow evolving (phyloP CONACC score >0). In Figure 4a, we present average conservation scores for the top 10% fastest (836 genes with average phyloP score <= -0.195) and slowest (1529 genes with average phyloP score >= 0.317) evolving loci. Across the meagre genome, chromosome 22 has the highest total number of the top 10% fastest loci (85 genes; 10.1% of fastest genes) and the highest ratio relative to its size (4.27 genes/Mb), followed by chromosome 3 with the second highest total (82 genes; 9.8% of fastest genes) and ratio (2.53 genes/Mb).

We then investigated the conservation of 1,051 meagre genes that belong to orthogroups with duplications. This revealed that more than 66.4% (698) of these genes are fast evolving, with only 13.4% (141) characterised as slow evolving and the rest (212) as neutral (Fig 4b). In addition, the average conservation score of fast
evolving genes in duplication associated orthogroups was much lower than that of all fast-evolving genes (average score -0.16 VS -0.07, p-value = 5.72\textsuperscript{-68}), while the score of slow-evolving duplication related genes was higher on average than all slow-evolving genes, but with lower significance (average score 0.19 VS 0.12, p-value = 9.38\textsuperscript{-5}) (Fig.4c).

The top GO terms associated with fast evolving genes included immune system related functions, leukocyte activation and adhesion, cytokine production, response to external biotic stimulus, largely overlapping with terms enriched in duplication associated orthogroups (Table 3). The orthogroups associating with these functions included MHC2, interleukin, interferon and toll-like receptor genes, including \textit{tlr22}, \textit{ifng1}, \textit{il4r} and \textit{il13ra}, which are also among the top 10% fastest evolving genes (Fig.4a). This analysis also revealed enrichment of functions related to lipid/ fatty acid metabolism and specifically to the negative regulation of lipid catabolism (Table 3). Locus \textit{apoc1}, which is associated with these functions, is also found among the top 10% fastest evolving genes (Fig. 4a; phyloP score -0.397). Finally, we also detected fast evolution signatures on cancer related loci \textit{tp53} (phylop score -0.145) and \textit{pim2} (Fig. 4a; phylop score -0.708).

Sciaenid Synteny

To study synteny conservation within Sciaenidae, the genomes of \textit{Larimichthys crocea} and \textit{Collichthys lucidus} were selected for their high contiguity and completeness. First, we investigated conservation of macrosynten
orthogroups with a single locus in the meagre and each other member of the family, plotting the position of these single copy orthologs in the 24 chromosomes of each species as a circos plot presented in Figure 5. Synteny is highly conserved within the family, especially between the meagre and *C. lucidus*. However, differences in synteny patterns can be seen, particularly with *L. crocea* chromosomes 1,2,3,4,6 and *C. lucidus* chromosome 1.

**Sex-related dmrt1 Locus Evolution**

Capitalising on the quality of the new meagre assembly and the observed conserved synteny in the family, we studied microsynteny in the *dmrt1* neighbourhood, aiming to obtain insight into the evolution of sex determination in the Sciaenidae family. In this analysis we focused on *dmrt1* and 9 syntenic genes: *hank1a, cfap157, dmrt3, dmrt2a, fam102a, odf2b, rnf183, piwil2* and *golga3*. To obtain a view of the state of the neighbourhood before the Sciaenid ancestor, we used the European sea bass *Dicentrarchus labrax* and the Nile tilapia *Oreochromis niloticus* as outgroups. At the same time, we also searched the recently published genome of *Argyrosomus japonicus*, to better understand the origin of *A. regius* duplications. Both outgroup species exhibited a highly similar structure in the neighbourhood, though a tandem duplication of *piwil2* occurred in *D. labrax* (Fig. 6). Comparison of the four Sciaenidae to these outgroup genomes revealed a series of genomic alterations in the neighbourhood, including translocations, inversions and duplications (Fig. 6). Parsimoniously, the *L. crocea, A. regius* and *A. japonicus* loci suggest an initial translocation of a block comprising *dmrt2a, odf2b* and *fam102a* and an inversion of *odf2b* and *fam102a* in the Sciaenid ancestor. In *L. crocea* there was a subsequent inversion of *dmrt3* and *dmrt1*. In *C. lucidus*, a segmental block including *dmrt3, cfap157, dmrt1, rnf183, piwil2* duplicated in tandem and was then inverted, with an
additional duplication of cfap157 upstream of these duplicate blocks, while there was an
inversion and a translocation of the dmrt2a, odf2b, fam102a block downstream of golga3, with
additional tandem duplications of fam102a (3 copies) and odf2b (2 copies). In A. regius, a
series of tandem duplications have resulted in 3 copies of dmrt3, 2 copies of cfap157 and 2
copies of dmrt1. Based on the fragmented structure of the second dmrt1 copy and the short
length of the remaining fragments, it is possible that the copy may be pseudogenizing (dmrt1-
fr in Fig.6), while comparison to A. japonicus suggests that a block duplication of cfap157 and
dmrt1 happened early in the evolution of the Argyrosomus genus. Strikingly, it also reveals
that the two tandem duplications of dmrt3 are very recent and specific to A. regius.
Phylogenomic reconstruction of the relationships of dmrt1, dmrt3 and dmrt2a (outgroup)
proteins for these loci and species also support the duplications of dmrt1 and dmrt3 being
specific to C. lucidus and the Argyrosomus genus respectively and not shared. In contrast, the
dmrt1 duplication is shared by A. regius and A. japonicus (Supplementary Fig.2).

Discussion

Assembly and Annotation

This work presents the first nuclear genome for the meagre, Argyrosomus regius. This
new chromosome level genomic assembly has very high completeness, as assessed
by two independent methods, kmer counting and conserved ortholog
presence/absence. Furthermore, we offer high quality annotation sets for genes,
transcripts and repeats, providing the means for further downstream dissection of
meagre adaptions and important tools for follow-up experiments on meagre biology.

The assembly and scaffolding of the genomic data to chromosomal contiguity was powered by combining our short-long read sequencing strategy with the high-quality linkage map previously constructed for the species. In contrast to other more demanding technologies for sequencing and scaffolding (e.g., Hi-C), our approach achieved optimal results with minimised sequencing effort in a significantly more time and cost-efficient manner.

It is also important to underline how the two-way exchange of information between linkage map and genomic assembly building allowed for the improvement of both resources. Notably, the new genomic assembly was used to assess and validate linkage group structure, in order to subsequently use the linkage groups for scaffolding.

As seen in Supplementary Fig. 1 and detailed in materials and methods, contig mapping on the linkage map guided the breakage of large LG I into two separate linkage groups, while supporting the discarding of LG XXIV, due to poor and uneven mapping of genomic regions. Following these two amendments, the linkage map successfully guided the scaffolding of the genome to a chromosomal level, with 92.85% of genome length in the 24 chromosomes.

The genome size (696Mb) suggested by the new assembly is slightly larger than that of the large yellow croaker L. crocea (679Mb) and smaller than that of C. lucidus (877Mb) and the repeat content (25.86%) also ranges between that of the former (18.1%) and the latter (34.68%). The number of annotated genes (24,589) is closer to the number predicted in L. crocea (25,401), while C. lucidus has more predicted genes than either (28,602). Interestingly, we identified a much larger number of single-copy orthologs between the meagre and L. crocea (11,225 genes), than C. lucidus (8,175).
Meagre adaptations highlighted by genomic analyses

The reconstructed phylogeny built from the assigned 2,104 single copy ortholog proteins confirmed the placement of the meagre and Sciaenidae in the expected positions, with *L. crocea* and *C. lucidus* grouping together and the European sea bass placed as the outgroup to the family. Based on the branch length (number of substitutions per site), Sciaenidae genomes appear to be more slowly evolving, on average, compared to other teleost species included, with the meagre genome evolving slower than *L. crocea* and *C. lucidus*.

The short branch of meagre prompted us to look for signatures of selection more closely. For this purpose, we coupled gene duplication search with base-wise conservation analysis both in duplicates and genome-wide, to identify patterns indicating adaptive forces at play. First, ontology analysis of orthogroups with duplications revealed enrichment in many immune system-related processes. However, such duplication patterns can be common, because of the continuous interaction of the immune system with the environment of a species. Therefore, we used a strict, comparative approach to filter and isolate the terms specifically associated with meagre duplications. After shortlisting orthogroups with two or more duplications in CAFE that GeneRax also confirms, we selected ontology terms significantly enriched in the meagre and a maximum of three other species. While this may have discarded processes that correspond to potentially real meagre adaptations, we strived to offer high confidence candidates for targeted follow up analyses. Among
top biological functions shortlisted in this fashion, categories with potential for such focused studies include cell-cell adhesion (leukocyte cell-cell adhesion, cell-cell adhesion via plasma-membrane adhesion molecules), immune signalling cascade (immune response-activating signal transduction, interferon-gamma-mediated signaling pathway, negative regulation of protein activation cascade) and antigen sampling in mucosal-associated lymphoid tissue.

Complementing the gene duplication data, enrichment analysis carried out on fast evolving genes ordered by conservation score provides further support on meagre adaptations relating to these and other immune system functions, with top enrichment terms in this analysis including leukocyte activation, positive regulation of cytokine production and leukocyte cell-cell adhesion. Interestingly, though not unexpectedly, the majority of genes in orthogroups with duplications exhibit lower than average conservation score. The overlap between these datasets may contribute to the shared enrichment terms observed, though the list of fast evolving genes is much larger than the duplication containing orthogroups. Noteworthy fast evolving immune related genes include interferons and their receptors (ifng1, ifngr2, igfng1r, igfngr2), interleukins and their receptors (e.g. il6st, il6r, il17a,il17c, il17rc, il16, il15l, il15r, il12ba, il12bb, il12rb2, il10, il10ra, il10rb), toll-like receptors (tlr1, tlr5a, tlr7, tlr22) and major histocompatibility complex class genes, with mhc2 genes (classified as mhc2da) showing both rapid expansion and fast evolution.

In addition to immune adaptations, conservation analysis highlighted fast evolving loci that offer intriguing candidates for studying the evolution and regulation of fast growth and large body size in the meagre. Ontology enrichment revealed fast evolution in genes associated with lipid metabolism, catabolism and fatty acid derivative
biosynthesis. This list includes gene candidates with described functions in these processes, such as apoc1(Fuior & Gafencu, 2019)(also among the top 10% fastest evolving genes), bscl2(Chen et al., 2013), ces3(Wei et al., 2010), dbi(Bouyakdan et al., 2015) and fads2(Glaser, Heinrich, & Koletzko, 2010). This evolutionary signature could indicate adaptations on lipid catabolism, potentially connected to rapid growth, making these loci of great importance for studies on increasing aquaculture efficiency for the species. Intriguingly, we also detected low conservation score on two cancer related loci, tumour protein P53 (tp53) and pim2. These genes are involved in the regulation of cell cycle, cell death and proliferation(Amir et al., 2017; Kronschnabl, Grünweller, Hartmann, Aigner, & Weirauch, 2020; Thomas, Kelly, & Strasser, 2022; Yan et al., 2003), with tp53 duplications in elephants previously suggested to be important for countering cancer increase due to large size in these animals(Sulak et al., 2016). Importantly, pim2 has the fourth fastest evolving transcript in the meagre genome, indicating high selection pressure on the locus. Therefore, we speculate the gene could be involved in promoting cell survival, potentially an adaptation connected to the large body size of the meagre.

**Sciaenid sex related dmrt1 neighbourhood evolution**

Single-copy ortholog synteny between the meagre and the other two Sciaenidae is highly conserved, as seen in Fig.3. Strikingly, although C. lucidus is more closely related to L. crocea, we observe higher synteny conservation between C. lucidus and A. regius, with one-to-one chromosome correspondence between all chromosomes and only few small-scale translocations in C. lucidus compared to the meagre. In
contrast, larger scale rearrangements in *L. crocea* include a fusion of regions in chromosome 1 corresponding to areas from meagre chromosomes 9, 13 and 22, as well as a fusion of regions in chromosome 2 corresponding to meagre chromosomes 3 and 18, with some sizable rearrangements also found in *L. crocea* chromosomes 3, 4 and 6.

Previous studies have suggested Sciaenidae share a relatively stable karyotype of 2n=48 (Cai et al., 2019; H. Lin et al., 2021), which is observed in *L. crocea* (Ao et al., 2015), in our new meagre genome (built from a female individual) and previous studies on the meagre (Soares et al., 2012), as well as in the recent *Argyrosomus japonicus* genome (Zhao et al., 2021), which is built from a male individual. Building on this work, we hypothesised that *C. lucidus* may consist a divergent case, with *L. crocea* showing a more typical karyotype of the family. Previous work has suggested *L. crocea* has an XY sex determination system, indicating chromosome 3 as the putative sex chromosome and highlighting *dmrt1* as a sex determination locus candidate (H. Lin et al., 2021). This DM-domain containing transcription factor is a major regulator of male gonad development (Lindeman et al., 2015; Matson et al., 2011) and plays a key role in sex determination in mammals (Matson et al., 2011), birds (Ioannidis et al., 2021), turtles (Ge et al., 2017), frogs (Ma, Rodrigues, Sermier, Brelsford, & Perrin, 2016; Yoshimoto et al., 2010) and other teleost species, including medaka (Lutfalla, Roest Crollius, Brunet, Laudet, & Robinson-Rechavi, 2003; Nanda et al., 2002), zebrafish (Webster et al., 2017), Chinese tongue sole (Cui et al., 2017), spotted scat (Mustapha et al., 2018) and mandarin fish (C. Han et al., 2021).

Our analysis into synteny conservation of *dmrt1* and neighbouring genes revealed these 10 loci have remained in a single highly conserved block across teleost fish.
(Fig.6; Supplementary Fig.3), with all 10 genes linked in 7 non-Sciaenid teleosts, 9 of the genes retaining synteny in zebrafish and 4 other teleosts and 8 found in human chromosome 9 in two syntenic blocks (Supplementary Fig.3). Comparison to Nile tilapia and European sea bass indicated high synteny conservation in the neighbourhood outside Sciaenidae. As such, the translocation of the dmrt2a, odf2b, fam102a block in the Sciaenid ancestor is notable and we suggest the disconnect of dmrt3 and dmrt2a may reflect a change in the regulation of the neighbourhood that would be worth investigating in future studies.

Among Sciaenids, the L. crocea area is probably the most similar to the ancestral state, with an inversion of dmrt3-dmrt1, while C. lucidus presents an exception studied so far with copy number and synteny changes in the area, combining inversion and translocation events with segmental and tandem duplications. As mentioned, L. crocea uses an XY sex determination system and dmrt1 has been highlighted as the sex determination locus, with specific deletions in the neighbourhood found in male fish (A. Lin et al., 2017; H. Lin et al., 2021). In contrast, C. lucidus has a complex X1X1X2X2/X1X2Y system, with chromosomes 1 and 7 suggested as the X1 and X2 chromosomes, while the dmrt1 neighbourhood is located on chromosome 11 (Xiao et al., 2020). The evolution of C. lucidus sex determination must have happened in a relatively short time, after the split from the shared ancestor with L. crocea circa 14.5 million years ago. Thus, the comparably fast and extensive changes in the C. lucidus dmrt1 area may have accompanied the transition to this new system, maybe due to a disconnection of dmrt1 from its ancestral role in male sex determination. Alternatively, the highly modified dmrt1 neighbourhood could have unknown roles in sex determination or gonad differentiation that remain to be determined.
The structure of the dmrt1 neighbourhood in the Argyrosomus genus is similar to L. crocea and the ancestral state. Based on the fragmented, yet detectable sequence of the dmrt1-fr copy, we suggest the segmental duplication of cfap157 and dmrt1 to be relatively recent in the genus, while the two tandem duplications of dmrt3 appear specific to A. regius, based on the comparison to A. japonicus. Considering the roles of dmrt1 in sex determination and the functions of neighbouring genes in gonadal fate determination and spermatogenesis, these rapid changes in the neighbourhood are highly suggestive of adaptive evolution in an evolutionarily short timeframe. Therefore, we hypothesise the dmrt1 neighbourhood participates in meagre sex determination, comparable to L. crocea, while recent duplications possibly reflect recent evolutionary changes in gonadal differentiation or sex determination in the species.

Acknowledgements

This study has received funding from the Hellenic Republic and the EU through the "MeagreGen" project under the call "Special Actions – AQUACULTURE" in the Operational Programme "Competitiveness Entrepreneurship and Innovation (EPAnEK 2014-2020)". The computational work in this study was carried out through the resources of the IMBBC (Institute of Marine Biology, Biotechnology and Aquaculture) HPC facility of the HCMR (Hellenic Centre for Marine Research). Funding for establishing the IMBBC HPC has been received by the MARBIGEN (EU Regpot) project, LifeWatchGreece RI and the CMBR (Centre for the study and sustainable exploitation of Marine Biological Resources) RI.

References


Cai, M., Zou, Y., Xiao, S., Li, W., Han, Z., Han, F., ... Wang, Z. (2019). Chromosome assembly of Collichthys lucidus, a fish of Sciaenidae with a multiple sex chromosome system. Scientific Data, 6(1). doi: 10.1038/s41597-019-0139-x


Sulak, M., Fong, L., Mika, K., Chigurupati, S., Yon, L., Mongan, N. P., ... Lynch, V. J. (2016). TPS3 copy number expansion is associated with the evolution of increased body size and an enhanced DNA damage response in elephants. doi: 10.7554/eLife.11994.001


**Data Accessibility**

Raw read data and the meagre assembly have been deposited to the European Nucleotide Archive (ENA) under study accession PRJEB56176. Gene and repeat annotations, functional annotation and gene ontology reference information can be found in the supplementary data files.

**Author Contributions**

VP carried out all computational work including genome assembly construction, gene and repeat annotation, phylogenomics analyses, whole genome alignments, duplication analyses, evolutionary analyses, synteny analysis and ontology enrichment and wrote the initial paper draft. CT and TM conceived and designed the experiments and sequencing strategy and supervised the work. JBK and AT performed dissections and nucleic acid extractions. JBK carried out MinION sequencing in IMBBC, HCMR. ON constructed and provided the meagre ddRAD linkage map and assisted with linkage map based scaffolding. CCM, CB, DC contributed to the interpretation of the findings. All authors contributed towards paper writing.
Figure 1 Sequencing and assembly quality assessment. a. GenomeScope kmer distribution (kmer length = 21) for short reads used for genome assembly. b. Scaffolding of 205 contigs (coloured and white bars) into 24 chromosomal scaffolds, based on the ddRAD linkage map. The red box marking Chromosome 12 corresponds to the N50 and L50 of the assembly. c. Merqury kmer distributions (kmer length = 21) for meagre genome assembly, plotting kmers found in 1 to 4 or more copies in the assembly, or only in reads separately. d. BUSCO analysis results identifying 98.7% of ultra-conserved single copy orthologs from the actinopterygii10 database, with 1.3% and 0.4% duplicated and fragmented BUSCO orthologs respectively.
Figure 2 Genome content in genes, repeats and GC. Meagre chromosomal scaffolds are plotted as coloured bars in the outer circle, followed by the gene density heatmap (over 1Mb windows) in the blue circle, the repeat content heatmap (over 50kb windows) in the red circle and the GC content distribution plot (over 50kb windows) in the innermost circle.
Figure 3 Phylogenomic placement of the meagre and gene duplications in Sciaenidae. a. Reconstruction of the evolutionary phylogenetic relationship of 25 fish species, based on 2,104 single copy orthologous loci, placing the meagre in the Sciaenidae family (blue). Bootstrap support for all branches was calculated at 100%. b. Focused ultra-metric tree of Sciaenidae and the European sea bass as the outgroup (clade in red box in Fig.3a), with predicted divergence times for each branch shown at branch nodes. Red numbers designate total number of orthogroups with gene duplication events. Selected top terms identified from gene ontology enrichment analysis of meagre duplications are shown in blue.
Figure 4 Conservation analysis of meagre gene duplicates. a. Circos plot of average phyloP (CONACC) score for top 10% fastest (red) and slowest (blue) evolving genes (red circle: 10% fastest: score <= -0.195, red triangle: 1% fastest: score <= -0.402, blue circle: 10% slowest: score >= 0.317, blue rectangle: 1% slowest: score >= 0.732). b. Number of genes from orthogroups with duplications that are fast evolving (phyloP score < 0), slow evolving (phyloP score > 0) or neutrally evolving. c. phyloP score of all non-neutrally evolving transcripts and transcripts of genes from orthogroups with duplications, with the average for fast evolving (phyloP score < 0) transcripts on the left and slow evolving (phyloP score > 0) transcripts on the right.
Figure 5 *Sciaenidae* single copy ortholog synteny. Circos plot of one-to-one synteny of single copy orthologous loci in Sciaenidae genomes (CLUC: *Collichthys lucidus*, LCRO: *Larimichthys crocea*, AREG: *Argyrosomus regius* – meagre). Loci in each meagre chromosome (colour coded differently) are connected with coloured lines (corresponding to meagre chromosomal colour codes) to orthologous loci in other genomes (orange for *C. lucidus*, green for *L. crocea*).
Figure 6 Sciaenidae sex-related dmrt1 locus synteny and evolution. Synteny of genes in the dmrt1 neighbourhood. The neighbourhood in each taxon is contained in a large white or grey rectangle, with taxon names on the bottom left. Each gene is represented as a coloured box, with distances and sizes non representative of real quantities. Duplications are indicated by bold black boxes, translocations are indicated by dotted boxes and arrows, inversions of gene blocks are indicated by two-way overlapping black arrows.
### Table 1 Assembly and annotation metrics.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Initial Assembly</th>
<th>Chromosome Assembly</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Length (Mb)</td>
<td>696249749</td>
<td>696267849</td>
</tr>
<tr>
<td>Contigs /Scaffolds</td>
<td>1012</td>
<td>831</td>
</tr>
<tr>
<td>Largest Contig/Scaffold GC (%)</td>
<td>41.15</td>
<td>41.15</td>
</tr>
<tr>
<td>N50 (Mb)</td>
<td>7.81</td>
<td>27.87</td>
</tr>
<tr>
<td>N75 (Mb)</td>
<td>2.80</td>
<td>24.50</td>
</tr>
<tr>
<td>L50</td>
<td>23</td>
<td>12</td>
</tr>
<tr>
<td>L75</td>
<td>61</td>
<td>19</td>
</tr>
<tr>
<td># Ns / 100kb</td>
<td>0</td>
<td>2.6</td>
</tr>
<tr>
<td>Length in 24 Scaffolds (Mb)</td>
<td>360890899</td>
<td>646448193</td>
</tr>
<tr>
<td>% of Genome in 24 Scaffolds</td>
<td>51.83%</td>
<td>92.85%</td>
</tr>
</tbody>
</table>

### Table 2 Gene Duplications Ontology Enrichment. Selected top terms from GO enrichment analysis of meagre duplications. Certain terms were omitted to avoid redundancy, with full table available in Supplementary Table 6.

<table>
<thead>
<tr>
<th>term_id</th>
<th>term_name</th>
<th>adjusted p_value</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0002253</td>
<td>activation of immune response</td>
<td>1.00E-04</td>
</tr>
<tr>
<td>GO:0007159</td>
<td>leukocyte cell-cell adhesion</td>
<td>3.88E-04</td>
</tr>
<tr>
<td>GO:0002757</td>
<td>immune response-activating signal transduction</td>
<td>3.43E-03</td>
</tr>
<tr>
<td>GO:0060333</td>
<td>interferon-gamma-mediated signaling pathway</td>
<td>5.88E-03</td>
</tr>
<tr>
<td>GO:0031347</td>
<td>regulation of defense response</td>
<td>8.02E-03</td>
</tr>
<tr>
<td>GO:0098742</td>
<td>cell-cell adhesion via plasma-membrane adhesion molecules</td>
<td>8.48E-03</td>
</tr>
<tr>
<td>GO:0002404</td>
<td>antigen sampling in mucosal-associated lymphoid tissue</td>
<td>9.47E-03</td>
</tr>
<tr>
<td>GO:2000258</td>
<td>negative regulation of protein activation cascade</td>
<td>9.48E-03</td>
</tr>
<tr>
<td>GO:0002218</td>
<td>activation of innate immune response</td>
<td>1.22E-02</td>
</tr>
<tr>
<td>GO:0050870</td>
<td>positive regulation of T cell activation</td>
<td>1.38E-02</td>
</tr>
</tbody>
</table>
Table 3 Fast Evolving Genes Ontology Enrichment. The top table shows selected top terms from GO enrichment analysis of meagre fast evolving genes (phyloP score < 0), with certain terms omitted to avoid redundancy. The full table available in Supplementary Table 10. The bottom table shows terms from GO enrichment analysis of meagre fast evolving genes that are associated with lipid metabolism.

<table>
<thead>
<tr>
<th>term_id</th>
<th>term_name</th>
<th>adjusted_p_value</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0006955</td>
<td>immune response</td>
<td>4.61E-67</td>
</tr>
<tr>
<td>GO:0002684</td>
<td>positive regulation of immune system process</td>
<td>3.09E-57</td>
</tr>
<tr>
<td>GO:0045321</td>
<td>leukocyte activation</td>
<td>3.06E-46</td>
</tr>
<tr>
<td>GO:0006952</td>
<td>defense response</td>
<td>5.32E-44</td>
</tr>
<tr>
<td>GO:0051707</td>
<td>response to other organism</td>
<td>2.27E-41</td>
</tr>
<tr>
<td>GO:0043207</td>
<td>response to external biotic stimulus</td>
<td>2.61E-41</td>
</tr>
<tr>
<td>GO:0002274</td>
<td>myeloid leukocyte activation</td>
<td>8.53E-41</td>
</tr>
<tr>
<td>GO:0002443</td>
<td>leukocyte mediated immunity</td>
<td>1.09E-40</td>
</tr>
<tr>
<td>GO:0001819</td>
<td>positive regulation of cytokine production</td>
<td>2.03E-40</td>
</tr>
<tr>
<td>GO:1903037</td>
<td>regulation of leukocyte cell-cell adhesion</td>
<td>3.01E-37</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>term_id</th>
<th>term_name</th>
<th>adjusted_p_value</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:1901568</td>
<td>fatty_acid_derivative_metabolic_process</td>
<td>3.09E-06</td>
</tr>
<tr>
<td>GO:0050995</td>
<td>negative_regulation_of_lipid_catabolic_process</td>
<td>1.05E-05</td>
</tr>
<tr>
<td>GO:0001676</td>
<td>long-chain_fatty_acid_metabolic_process</td>
<td>2.01E-04</td>
</tr>
<tr>
<td>GO:0035336</td>
<td>long-chain_fatty-acyl-CoA_metabolic_process</td>
<td>2.23E-03</td>
</tr>
<tr>
<td>GO:0050994</td>
<td>regulation_of_lipid_catabolic_process</td>
<td>3.02E-03</td>
</tr>
<tr>
<td>GO:0045833</td>
<td>negative_regulation_of_lipid_metabolic_process</td>
<td>1.14E-02</td>
</tr>
<tr>
<td>GO:0006631</td>
<td>fatty_acid_metabolic_process</td>
<td>1.25E-02</td>
</tr>
<tr>
<td>GO:0016042</td>
<td>lipid_catabolic_process</td>
<td>2.50E-02</td>
</tr>
<tr>
<td>GO:0006638</td>
<td>neutral_lipid_metabolic_process</td>
<td>3.09E-02</td>
</tr>
<tr>
<td>GO:1901570</td>
<td>fatty_acid_derivative_biosynthetic_process</td>
<td>4.37E-02</td>
</tr>
</tbody>
</table>