UNIVERSITY OF CRETE

Gilthead seabream (*Sparus aurata*) genome assembly, annotation and evolution of gene duplications

https://conxemar.com/en/gilthead-sea-bream

September, 2021

by

Klara Eleftheriadi

A thesis submitted to School of Medicine, University of Crete
for the degree of M.Sc. in Bioinformatics

Supervised by:

Researcher C', ***Dr. Tereza Manousaki*** [1]

Examination Committee:

Associate Professor, ***Dr. Ioannis Iliopoulos*** [2]

Researcher B', ***Dr. Christoforos Nikolaou*** [3]

Researcher C', ***Dr. Tereza Manousaki*** [1]

[1] Institute of Marine Biology, Biotechnology and Aquaculture, Hellenic Centre for Marine Research, Heraklion, Greece

[2] School of Medicine, University of Crete

[3] Institute of Bioinnovation, BSRC "Alexander Fleming"

Η παρούσα διπλωματική εργασία πραγματοποιήθηκε στα πλαίσια του Διϊδρυματικού Προγράμματος Μεταπτυχιακών Σπουδών στη Βιοπληροφορική της Ιατρικής Σχολής του Πανεπιστημίου Κρήτης και του Ιδρύματος Τεχνολογίας & Έρευνας (ΙΤΕ).

Σεπτέμβριος, 2021

ACKNOWLEDGEMENTS

First of all, I would like to thank my main supervisor, Tereza Manousaki, for the opportunity to join the "Genome nerds" (that's how we like to call ourselves in the lab) and for the excellent guidance and support. During these difficult "COVID-19" times and one hard lockdown, she was always there providing me with all the appropriate advice and assistance, trying to always keep me motivated.

I would also like to thank my second supervisor, Costas Tsiggenopoulos, whose great experience and guidance was valuable.

A special thanks to the great team of our lab! Thodoris Danis for his amazing advice in assembly expertise and for his continuous encouragement, Vasilis Papadogiannis for our amazing discussions through coffee breaks and lunch and his assistance, Elsa Iliopoulou for listening when I was moaning in the lab and Nellina Angelova for passing her tech-knowledge in the beginning and beer-drinking and strolls.

I would like to acknowledge our high performance computing cluster admins, Stelios Ninidakis and Antonis Potirakis, because without their help many of my analyses would be impossible, but also the people who created it Evaggelos Pafilis and Dimitrios Sidirokastritis.

One more thanks to my friends, Sophia, Polyxeni and Katerina, who always encourage me and support me no matter how far away we are.

Last but not least, in this section will always be the family. So, the biggest Thank You belongs to my parents and my sister, for supporting and understanding every step of mine.

CONTENTS

ΠΕΡΙΛΗΨΗ

Η Τσιπούρα (*Sparus aurata*) είναι ένα είδος τελεόστεων που ανήκει στην οικογένεια Σπαρίδες, της τάξης των Περκόμορφων. Θεωρείται το ψάρι με την υψηλότερη οικονομική αξία στη μεσογειακή υδατοκαλλιέργεια και έτσι έχει γίνει στόχος δύο προσπαθειών αλληλουχίας πριν από την παρούσα εργασία.

Εδώ, παρουσιάζουμε ένα ενημερωμένο γονιδίωμα υψηλής ποιότητας καθώς και τα γονίδια της Τσιπούρας. Αυτό το νέο γονιδίωμα χρησιμοποιήθηκε για την επιθεώρηση της εξέλιξης των διπλασιασμένων γονιδίων συγκεκριμένα για την Τσιπούρα, χρησιμοποιώντας την ανακατασκευή δέντρων γονιδίων και ανάλυση των διευρυμένων οικογενειών γονιδίων από ένα σύνολο 24 ειδών ψαριών. Το γονιδίωμα που προέκυψε ήταν σε επίπεδο χρωμοσωμάτων (217 contigs, με 24 από αυτά να περιέχουν 98,6% του γονιδιώματος, N50 ίσο με 37 Mbp, BUSCO 98% και 26,919 μοντέλα γονιδίων). Η ανάλυση για την εύρεση των διπλασιασμένων γονιδίων αποκάλυψε ότι το Gilthead seabream έχει ένα τυπικό μοτίβο τελεόστεων σχετικά με τα διπλασιασμένα γονίδια.

Λέξεις -κλειδιά: Τσιπούρα, τελεόστεοι, γονιδίωμα, εύρεση γονιδίων, διπλασιασμένα γονίδια, φυλογενετική ανάλυση, δέντρα γονιδίων

ABSTRACT

Gilthead seabream (*Sparus aurata*) is a teleost fish belonging to the Sparidae family, order Perciformes. It is considered the fish with the highest economic value in the Mediterranean aquaculture and has thus been the target of two sequencing efforts prior to our work.

Here, we present an updated high quality reference genome assembly and annotation for Gilthead seabream. This new resource was used to inspect the evolution of species-specific gene duplications, using gene tree reconciliation and gene family evolution analyses from a dataset of 24 other fish species. The genome assembly pipeline resulted in a high quality chromosome level genome assembly (217 scaffolds, with 24 of them containing 98.6% of the genome, N50 equal to 37 Mbp, BUSCO completeness score 98% and 26,919 gene models). The gene duplication analysis revealed that Gilthead seabream has a typical teleost gene duplication pattern.

Keywords: Gilthead seabream, Sparus aurata, teleosts, genome assembly, genome annotation, gene duplications, phylogenomic analysis, gene tree reconciliations

1.INTRODUCTION

Genome structures like genes, chromosomes or the entire genome could be affected by duplications (Glasauer and Neuhauss, 2013). Duplicated genes or entire genomes contribute to the increasing diversity, complexity of the organisms and functional diversification (Meyer and Schartl, 1999). A third whole-genome duplication (WGD - "3R Hypothesis) took place in the common ancestor of all teleosts (teleost-specific WGD - TSGD) making ray-finned fish the most evolutionarily diverse group of species (Meyer and Van de Peer, 2005). The TSGD is still present in teleost genomes (Jaillon et al., 2004; Berthelot et al., 2014; Brawand et al., 2014) and the rediploidization process may be more gradual and slower than expected (Lien et al. 2016).

The Gilthead seabream (*Sparus aurata*, Linnaeus 1758) is a sequential hermaphrodite fish species of Sparidae, order Perciformes. It is one of the most important teleost species in aquaculture, making it one of the most well-studied sequential hermaphrodite fish models (Mylonas et al., 2011). Its hermaphroditism is characterized as protandrous, maturing as male by reaching the age of two years and transforming their sex to female at the age of three years (Mylonas et al., 2011).

Most studies focus on immunology, reproductive physiology and nutrition (Pauletto et al., 2018), but the genomic structural information of features of these responses are still being unveiled. Many studies have been conducted by searching specific gene families for duplications and paralogs identification (Sunyer et al., 1997, Tan and Du, 2002, Angotzi et al., 2020), specific type of genes, such as sex biased genes (Pauletto et al., 2018), and gene duplication rate across the whole genome (Pérez-Sánchez et al., 2019).

There are two genome assemblies currently available (Pauletto et al., 2018 and Vertebrates Genome Project) and one genome browser (www.nutrigroup-iats.org/seabreamdb Pérez-Sánchez et al., 2019). Given the economic importance of the species, we embarked this project with the main aim to construct a *de-novo* high quality and contiguous genome assembly, incorporating many different sequencing technologies, resulting in an updated genome assembly for Gilthead seabream. Also, we annotated the generated genome and based on this annotation we scanned the genome of Gilthead seabream for species-specific gene duplications and expanded gene families.

2. MATERIALS AND METHODS

2.1 Genome building

2.1.1 *De-novo* genome assembly

All the analyses were performed on the IMBBC HPC cluster, HCMR, Heraklion, Greece (Zafeiropoulos et al. 2021).

For the construction of the primary *de-novo* genome assembly, we used the LSGA automated pipeline from https://github.com/genomenerds/SnakeCube (Angelova et al., 2021), which uses long read data from minION sequencer along with short read data from Illumina. This combinatorial strategy leverages the long read data to acquire high contiguity and short read data to improve fidelity.

The initial steps included the preprocessing of the raw reads. The quality assessment of the raw Illumina sequence data was performed with FastQC v0.11.8 (Andrews et al., 2010) and the low quality reads and adapters were removed using Trimmomatic v0.39 (Bolger et al., 2014). (MINLEN: 75, SLIDINGWINDOW: 4:15, LEADING: 10, TRAILING: 10, AVGQUAL: 10). The corresponding preprocessing for the ONT reads was performed with Porechop v0.2.4 (https://github.com/rrwick/Porechop). The preliminary assembly was built using a repeat graph assembler, Flye v2.6 (Kolmogorov et al., 2019) and was polished with two rounds of Racon v1.4.12 (Vaser et al., 2017), following up of one round of Medaka v0.9.2 (https://github.com/nanoporetech/medaka) and the final step includes polishing with the short Illumina reads with two rounds of Pilon v1.23 (Walker et al., 2014) (primary assembly).

2.1.2 Scaffolding the *de-novo* genome assembly

Trying to build a chromosome-level assembly we took advantage of almost all the publicly available genomic data of the Gilthead seabream. The total of the SRA numbers used are shown in Table 1. To be able to construct a high-quality genome assembly, we combined four cutting

edge technologies (2 Mate-Pair libraries, 10 PacBio libraries, 5 Arima Genomics HiC libraries and 3 Linkage Maps). Many approaches and combinations of the aforementioned technologies were tested in order to get the expected result, emerging in 3 workflows and 6 produced draft assemblies (Figure 1). Since we obtained a high quality and contiguous assembly from the containerized pipeline (primary assembly), we set three starting points in our pipeline for further scaffolding: (1) Mate-Pair reads, (2) HiC reads and (3) PacBio reads.

In the first approach (Figure 1 - Workflow 1), after filtering and adapter trimming of Mate- Pair raw reads with fastp v0.20.0 (Chen et al., 2018), the preprocessed reads were used to scaffold the primary assembly with SSPACE (-k=3) (Boetzer et al., 2011), which scaffolds the contigs using the SSAKE short-read assembler (Warren et al., 2007). Next, the HiC reads were mapped to the generated Mate-Pair contig set using bwa (Li, 2013) and filtered with Arima Mapping Pipeline (https://github.com/ArimaGenomics/mapping_pipeline) for SALSA2 (Ghurye et al., 2019) scaffolder and with scripts PreprocessSAMs.pl and filterBAM_forHiC.pl from https://github.com/tangerzhang/ALLHiC for ALLHiC (Zhang et al. 2019) scaffolder. SALSA2 uses a maximum matching algorithm. An *a priori* estimation of the chromosome number is not necessary, and determines the potential order and orientation of each contig by analysing the normalized frequency of HiC links between the ends of contigs (Ghurye et al., 2019). On the contrary, ALLHiC requires the number of groups (chromosomes) to partition the contigs with an agglomerative hierarchical clustering algorithm followed by an optimization step to order and orient the contigs within the partitions alongside a Genetic Algorithm. Finally, the genome assembly is constructed in the building step (Zhang et al., 2019). After mapping and filtering, the resulting contig set from SSPACE was subjected to SALSA2 with default parameters (-e GATC, GANTC) and ALLHiC scaffolding in simple diploid model (partition -k 24). PBJelly, a highly automated pipeline, consists of five stages (setup, mapping, support, extraction, assembly, output), maps the long-reads to the draft assembly (English et al., 2012), using BLASR (Basic Local Alignment and Serial Refinement) (Chaisson and Tesler, 2012) and was incorporated in our pipeline to fill in gaps by taking advantage of the long read information from PacBio, after both SALSA2 and ALLHiC scaffold sets, generating two different scaffold-level draft assemblies. As a finishing step, both the draft assemblies were scaffolded against the current reference genome of Gilthead seabream (GCA_900880675.2) with the assistance of RagTag v1.1.1 (ragtag.py scaffold) (Alonge, 2020), which orders and orients the sequences of a draft assembly based on the alignment

to a reference genome. This approach resulted in two different, high quality genome assemblies, assembly saur v1 and assembly saur v2.

For the second approach (Figure 1 - Workflow 2) of our pipeline we used HiC data as an initial point, in order to decipher the order and the orientation of the chromosomes. First, the raw reads were mapped in the primary assembly and then filtered, as described above and the preliminary scaffolding step was carried out using SALSA2, since it was proposed by the developers to scaffold draft assemblies generated from long reads, e.g. Oxford Nanopore, like the one we generated with Flye assembler. Next, PBJelly was used to fill in gaps and RagTag v1.1.1 was used for scaffolding based on the published reference genome (assembly saur v4). During the evaluation process the assembly saur v3 metrics were better, so we rejected saur v1 and v2.

In the third and final approach (Figure 1 - Workflow 3), we tried to extend the contigs of the primary assembly, correct misjoints, fix structural rearrangements and fill the gaps using PBJelly. After that, we combined HiC and Mate-Pair reads, alternately. This means that SALSA2 and SSPACE (assembly saur v4) were used and then vice versa SSPACE and SALSA2 (assembly saur v5). Finally, we evaluated the assemblies, and saur v5 resulted in a slightly better genome compared to assembly saur v3 and v4, since we have already rejected v1 and v2.

2.1.3 Chromosome-level assembly using Linkage maps

Despite the fact that the already produced genome was highly contiguous, we further improved the quality by using linkage mapping. Three high-density linkage maps (Pauletto et al., 2018) with 11,572, 14,481 and 14,506 single-nucleotide polymorphism (SNP) markers mapped against the assembly saur v6 using HISAT2 and used to join and order the scaffolds based on the 24 chromosomes with the software ALLMAPS (Tang et al., 2015). The ALLMAPS algorithm is implemented in two phases. In phase 1, first, the orientation of the scaffolds is calculated through eigenvectors and secondly, the ordering is considered analogous to the 'Traveling Salesman Problem' and thus determined using CONCORDE algorithm (Mulder and Wunsch II, 2003). Sometimes, the complexity of a genome increases the suboptimal results of Phase 1. However, phase 2 refines the order and orientation of the scaffolds using a Genetic Algorithm, reaching at the chromosome level assembly.

All the resulting assembly statistics were calculated with QUAST v5.0.2 (Gurevich et al., 2013) and assessed with BUSCO v4.1.4 (Simão et al., 2015) against the Actinopterygii odb10 dataset in genome mode.

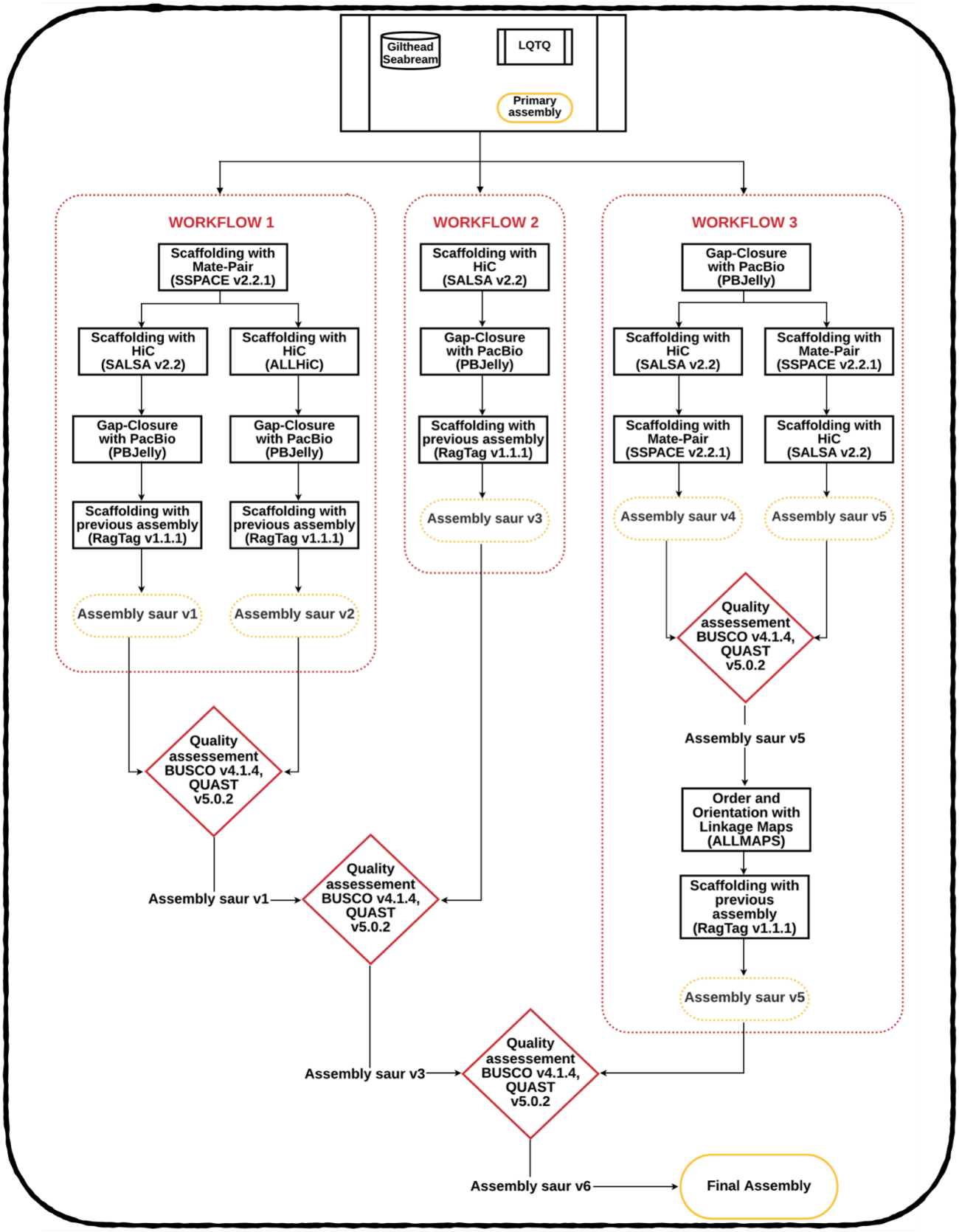
Gilthead Seabream
LQTQ
Primary assembly
WORKFLOW 1
Scaffolding with Mate-Pair (SSPACE v2.2.1)
Scaffolding with HiC (SALSA v2.2)
Scaffolding with HiC (ALLHiC)
Gap-Closure with PacBio (PBJelly)
Gap-Closure with PacBio (PBJelly)
Scaffolding with previous assembly (RagTag v1.1.1)
Scaffolding with previous assembly (RagTag v1.1.1)
Assembly saur v1
Assembly saur v2
Quality assessement BUSCO v4.1.4, QUAST v5.0.2
Assembly saur v1
WORKFLOW 2
Scaffolding with HiC (SALSA v2.2)
Gap-Closure with PacBio (PBJelly)
Scaffolding with previous assembly (RagTag v1.1.1)
Assembly saur v3
Quality assessement BUSCO v4.1.4, QUAST v5.0.2
Assembly saur v3
WORKFLOW 3
Gap-Closure with PacBio (PBJelly)
Scaffolding with HiC (SALSA v2.2)
Scaffolding with Mate-Pair (SSPACE v2.2.1)
Scaffolding with Mate-Pair (SSPACE v2.2.1)
Scaffolding with HiC (SALSA v2.2)
Assembly saur v4
Assembly saur v5
Quality assessement BUSCO v4.1.4, QUAST v5.0.2
Assembly saur v5
Order and Orientation with Linkage Maps (ALLMAPS)
Scaffolding with previous assembly (RagTag v1.1.1)
Assembly saur v5
Quality assessement BUSCO v4.1.4, QUAST v5.0.2
Assembly saur v6
Final Assembly

Figure 1. The Gilthead seabream (*Sparus aurata*) genome assembly pipeline, with the three included workflows.

2.2 Genome annotation

2.2.1 Genome masking

We identified the repetitive elements of the genome of Gilthead seabream, by constructing a *de-novo* library using RepeatModeler2 (Flynn et al., 2020), which employs RepatScout (Price et al., 2005), RECON (Bao & Eddy, 2002), LTRHarvest (Ellinghaus et al., 2008) and Ltr_retreiver (Ou and Jiang, 2018), along with the extra LTRStruct pipeline. The software RepeatMasker v4.1.0 (Tarailo-Graovac and Chen, 2009) combined with RepBase v17.01 was used to annotate the repeat elements, based on the previously described *de-novo* library.

2.2.2 Structural and Functional annotation

After masking the repetitive elements, protein coding gene models were annotated. The initial step of gene prediction was performed with the MAKER3 pipeline (Holt and Yandell, 2011), by integrating transcriptome, homology-based and *ab-initio* gene prediction evidence. Transcripts passed to MAKER3 (est2genome=1) were prior aligned to the genome assembly using HISAT2 v2.2.0 (Kim et al., 2019) and assembled using StringTie v2.2.1 (Pertea et al., 2015). Protein sequences from *Danio rerio, Tetraodon nigroviridis, Tetraodon rubripes, Oryzias latipes, Lamirichtys crocea* and *Gasterosteus aculeatus* were downloaded from UniProtKB/Swiss-Prot (www.uniprot.org) and passed to MAKER3 as protein homology evidence (protein2genome=1). Additionally, to the MAKER3 repeat library, we also used the previous custom repeat library and the masked genome from RepeatMasker v4.1.0. We did not proceed with a second iterative round of MAKER3, since it is a time consuming and computer memory intensive software.

We further continued with a custom pipeline. We used the protein coding gene models from MAKER3 to train AUGUSTUS v3.3.3 (Stanke et al., 2006). We chose the training set for AUGUSTUS by using the AGAT toolkit (https://github.com/NBISweden/AGAT/) and by selecting:

(1) Only protein coding genes (agat_sp_separate_by_record_type.pl)

(2) Genes with AED score under 0.2 (agat_sp_filter_feature_by_attribute_value.pl)
(3) The longest isoforms of mRNAs (agat_sp_keep_longest_isoform.pl)
(4) The complete gene models (with start and stop codon) and removing the incomplete ones (agat_sp_filter_incomplete_gene_coding_models.pl)
(5) Gene models that have distance more than 500 bp from neighbouring genes, in order to train properly the intergenic regions (agat_sp_filter_by_locus_distance.pl)
(6) Removed redundant genes, by running recursive BLAST and filtering the result (agat_sp_filter_by_mrnaBlastValue.pl).

This generated training set included the selected transcripts to train AUGUSTUS for two rounds of optimisation (--optrounds=2), in order to achieve better results concerning the sensitivity and specificity.

We also generated gene hints from the output of MAKER3. We used Portcullis v1.2.0 (Mapleson et al., 2018) to generate splice junctions and with the bam2hints script (-intronsonly, --minintronlen=15) from AUGUSTUS we created species-specific intron hints. We also created species-specific exon hints by keeping only the exon features from the MAKER3 output file and spliced protein alignments by aligning the *Dicentrarchus labrax* proteome (Proteome ID: UP000279273 from UniProt) to Gilthead seabream masked genome with Exonerate v2.4.0 (Slater and Birney, 2005) in protein2genome model. Then, we run AUGUSTUS for ab-initio gene predictions in the masked genome. We loaded the proteins predicted by AUGUSTUS to the PASA v2.4.1 pipeline (Haas et al., 2008) and we updated the PASA database with the result from MAKER3. Finally, we filtered the resulting gene models for spurious gene predictions and genes that overlap with repeats and hit transposable elements (TEs).
We evaluated the putative genes using BUSCO v4.1.4 (Simão et al., 2015) against the Actinopterygii odb10 database in protein mode.

The functional annotation of the dataset was performed with the online service (http://eggnog-mapper.embl.de) of eggNOG-mapper (Huerta-Cepas et al., 2017), a tool based on fast orthology assignments using precomputed clusters and phylogenies from the eggNOG database (Huerta-Cepas et al., 2016). The tool is synchronized with the eggNOG database, ensuring that the

annotation sources and taxonomic ranges will be kept up-to-date with future eggNOG versions (Huerta-Cepas et al., 2017).

2.3 Phylogenomic analysis

For the phylogenomic analysis (Figure 2), we compared 33 teleost species (Table 2) with OrthoFinder2 v2.5.4 (Emms and Kelly, 2018). We downloaded the annotation gff3 files along with the genomes for the 33 species, we kept only the longest isoforms using the script from the AGAT toolkit (agat_sp_keep_longest_isoforms.pl) and we produced 33 proteomes, in order to cluster the longest isoform of each protein coding gene for each species. These isoform-filtered proteomes were clustered into orthogroups with OrthoFinder2 v2.5.4. In order to construct a robust species tree, we removed those clusters (orthogroups) that contained paralogous sequences. Only 1533 orthogroups containing single copy genes were retained.

The single copy genes for each orthogroup were used for multiple sequence alignment (MSA) using MAFFT v7.486 (Katoh and Daron, 2013), trimmed using trimAl v1.4.1 (Capella-Gutiérrez et al., 2009). After alignment and trimming, the one-to-one orthologous genes were concatenated into a single supergene per species. An extra step of trimming was performed after the construction of the supermatrix. Then, ModelTest-NG v0.1.6 (Darriba et al., 2020) was used for the selection of the substitution model and FastTree 2 (Price et al., 2010) for the species tree reconstruction, which uses maximum likelihood to infer phylogeny. A second run of OrthoFinder2 was performed with input the precomputed blast results from the first run and the species tree reconstructed from FastTree 2.

2.3.1 Gene tree inference

Gene trees were constructed based on the Phylogenetic Hierarchical Orthogroups (HOGs) from the second round of OrthoFinder2. HOGs are the inferred orthogroups at each hierarchical level in the species tree (https://github.com/davidemms/OrthoFinder) by analysing the rooted gene tree and are 12% - 20% more accurate than the initial orthogroups (OGs) produced by OrthoFinder2. Since OrthoFinder2 did not produce the fasta files of HOGs, we used custom python scripts to create a fasta file per HOG, each one composed of the corresponding genes. First, we used prepare_file.py to extract from N0.tsv (the file that contains the genes belonging to a single HOG

per species) a subset of 24 species closer to Gilthead seabream, those HOGs that had at least one gene present in Gilthead seabream and the HOGs with more than 80% of the species (>18 species) present. Subsequently, the script hogsToFasta.py was used to create one fasta file per HOG. The filtered HOGs were aligned using MAFFT v7.486 (Katoh and Daron, 2013). The resulted alignments were trimmed using trimAl v1.4.1 (Capella-Gutiérrez et al., 2009) and the phylogenetic inference per gene family was conducted using IQ-TREE v1.6.12 (Nguyen et al., 2015) with "-m TEST" option to infer, directly from each alignment, the best substitution model with ModelFinder (Kalyaanamoorthy et al., 2017).

2.3.2 Gene duplication analysis

Gene duplications were detected using GeneRax v2.0.2 (Morel et al., 2020), a tool that reconciles gene trees based on species tree and applies a heuristic tree search in each gene family independently. The resulting gene trees and best-fitting models from IQ-TREE, were used as starting gene trees and models, respectively, along with the corresponding trimmed alignments for input in GeneRax. Also, we used the species tree that we produced with one-to-one orthologs using FastTree 2. GeneRax was run in mpi mode with options --rec-model UndatedDL, --per-family-rates, --max-spr-radius 3.

2.3.3 Gene Family Evolution analysis

We employed the CAFE v4.2.1 (De Bie et al., 2006) pipeline to investigate the gene gain and loss of Gilthead seabream genome compared to the 24 teleost fish previously used for the gene tree reconciliations. We used the clustered HOGs derived from OrthoFinder2 and we produced an ultrametric tree with calibration.R using the phylogenetic tree produced in the phylogenomic analysis. The divergence time for four pairs of species (*Oryzias latipes* and *Takifugu rubripes, Lates calcarifer* and *Takifugu rubripes, Larimichthys crocea* and *Takifugu rubripes, Oreochromis niloticus* and *Takifugu rubripes*) was taken from TIMETREE (http://www.timetree.org/). Finally, the CAFE pipeline was run with default parameters.

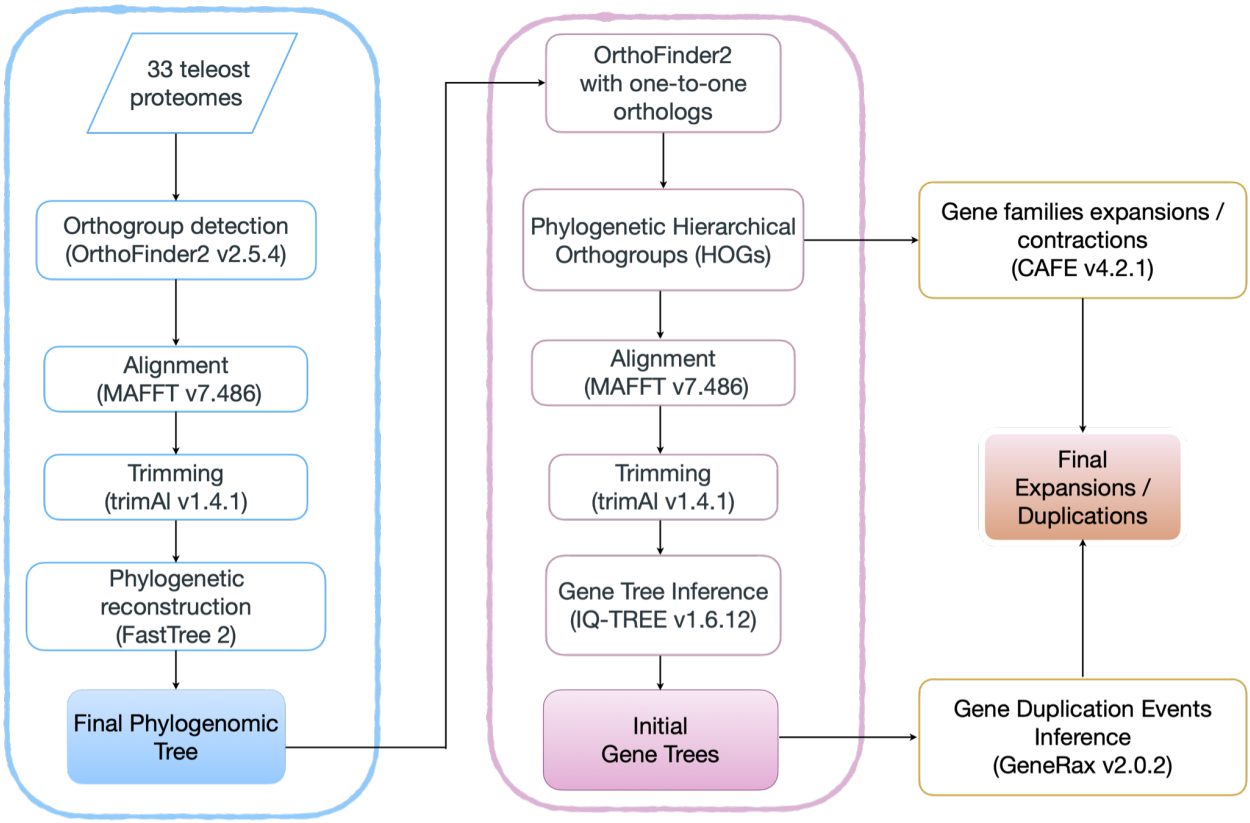

Figure 2. The complete phylogenomic analysis pipeline. The reconstructed phylogenomic tree with one-to-one orthologs was used for the final orthogroups detection. The initial gene trees were used as input for the gene tree reconciliations.

3. RESULTS

3.1 Genome assembly

Sequencing of Nanopore reads yielded 64,595.5 Mbp reads above Q7 with N50 10,442 bp and after quality filtering and trimming, 64,102.7 Mbp remained with N50 10,409 bp above Q7. The total reads of the publicly available data that were used for polishing and scaffolding are shown in Table 1.

The LSGA automated pipeline (https://github.com/genomenerds/SnakeCube) was used for constructing the primary assembly. This preliminary assembly contained 1609 contigs with total length ~883 Mbp, N50 value more than 4 Mbp with L50 52 and largest contig of ~23 Mbp. The

assessment using BUSCO v4.1.4 (Simão et al., 2015) yielded 98.7% against the Actinopterygii dataset odb10.

The workflow 1 (Figure 1) produced the assemblies saur v1 and saur v2. The assembly saur v1 yielded 615 contigs with 883,379,118 total length, N50 at 37,103,224 bp, L50 at the 12th contig and the largest contig was 44,334,261 bp. The assembly saur v2 resulted in 234 contigs with 886,345,629 total length as previously, N50 at 42,405,076 Mbp, L50 at the 8th contig and the largest contig was 77,868,326 bp. The BUSCO completeness score was 98% in both draft assemblies. We aligned the genomes with the current reference genome from VGP and used the synteny_plot.py script from https://github.com/genomenerds/Lagocephalus-sceleratus to visualise our assemblies compared to the current reference genome. The assembly saur v2, with ALLHiC scaffolder, resulted in an over-scaffolded assembly (Figure 3). One scaffold from our assembly aligns in chromosome 2 and 4 in the reference.

The workflow 2 (Figure 1) generated assembly saur v3 with 614 scaffolds, N50 at 37,709,259 bp, L50 at the 11th scaffold and the largest contig was 44,306,822 Mbp. Since workflow 2 yielded a slightly better assembly than assembly saur v1, we kept assembly saur v3 for further comparisons.

The final workflow 3 (Figure 1) produced the assemblies saur v4, v5 and v6. The assembly saur v4 comprised 1008 contigs, with 884,410,066 bp total genome length, N50 24,757,580 bp at the 15th contig. The assembly saur v5 includes 1009 scaffolds in total of 884,104,686 bp, with N50 24,359,146 bp and L50 15. We further proceed with assembly saur v5, by ordering and orienting the information in 24 chromosomes using the three linkage maps, followed by scaffolding based on synteny with the current reference genome. The final assembly (assembly saur v6) produced 451 scaffolds, spanned in 884,161,186 bp, an N50 value equal to 37,690,185 bp and L50 at the 11th scaffold. Also, the 24 chromosomes contain 98.6% of the complete genome. The completeness of the genome was performed using BUSCO v1.4.1 against the Actinopterygii odb10 dataset (3640 genes) for the final assembly. We found 97.5% (3551 out of 3640) complete genes, 0.5% (19 out of 3640) fragmented genes and 2% (70 out of 3640) genes of the dataset were missing.

All the statistics and BUSCO scores at each step of the pipeline are shown in Table 3.

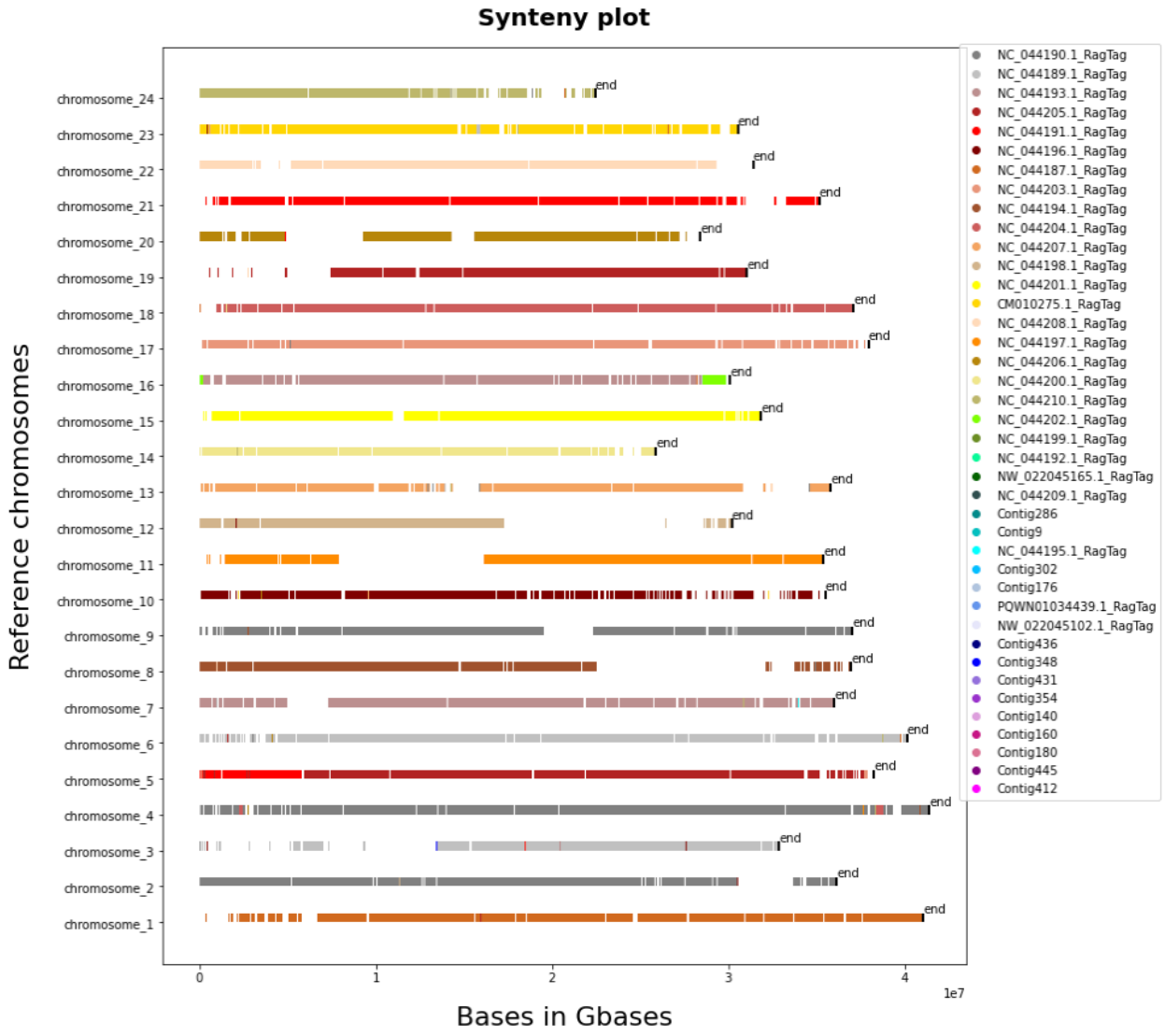

Figure 3. Assembly with ALLHiC scaffolder led to over-scaffolded assembly. The chromosomes of the reference genome are represented in the y axis and the scaffolds of our assembly are represented horizontally in the x axis. The NC_044190.1_RagTag scaffold from our assembly mapped in both chromosomes 2 and 4 of the reference genome, leading to an over-scaffolded genome assembly.

3.2 Repeat Annotation

Repetitive sequences of the genome assembly of Gilthead seabream were masked with RepeatMasker. The repeats covered 33,84% (299.2 Mbp) of the genome. The class of Retroelements and DNA transposons accounted for 5.29% and 10.27% of the complete genome, respectively (Table 4).

3.3 Gene prediction and functional annotation

The combinatorial pipeline we used for the structural annotation of the genome predicted 26,919 protein coding genes with mean gene length of 12,922 bp and mean exon length of 263 bp. From the functional annotation 23,606 genes were successfully annotated, representing around 87,7% of the complete gene set.

The gene set was assessed using BUSCO v4.1.4. From a total set of 3,640 single-copy ortholog genes from the Actinopterygii odb10 dataset, 92.3% were complete (91.1% single-copy genes, 1.2% duplicated genes), 3.2% were fragmented and 4.5% were missing.

3.4 Phylogenomic analysis

33 species were used for the comparative genomic analysis using OrthoFinder2. OrthoFinder2 assigned 756,930 of the total 777,819 genes (97.3%) in 22,958 orthogroups, with 1,533 of these to be single-copy groups. The 1,533 orthogroups with the single copy genes were used for the phylogenomic reconstruction.

ModelTest-NG identified JTT+I+G4+F as the best substitution model and FastTree 2 resulted in the species tree shown in Figure 4. The phylogenetic position of Gilthead seabream comes in agreement with previous studies (Pauletto et al., 2018, Natsidis et al., 2019, Danis, 2021).

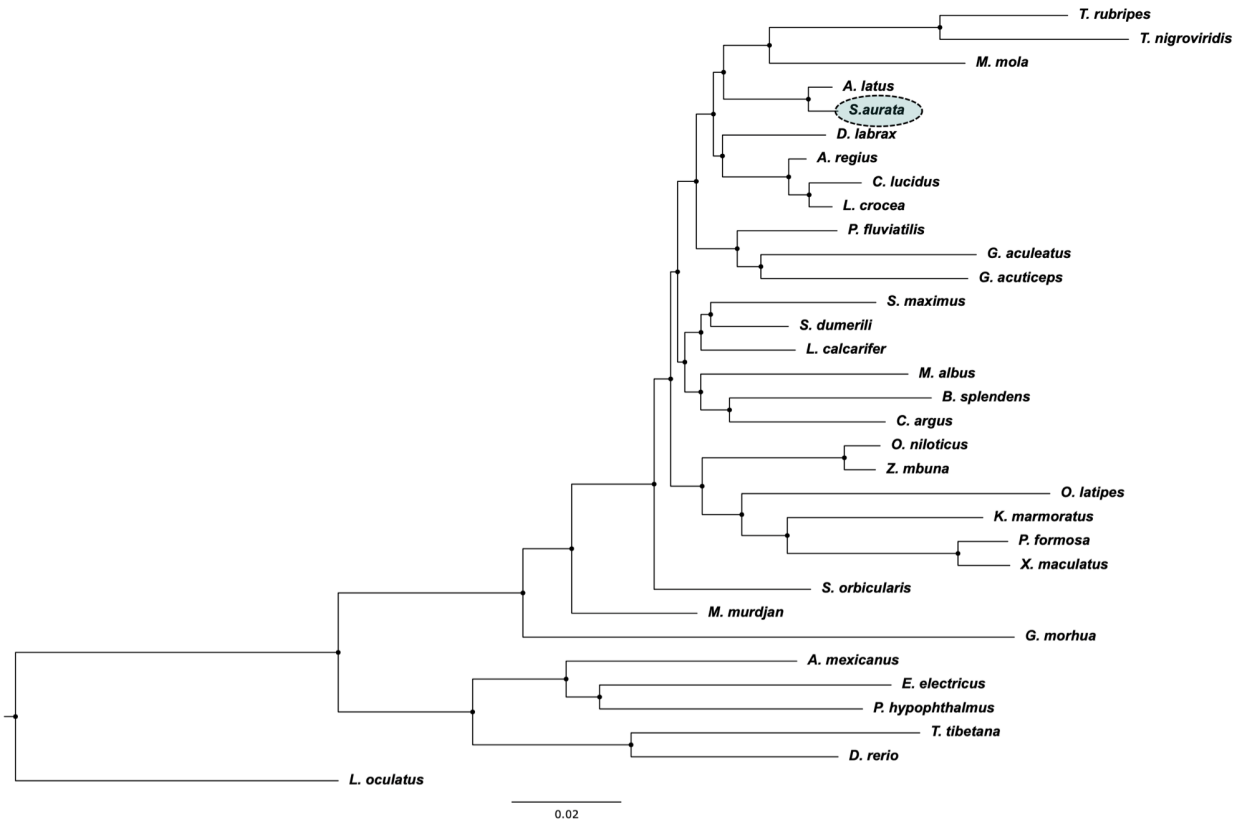

Figure 4. Maximum likelihood tree reconstruction using JTT +I+ G4+F model. The tree was visualized using FigTree v1.4.4 (http://tree.bio.ed.ac.uk/software/figtree/)

3.5 Gene duplication analyses

In order to study further the species-specific gene duplications in Gilthead seabream, we used the hierarchical orthogroups that resulted from OrthoFinder2. After filtering the HOGs by retaining only those that Gilthead seabream was present and those where more than 80% of the total species were also present, we ended up with 15,492 HOGs. These 15,492 HOGs were used for MSA with MAFFT, trimming with trimAl and gene tree inference with IQ-TREE. The gene trees along with the reconstructed species tree were used as input in GeneRax, which recovered 1,067 duplication events in 489 HOGs for our species of interest.

CAFE was used to identify expanded and contracted gene families, indicating for Gilthead seabream 625 expanded and 321 contracted gene families, while 103 and 27 of them were rapidly evolving, respectively.

We used the script intersect_gene_families.py to extract the intersected gene families that GeneRax produced as duplicated and CAFE as rapidly expanded, finding 80 gene families shared between the two different analyses. Finally, with the script violin_plots.py, we created violin plots for both GeneRax (Figure 5) and CAFE (Figure 6) results, in order to compare the gene duplications with the other species used in phylogenomic analysis, revealing that Gilthead seabream does not have more duplicated genes comparing to other teleosts.

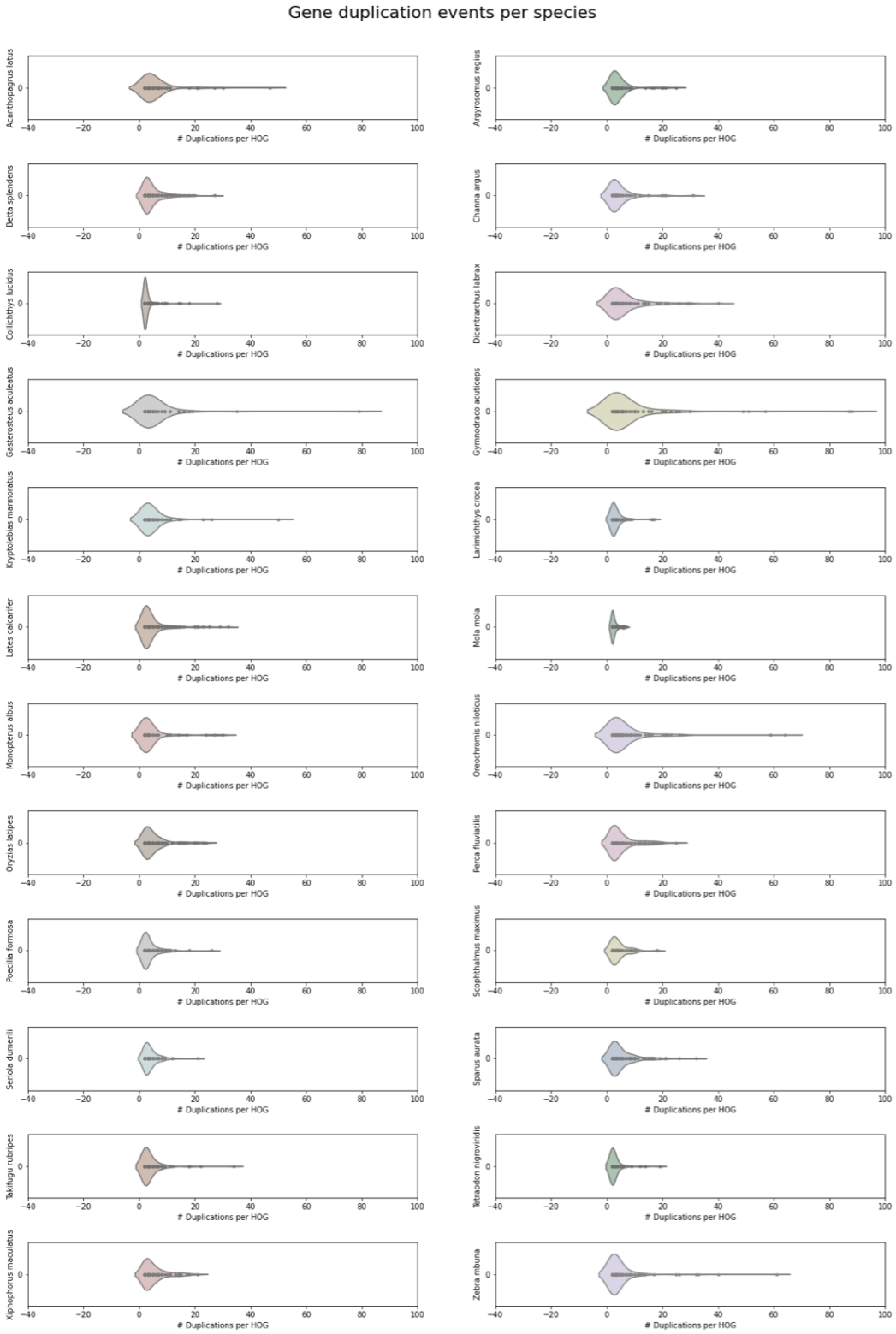

Figure 5. Violin plots for the gene duplication events, inferred from GeneRax, of the 24 species used in phylogenomic analysis.

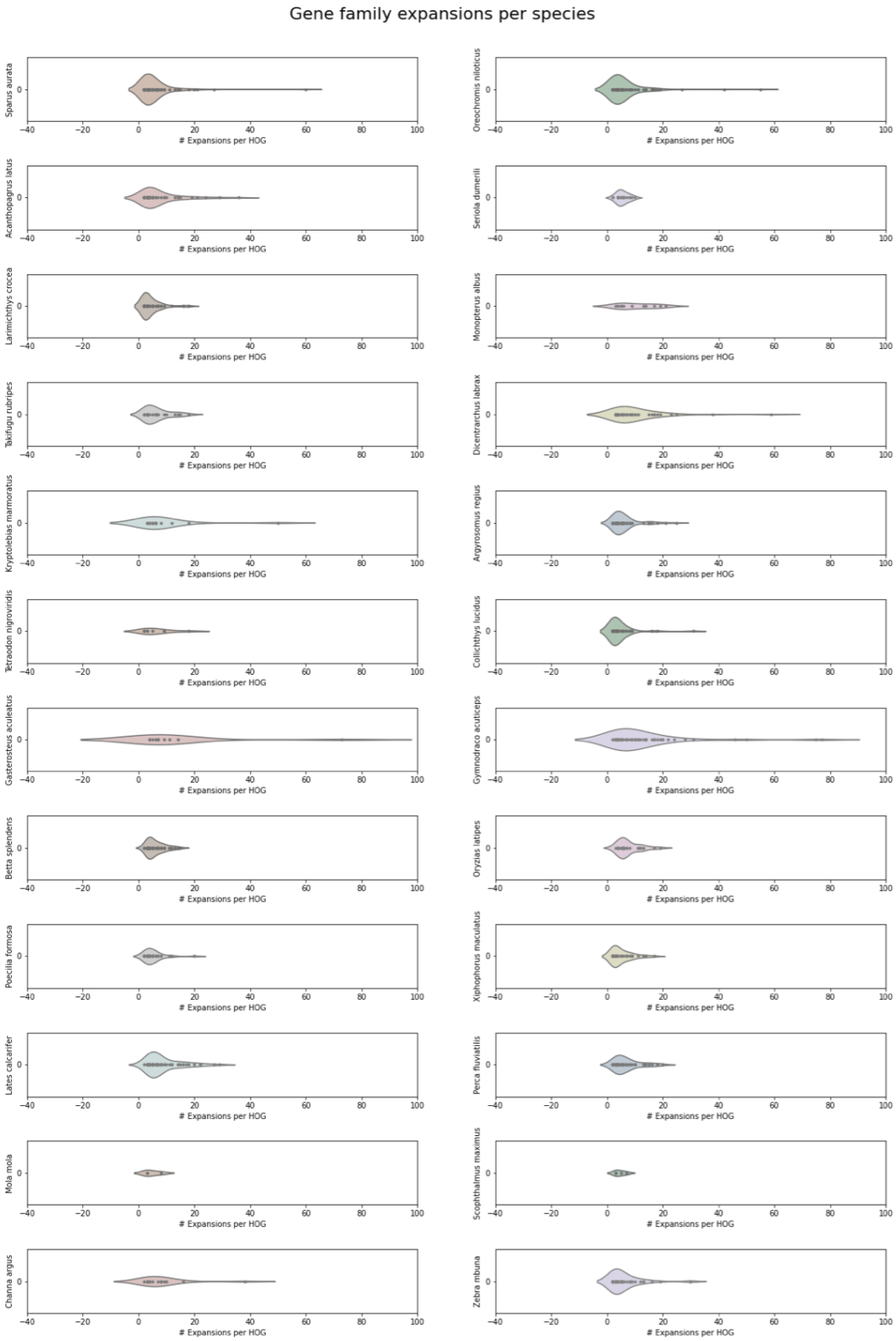

Figure 6. Violin plots for the expanded gene families of CAFE of the 24 species used in phylogenomic analysis.

4. DISCUSSION

Genome size and quality of the assembly

The complete pipeline of our assembly construction yielded ~883 Mbp genome size for Gilthead seabream, which is almost at the same levels as other studies predicted (Kuhl et al., 2011, Pauletto et. al., 2018, current reference genome assembly), while there is one study (Pérez-Sánchez et al., 2019) that the genome size estimation was ~1.59 Mbp. The Gilthead seabream assembly size is comparable to other members of Sparidae family, such as *Acantoragrus latus* (~685 Mbp; VGP), *Pagrus major* (~875 Mbp; Sawayama et al., 2017), *Diplodus sargus* (~785 Mbp*;* Fietz et al., 2020*), Spondyliosoma cantharus* (~767 Mbp; Malmstrøm et al., 2017).

Compared to the two previous efforts of genome assembly construction (GCA_003309015.1, GCA_900880675.2) our genome contains more information. VGP has produced a high quality chromosome-level reference genome assembled in 175 scaffolds above 10kbp, spanned in ~833 Mbp with N50 and L50 values equal to ~35.8 Mbp and 11, respectively. Our genome assembly strategy resulted in 497 scaffolds in total, with 219 over 10 Kbp. However, the total genome length is 50Mbp more (~882Mbp in scaffolds more than 10 Kbp), compared to the current reference. The number of N's per 100kbp is 41.06 in the VGP genome and 24.65 in our genome. Another statistic we calculated is the percent of the genome contained in 24 chromosomes. In the VGP genome 98.19% of the assembly is represented in 24 chromosomes, while 98.6% of our assembly is represented in 24 chromosomes. The assessment using BUSCO, indicated the same score at 98%. We conclude that our assembly is of higher quality and contiguity, potentially containing more information for the Gilthead seabream genome, when compared to the current reference genome.

Repeat Content

In total, the repetitive sequences of Gilthead seabream cover 33.84% of the genome. Similar studies in other teleost species of the Sparidae family, order Perciformes, indicate a repeat content almost in the same level as our species of interest. Red seabream (*Pagrus major*) genome consists of 31.1% repeats (Shin et al., 2018), while Yellowfin seabream (*Acanthopagrus latus*) total repeats

are at a lower, but still a high level, at 21.24% of the genome (Gao et al., 2021). Previous study has shown the positive correlation between the genome size and the repeat content (Yuan et al., 2018), including the order Perciformes. Our species seems to apply to this positive correlation, so we could infer that the genome size may be analogous to the repetitive elements.

Gene duplications

Since the cost of sequencing technologies has been reduced, there are many initiatives worldwide, e.g. Earth Biogenome Project (https://www.earthbiogenome.org), Vertebrates Genome Project (https://vertebrategenomesproject.org), Darwin Tree of Life (https://www.darwintreeoflife.org), European Reference Genome Atlas (https://www.erga-biodiversity.eu), that are trying to generate high quality, chromosome level and error-free genome assemblies of all the lineages.

Gilthead seabream is one of the most economically important fish in aquaculture in the Mediterranean area, with global production accounting for ~186,000 metric tonnes by 2016 (http://www.fao.org/fishery/culturedspecies/Sparus_aurata/en). Hence, it is a well studied species. Due to the lack of a reference genome until 2018 (Pauletto et al., 2018) the studies considering paralogs and duplicated genes in the Gilthead seabream genome, had been restricted in specific gene families (Sunyer et al., 1997; Angotzi et al., 2020) or in a few genes (Tan and Du, 2002). Moreover, Natsidis et al., 2019 has shown through BUSCO scores that Gilthead seabream has the higher percentage of duplications among 31 other teleosts and Pauletto et al. 2018 has shown that there are male-biased genes that are duplicated. To our knowledge, there is only one study, supporting whole-genome level gene duplication rate (Pérez-Sánchez et al., 2019).

Through our high quality and contiguous genome assembly, we managed to annotate the genes that are present in Gilthead seabream and identify some expanded gene families. The highest levels of duplications were found in gene families composed of C-type lectin genes, which are known to be specialized in pathogen recognition (Gambi et al., 2005), as a part of the innate immune system (Mayer et al., 2017).

Moreover, lectins, in some cases, can cause the activation of the complement system (Fujita, 2002; Forn-Cuni et al., 2014), in which we also found expansions in the third component (C3) gene

family. Multiple studies have previously reported duplications in this specific gene family (Najafpour et al., 2020; Forn-Cuni et al., 2014) across teleosts, including Gilthead seabream (Najafpour et al., 2020; Sunyer et al., 1997). Forn-Cuni et al., 2014 proposed that two paralogs of C3 gene occurred in the base of the teleost fish clade, during TSGD. Although, Forn-Cuni et al. 2014, through phylogenomic analysis revealed that the number of paralogs differs among the teleost species and each C3 gene may be a result of species-specific duplication events, providing capabilities to survive and develop a robust innate immune system confronting with a wide range of pathogens. Along with these expansions, in our resulting dataset of inferred gene duplications, the regulatory factor of C3 complement, the factor H (CFH) gene family, seems to be expanded. A large expansion of CFH genes has been reported previously in ray-finned fish, suggesting that they may fulfill the gap of the less developed adaptive immune system (Najafpour et al., 2020).

One more putative gene family, in which we found gene duplications and expansions, is the CC chemokines, which are considered the largest subfamily of chemokines. Chemokines are key regulators to immune responses and are considered to be a connector between the innate and adaptive responses (Cuesta et al., 2010). CC chemokines are expanded in teleosts, indicating gene duplications, which are likely to be the result of lineage-specific tandem duplications depending on the aquatic environment and the emerging immune responses (Peatman and Liu, 2007; Fu et al., 2017). Specifically in Gilthead seabream, six CC chemokines have been reported previously, suggesting an important role of these CC chemokines during a viral infection (Cuesta et al., 2010).

Furthermore, the major histocompatibility complex (MHC) class I gene family was detected as expanded in Gilthead seabream. MHC genes, a part of the adaptive immune system, act as initiators of the immune responses, when an intracellular pathogen is detected (Grimholt et al., 2015).

Most gene families that our analysis reported with gene duplications were related to immune responses. As an euryhaline and eurythermal species, the Gilthead seabream is capable of surviving to new habitats and the challenging farming environments. In addition, the wide range of duplicated immune genes that we have found, increase and expand its ability of migrating and conquering successfully among all these habitats, dealing with a variety of unknown pathogens. At the end, we suggest according to the comparison among other teleosts (Fig. 5 and 6) that the

Gilthead seabream follows a typical, teleost gene duplication pattern, which pattern is also supported by another study (Lu et al., 2012).

5. CONCLUSIONS

Gilthead seabream represents a broadly farmed species in the aquaculture industry. It is cultured intensively, thus, a high-quality, contiguous and chromosome-level genome assembly is of great importance for additional studies. In this project, we achieved chromosome-level genome assembly, provided the structural and functional annotations of the genome and shed light into the gene duplications evolution. All the analyses that were performed here, suggested that Gilthead seabream has some expanded/duplicated gene families, mostly related to immune responses and typical levels of duplications with other teleosts across the whole genome, that maybe are highly correlated with its dynamic behavior on changing and conquering different environments.

CODE AVAILABILITY

All the custom scripts that were used in this project are available at the below GitHub repository: https://github.com/genomenerds/Sparus-aurata

Table 1. Raw reads used for the genome and transcriptome assembly construction.

Sequencing Technology	Total Raw Reads	Source	SRA accession number
Oxford Nanopore	7,337,917	Current study	-
Illumina	2,068,689,358	Pauletto et al., (2018) Pérez-Sánchez et al., (2019)	SRR6244977 - SRR6244979 SRR9615483 - SRR9615484 SRR9615489 - SRR9615496 SRR9615501 - SRR9615504 SRR9615517 - SRR9615520 SRR9615522 - SRR9615526
Pacific Biosciences	1,020,739	Pérez-Sánchez et al., (2019)	SRR9615481 - SRR9615482 SRR9615485 - SRR9615488 SRR9615497 - SRR9615498 SRR9615505 - SRR9615506 SRR9615509 - SRR9615510
Mate-Pair	384,154,816	Pérez-Sánchez et al., (2019)	SRR9615497 - SRR9615498
HiC	668,625,500	Vertebrates Genome Project (VGP)	ERR4179369 - ERR4179373
Linkage Maps	40,559 markers	Pauletto et al., (2018)	-
RNA-seq	1,088,167,416	Pauletto et al., (2018)	SRR6223527 - SRR6223532 SRR6223535 - SRR6223542 SRR6237494 - SRR6237500

Table 2. Species included in the phylogenomic analysis.

Species	# of Proteins (longest Isoforms)
A. latus	23,914
A. regius	24,443
A. mexicanus	26,698
B. splendens	23,126
C. argus	22,568
C. lucidus	28,508
D. rerio	25,208
D. labrax	23,380
E. electricus	22,430
G. morhua	23,513
G. aculeatus	20,779
G. acuticeps	25,028
K. marmoratus	22,228
L. crocea	22,925
L. calcarifer	25,109
L. oculatus	18,339
M. mola	21,404
M. albus	22,143
M. murdjan	23,659
O. niloticus	28,186
O. latipes	23,620
P. hypophthalmus	21,245
P. fluviatilis	24,324
P. formosa	23,615
S. maximus	21,737

S. dumerili	23,276
S. aurata	26,919
S. orbicularis	24,339
T. rubripes	21,411
T. nigroviridis	19,600
T. tibetana	24,310
X. maculatus	23,772
Z. mbuna	26,063

Table 3. *De-novo* genome assembly pipeline and current reference genome statistics.

Assembly version: Primary assembly

Software	Flye	Racon/Medaka	Pilon 1st round	Pilon 2nd round
# contigs (>= 1000 bp)	1490	1555	1555	1555
# contigs (>= 5000 bp)	1349	1388	1386	1385
# contigs (>= 10000 bp)	1082	1115	1114	1115
# contigs (>= 25000 bp)	898	907	907	907
# contigs (>= 50000 bp)	769	777	774	773
Total length (>= 0 bp)	887505396	886196235	883536162	883244168
N50	4550458	4421790	4399725	4397770
L50	45	52	52	52
Largest Contig	23560554	23316179	23298514	23296925
# N's per 100 kbp	0.66	0.00	0.00	0.00
BUSCO	91.8%	97.9%	98.6%	98.7%

Assembly version: Assembly saur v1

Software	SSPACE	SALSA	PBJelly	RagTag (VGP)
# contigs (>= 1000 bp)	1505	1091	1035	615
# contigs (>= 5000 bp)	1347	934	885	485
# contigs (>= 10000 bp)	1088	688	658	329
# contigs (>= 25000 bp)	886	504	465	197
# contigs (>= 50000 bp)	755	410	375	141
Total length (>= 0 bp)	883246223	881592705	883336518	883379118
N50	4397770	25933391	25981191	37103224
L50	51	15	15	12
Largest Contig	23296925	35095099	35108077	44334261
# N's per 100 kbp	0.21	25.11	11.12	15.94
BUSCO	98.1%	98.1%	98%	98.1%

Assembly version: Assembly saur v2

Software	SSPACE	ALLHiC	PBJelly	RagTag (VGP)
# contigs (>= 1000 bp)	1505	404	403	234
# contigs (>= 5000 bp)	1347	296	300	152

# contigs (>= 10000 bp)	1088	190	197	98
# contigs (>= 25000 bp)	886	120	121	65
# contigs (>= 50000 bp)	755	78	77	47
Total length (>= 0 bp)	883246223	883356623	886328029	886345629
N50	4397770	37689299	37847581	42405076
L50	51	11	11	8
Largest Contig	23296925	45618088	45715280	77868326
# N's per 100 kbp	0.21	12.71	10.69	12.67
BUSCO	98.1%	98.3%	98.2%	98%

Assembly version: Assembly saur v3

Assembly	**SALSA**	**PBJelly**	**RagTag VGP**
# contigs (>= 1000 bp)	1129	1059	614
# contigs (>= 5000 bp)	959	896	475
# contigs (>= 10000 bp)	699	650	306
# contigs (>= 25000 bp)	509	456	185
# contigs (>= 50000 bp)	415	369	134
Total length (>= 0 bp)	881594689	883151188	883196288
N50	24697536	26107799	37709259
L50	16	15	11
Largest Contig	35086880	35096103	44306822
# N's per 100 kbp	25.35	10.74	15.85
BUSCO	98.2%	98.2%	98.2%

Assembly version: Assembly saur v4

Assembly	**PBJelly**	**SALSA**	**SSPACE**
# contigs (>= 1000 bp)	1319	1013	1008
# contigs (>= 5000 bp)	1165	861	858
# contigs (>= 10000 bp)	916	616	615
# contigs (>= 25000 bp)	715	425	425
# contigs (>= 50000 bp)	604	338	338
Total length (>= 0 bp)	883943163	884107663	884108066
N50	5301358	24757580	24757580
L50	45	15	15
Largest Contig	23533739	38026669	38026669

# N's per 100 kbp	0.00	18.61	18.65
BUSCO	98.1%	98.2%	98.2%

Assembly version: Assembly saur v5

Assembly	PBJelly	SSPACE	SALSA
# contigs (>= 1000 bp)	1319	1307	1009
# contigs (>= 5000 bp)	1165	1155	860
# contigs (>= 10000 bp)	916	906	614
# contigs (>= 25000 bp)	715	709	425
# contigs (>= 50000 bp)	604	600	338
Total length (>= 0 bp)	883943163	883943686	884104686
N50	5301358	5301358	24359146
L50	45	44	15
Largest Contig	23533739	23533739	38858959
# N's per 100 kbp	0.00	0.05	18.26
BUSCO	98.1%	98.1%	98.1%

Assembly version: Assembly saur v6 (Final assembly)

Assembly	Assembly saur	v5 + ALLMAPS	RagTag VGP	Reference Genome (GCA_900880675.2)
# contigs (>= 1000 bp)		874	451	175
# contigs (>= 5000 bp)		726	333	175
# contigs (>= 10000 bp)		484	217	175
# contigs (>= 25000 bp)		300	133	160
# contigs (>= 50000 bp)		220	96	103
Total length (>= 0 bp)		884118186	884161186	833595063
N50		36434643	37690185	35791275
L50		12	11	11
Largest Contig		42506557	45126894	41392777
# N's per 100 kbp		19.79	24.65	41.06
BUSCO		98%	98%	98%

Table 4. Repeat annotation statistics.

Repetitive elements	Number of elements	Length occupied	Percentage of sequence
Retroelements	188736	46753174 bp	5.29 %
SINEs:	8078	1036074 bp	0.12 %
Penelope:	8365	1838381 bp	0.21 %
LINEs:	97471	31439326 bp	3.56 %
CRE/SLACS	0	0 bp	0.00 %
L2/CR1/Rex	51937	16290147 bp	1.84 %
R1/LOA/Jockey	6839	1249311 bp	0.14 %
R2/R4/NeSL	667	582226 bp	0.07 %
RTE/Bov-B	8533	5048244 bp	0.57 %
L1/CIN4	15124	3834500 bp	0.43 %
LTR elements:	83187	14277774 bp	1.61 %
BEL/Pao	1685	937275 bp	0.11 %
Ty1/Copia	0	0 bp	0.00 %
Gypsy/DIRS1	16568	5684875 bp	0.64 %
Retroviral	59810	6387435 bp	0.72 %
DNA transposons	494141	90802967 bp	10.27 %
hobo-Activator	219662	41910888 bp	4.74 %
Tc1-IS630-Pogo	38584	9280829 bp	1.05 %
En-Spm	0	0 bp	0.00 %
MuDR-IS905	0	0 bp	0.00 %
PiggyBac	12605	2258521 bp	0.26 %
Tourist/Harbinger	25776	6710963 bp	0.76 %
Other (Mirage, P-element,Transib)	3108	513322 bp	0.06 %
Rolling-circles	16195	6714146 bp	0.76 %
Unclassified:	706798	153793868 bp	17.39 %

Total interspersed repeats:		291350009 bp	32.95 %
Small RNA:	0	0 bp	0.00 %
Satellites:	15883	1174004 bp	0.13 %
Simple repeats:	0	0 bp	0.00 %
Low complexity:	0	0 bp	0.00 %

REFERENCES

Alonge M: **Ragtag: Reference-guided genome assembly correction and scaffolding.** GitHub archive. 2020.

Andrews, Simon. "FastQC: a quality control tool for high throughput sequence data. 2010." (2017): W29-33.

Angotzi, A. R., Puchol, S., Cerdá-Reverter, J. M., & Morais, S. (2020). Insights into the Function and Evolution of Taste 1 Receptor Gene Family in the Carnivore Fish Gilthead Seabream (Sparus aurata). *International journal of molecular sciences*, *21*(20), 7732.

Bao, Z., & Eddy, S. R. (2002). Automated de novo identification of repeat sequence families in sequenced genomes. *Genome research*, *12*(8), 1269-1276.

Berthelot, C., Brunet, F., Chalopin, D., Juanchich, A., Bernard, M., Noël, B., et al. (2014). The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates. *Nature communications*, *5*(1), 1-10.

Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D., & Pirovano, W. (2011). Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics*, *27*(4), 578-579.

Bolger, Anthony M., Marc Lohse, and Bjoern Usadel. "Trimmomatic: a flexible trimmer for Illumina sequence data." *Bioinformatics* 30.15 (2014): 2114-2120.

Brawand, D., Wagner, C. E., Li, Y. I., Malinsky, M., Keller, I., Fan, S., et al. (2014). The genomic substrate for adaptive radiation in African cichlid fish. *Nature*, *513*(7518), 375-381.

Cambi, A., Koopman, M., & Figdor, C. G. (2005). How C-type lectins detect pathogens. *Cellular microbiology*, *7*(4), 481-488.

Capella-Gutiérrez, S., Silla-Martínez, J. M., & Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, *25*(15), 1972-1973.

Chaisson, M. J., & Tesler, G. (2012). Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC bioinformatics*, *13*(1), 1-18.

Chen, N. (2004). Using Repeat Masker to identify repetitive elements in genomic sequences. *Current protocols in bioinformatics*, *5*(1), 4-10.

Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, *34*(17), i884-i890.

Cuesta, A., Dios, S., Figueras, A., Novoa, B., Esteban, M. A., Meseguer, J., & Tafalla, C. (2010). Identification of six novel CC chemokines in gilthead seabream (Sparus aurata) implicated in the antiviral immune response. *Molecular immunology*, *47*(6), 1235-1243

Danis, T. (2021) Genome analysis of *Lagocephalus sceleratus*: unraveling the genomic aspects of a successful invader [Master's Thesis, University of Crete] https://elocus.lib.uoc.gr//dlib/4/f/7/attached-metadata-dlib-1625476411-502612-31334/Danhs.pdf

Darriba, D., Posada, D., Kozlov, A. M., Stamatakis, A., Morel, B., & Flouri, T. (2020). ModelTest-NG: a new and scalable tool for the selection of DNA and protein evolutionary models. *Molecular Biology and Evolution*, *37*(1), 291-294.

De Bie, T., Cristianini, N., Demuth, J. P., & Hahn, M. W. (2006). CAFE: a computational tool for the study of gene family evolution. *Bioinformatics*, *22*(10), 1269-1271.

Ellinghaus, D., Kurtz, S., & Willhoeft, U. (2008). LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC bioinformatics*, *9*(1), 1-14.

Emms, D. M., & Kelly, S. (2018). OrthoFinder2: fast and accurate phylogenomic orthology analysis from gene sequences. *BioRxiv*, 466201.

English, A. C., Richards, S., Han, Y., Wang, M., Vee, V., Qu, J., et al. (2012). Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. *PloS one*, *7*(11), e47768.

Fietz, K., Trofimenko, E., Guerin, P. E., Arnal, V., Torres-Oliva, M., Lobréaux, S., et al. (2020). New genomic resources for three exploited Mediterranean fishes. *Genomics*, *112*(6), 4297-4303.

Flynn, J. M., Hubley, R., Goubert, C., Rosen, J., Clark, A. G., Feschotte, C., & Smit, A. F. (2020). RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences*, *117*(17), 9451-9457.

Forn-Cuni, G., Reis, E. S., Dios, S., Posada, D., Lambris, J. D., Figueras, A., & Novoa, B. (2014). The evolution and appearance of C3 duplications in fish originate an exclusive teleost c3 gene form with anti-inflammatory activity. *PLoS One*, *9*(6), e99673.

Fu, Q., Yang, Y., Li, C., Zeng, Q., Zhou, T., Li, N., et al. (2017). The chemokinome superfamily: II. The 64 CC chemokines in channel catfish and their involvement in disease and hypoxia responses. *Developmental & Comparative Immunology*, *73*, 97-108.

Fujita, T. (2002). Evolution of the lectin–complement pathway and its role in innate immunity.

Nature Reviews Immunology, *2*(5), 346-353.

Gao, D., Fang, W., Sims, Y., Collins, J., Torrance, J., Lin, G., et al. (2021). Chromosome-level genome assembly of Acanthopagrus latus using PacBio and Hi-C technologies. *bioRxiv*.

Ghurye, J., Rhie, A., Walenz, B. P., Schmitt, A., Selvaraj, S., Pop, M., et al. (2019). Integrating Hi-C links with assembly graphs for chromosome-scale assembly. *PLoS computational biology*, *15*(8), e1007273.

Glasauer, S. M., & Neuhauss, S. C. (2014). Whole-genome duplication in teleost fishes and its evolutionary consequences. *Molecular genetics and genomics*, *289*(6), 1045-1060.

Grimholt, U., Tsukamoto, K., Azuma, T., Leong, J., Koop, B. F., & Dijkstra, J. M. (2015). A comprehensive analysis of teleost MHC class I sequences. *BMC evolutionary biology*, *15*(1), 1-17.

Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, *29*(8), 1072-1075.

Haas, B. J., Salzberg, S. L., Zhu, W., Pertea, M., Allen, J. E., Orvis, J., et al. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome biology*, *9*(1), 1-22.

Holt, C., & Yandell, M. (2011). MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC bioinformatics*, *12*(1), 1-14.

Huerta-Cepas, J., Forslund, K., Coelho, L. P., Szklarczyk, D., Jensen, L. J., Von Mering, C., & Bork, P. (2017). Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. *Molecular biology and evolution*, *34*(8), 2115-2122.

Huerta-Cepas, J., Serra, F., & Bork, P. (2016). ETE 3: reconstruction, analysis, and visualization of phylogenomic data. *Molecular biology and evolution*, *33*(6), 1635-1638.

Jaillon, O., Aury, J. M., Brunet, F., Petit, J. L., Stange-Thomann, N., Mauceli, E., et al. (2004). Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. *Nature*, *431*(7011), 946-957.

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K., Von Haeseler, A., & Jermiin, L. S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature methods*, *14*(6), 587-589.

Katoh, Kazutaka, and Daron M. Standley. "MAFFT multiple sequence alignment software version 7: improvements in performance and usability." *Molecular biology and evolution* 30.4 (2013): 772-780.

Kim, D., Paggi, J. M., Park, C., Bennett, C., & Salzberg, S. L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature biotechnology*, *37*(8), 907-915.

Kleszczyńska, A., Vargas-Chacoff, L., Gozdowska, M., Kalamarz, H., Martínez-Rodríguez, G., Mancera, J. M., & Kulczykowska, E. (2006). Arginine vasotocin, isotocin and melatonin responses following acclimation of gilthead sea bream (Sparus aurata) to different environmental salinities. *Comparative Biochemistry and Physiology Part A: Molecular & Integrative Physiology*, *145*(2), 268-273.

Kolmogorov, Mikhail, et al. "Assembly of long, error-prone reads using repeat graphs." *Nature biotechnology* 37.5 (2019): 540-546.

Kuhl, H., Sarropoulou, E., Tine, M., Kotoulas, G., Magoulas, A., & Reinhardt, R. (2011). A comparative BAC map for the gilthead sea bream (Sparus aurata L.). *Journal of Biomedicine and Biotechnology*, *2011*.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303.3997*.

Lien, S., Koop, B. F., Sandve, S. R., Miller, J. R., Kent, M. P., Nome, T., et al. (2016). The Atlantic salmon genome provides insights into rediploidization. *Nature*, *533*(7602), 200-205.

Lu, J., Peatman, E., Tang, H., Lewis, J., & Liu, Z. (2012). Profiling of gene duplication patterns of sequenced teleost genomes: evidence for rapid lineage-specific genome expansion mediated by recent tandem duplications. *BMC genomics*, *13*(1), 1-10.

Malmstrøm, M., Matschiner, M., Tørresen, O. K., Jakobsen, K. S., & Jentoft, S. (2017). Whole genome sequencing data and de novo draft assemblies for 66 teleost species. *Scientific data*, *4*(1), 1-13.

Malmstrøm, M., Matschiner, M., Tørresen, O. K., Star, B., Snipen, L. G., Hansen, T. F., et al. (2016). Evolution of the immune system influences speciation rates in teleost fishes. *Nature Genetics*, *48*(10), 1204-1210.

Mapleson, D., Venturini, L., Kaithakottil, G., & Swarbreck, D. (2018). Efficient and accurate detection of splice junctions from RNA-seq with Portcullis. *GigaScience*, *7*(12), giy131.

Mayer, S., Raulf, M. K., & Lepenies, B. (2017). C-type lectins: their network and roles in pathogen recognition and immunity. *Histochemistry and cell biology*, *147*(2), 223-237.

Meyer, A., & Schartl, M. (1999). Gene and genome duplications in vertebrates: the one-to-four (-to-eight in fish) rule and the evolution of novel gene functions. *Current opinion in cell biology*, *11*(6), 699-704.

Meyer, A., & Van de Peer, Y. (2005). From 2R to 3R: evidence for a fish-specific genome duplication (FSGD). *Bioessays*, *27*(9), 937-945.

Morel, B., Kozlov, A. M., Stamatakis, A., & Szöllősi, G. J. (2020). GeneRax: a tool for species-tree-aware maximum likelihood-based gene family tree inference under gene duplication, transfer, and loss. *Molecular biology and evolution*, *37*(9), 2763-2774.

Mulder, S. A., & Wunsch II, D. C. (2003). Million city traveling salesman problem solution by divide and conquer clustering with adaptive resonance neural networks. *Neural Networks*, *16*(5-6), 827-832.

Mylonas, C. C., Zohar, Y., Pankhurst, N., & Kagawa, H. (2011). Reproduction and broodstock management. *Sparidae*, 95-131.

Najafpour, B., Cardoso, J. C., Canário, A. V., & Power, D. M. (2020). Specific evolution and gene family expansion of complement 3 and regulatory factor H in fish. *Frontiers in immunology*, *11*.

Natsidis, P., Tsakogiannis, A., Pavlidis, P., Tsigenopoulos, C. S., & Manousaki, T. (2019). Phylogenomics investigation of sparids (Teleostei: Spariformes) using high-quality proteomes highlights the importance of taxon sampling. *Communications biology*, *2*(1), 1-10.

Nelina Angelova, Theodoros Danis, Lagnel Jacques, Costas Tsigenopoulos, & Tereza Manousaki. (2021). SnakeCube: containerized and automated next-generation sequencing (NGS) pipelines for genome analyses in HPC environments. Zenodo. https://doi.org/10.5281/zenodo.4663112

Nguyen, L. T., Schmidt, H. A., Von Haeseler, A., & Minh, B. Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular biology and evolution*, *32*(1), 268-274.

Ou, S., & Jiang, N. (2018). LTR_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant physiology*, *176*(2), 1410-1422.

Pauletto, M., Manousaki, T., Ferraresso, S., Babbucci, M., Tsakogiannis, A., Louro, B., et al. (2018). Genomic analysis of Sparus aurata reveals the evolutionary dynamics of sex-biased genes in a sequential hermaphrodite fish. *Communications biology*, *1*(1), 1-13.

Peatman, E., & Liu, Z. (2007). Evolution of CC chemokines in teleost fish: a case study in gene duplication and implications for immune diversity. *Immunogenetics*, *59*(8), 613-623.

Pérez-Sánchez, J., Naya-Català, F., Soriano, B., Piazzon, M. C., Hafez, A., Gabaldón, T., et al. (2019). Genome sequencing and transcriptome analysis reveal recent species-specific gene duplications in the plastic gilthead sea bream (Sparus aurata). *Frontiers in Marine Science*, *6*, 760.

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T. C., Mendell, J. T., & Salzberg, S. L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature biotechnology*, *33*(3), 290-295.

Price, A. L., Jones, N. C., & Pevzner, P. A. (2005). De novo identification of repeat families in large genomes. *Bioinformatics*, *21*(suppl_1), i351-i358.

Price, M. N., Dehal, P. S., & Arkin, A. P. (2010). FastTree 2–approximately maximum-likelihood trees for large alignments. *PloS one*, *5*(3), e9490.

Sawayama, E., Tanizawa, S., Kitamura, S. I., Nakayama, K., Ohta, K., Ozaki, A., & Takagi, M. (2017). Identification of quantitative trait loci for resistance to RSIVD in red sea bream (Pagrus major). *Marine Biotechnology*, *19*(6), 601-613.

Shin, G. H., Shin, Y., Jung, M., Hong, J. M., Lee, S., Subramaniyam, S., et al. (2018). First draft genome for Red Sea bream of family Sparidae. *Frontiers in genetics*, *9*, 643.

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, *31*(19), 3210-3212.

Slater, G. S. C., & Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. *BMC bioinformatics*, *6*(1), 1-11.

Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S., & Morgenstern, B. (2006). AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic acids research*, *34*(suppl_2), W435-W439.

Sunyer, J. O., Tort, L., & Lambris, J. D. (1997). Structural C3 diversity in fish: characterization of five forms of C3 in the diploid fish Sparus aurata. *The Journal of Immunology*, *158*(6), 2813-2821.

Sunyer, J. O., Tort, L., & Lambris, J. D. (1997). Structural C3 diversity in fish: characterization of five forms of C3 in the diploid fish Sparus aurata. *The Journal of Immunology*, *158*(6), 2813-2821.

Tan, X., & Du, S. (2002). Differential expression of two MyoD genes in fast and slow muscles of gilthead seabream (Sparus aurata). *Development genes and evolution*, *212*(5), 207-217.

Tang, H., Zhang, X., Miao, C., Zhang, J., Ming, R., Schnable, J. C., et al. (2015). ALLMAPS: robust scaffold ordering based on multiple maps. *Genome biology*, *16*(1), 1-15.

Vaser, Robert, et al. "Fast and accurate de novo genome assembly from long uncorrected reads." *Genome research* 27.5 (2017): 737-746.

Walker, Bruce J., et al. "Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement." *PloS one* 9.11 (2014): e112963.

Warren, R. L., Sutton, G. G., Jones, S. J., & Holt, R. A. (2007). Assembling millions of short DNA sequences using SSAKE. *Bioinformatics*, *23*(4), 500-501.

Yuan, Z., Liu, S., Zhou, T., Tian, C., Bao, L., Dunham, R., & Liu, Z. (2018). Comparative genome analysis of 52 fish species suggests differential associations of repetitive elements with their living aquatic environments. *BMC genomics*, *19*(1), 1-10.

Zafeiropoulos, H., Gioti, A., Ninidakis, S., Potirakis, A., Paragkamian, S., Angelova, N., et al. (2021). 0s and 1s in marine molecular research: a regional HPC perspective. *GigaScience*, *10*(8), giab053.

Zhang, X., Zhang, S., Zhao, Q., Ming, R., & Tang, H. (2019). Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data. *Nature plants*, *5*(8), 833-845.