SHORT-TERM CONTRACT FOR BIOLOGICAL STUDIES
(ICCAT GBYP 08/2017-2) OF THE ATLANTIC-WIDE
RESEARCH PROGRAMME ON BLUEFIN TUNA
(GBYP Phase 7)

DELIVERABLE #4

Consortium **"Next-Generation BFT"**
composed by
Alma Mater Studiorum - University of Bologna
Dept. Biological, Geological and Environmental Sciences
Bologna, Italy (UNIBO-BiGeA)
&
Centro di Competenza sulla Biodiversità Marina
University of Cagliari
Cagliari, Italy (Com.Bio.Ma)

Bologna, 8th February 2018

INDEX

EXECUTIVE SUMMARY

The Partnership “NextGeneration BFT” is established within University of Bologna, Dept. Biological, Geological and Environmental Sciences, Bologna, Italy (UNIBO-BiGeA) and the Centro di Competenza sulla Biodiversità Marina, University of Cagliari, Cagliari, Italy (Com.Bio.Ma) to accomplish the activities awarded by the Call for Tenders ICCAT GBYP 08/2017 - BIOLOGICAL STUDIES and specifically for services concerning Sampling (b) and Genetic Analyses (d) as requested by the Call.

One of the main objective the ICCAT GBYP is to improve and enhance the understanding of key biological and ecological processes for Atlantic bluefin tuna. As part of this Programme the activities awarded by the Call for Tenders ICCAT GBYP 08/2017 - BIOLOGICAL STUDIES to the Partnership “NextGeneration BFT” contributed to both the biological sampling of the main bluefin fisheries and to future research and conservation efforts by providing a new and annotated genomic reference, that constitute a novel resource useful for the scientific community.

Developing genomic resources for Atlantic Bluefin tuna has been of crucial importance, not only for our target species, but also for other tuna species that are threatened by over-exploitation. Annotation and comparison of our reference against other tuna genomes will allow better functional characterisation of tuna genomes. Moreover, our genomic reference is of use for the mapping and positioning of current and future genomic markers, allowing comparison between markers dataset and analyses and the validation of population structure results.

During Phase 7, following sampling protocols agreed in earlier Phases, the Partnership “NextGeneration BFT” sampled a total of 248 bluefin tuna (85 young-of-the-year, 28 juveniles, and 135 medium sized fish) from different subareas of the Central Mediterranean (198) and Western Mediterranean (50). The sampling of juveniles was only partially accomplished because the scarcity of commercial catches of this stage. On the contrary, the great abundance of young-of-the-year individuals registered in the East Sicily and Ionian area led the opportunity to collect a high number of individuals at this stage (YOY). According to the mail correspondence occurred with the previous GBYP Coordinator in December 2018 and because the scarcity of YOY collected from other strata and other consortia, we added to the sampling strata the 12d Central Mediterranean - Sicily (East Sicily and Ionian Sea) – YOY. A number of YOYs was collected up to fill the amount of financial resources contracted with ICCAT GBYP.

A high-quality BFT genome assembly has been generated capitalizing on available genomic data previously generated by the research project “Consortium BFTbySNP”, by GBYP-phase2, and by the UNIBO-founded project FARB Assessing Temporal Genome Changes in the bluefin TUNA (ATGC-

TUNA). The resulting genome assembly reported very good contiguity statistics as well as a good level of completeness.

Genome-wide annotation of protein-coding genes has been performed using a BFT-specific gene statistical model, implemented ad hoc for this task. Overall, 41,508 protein-coding genes were identified.

The quality of the current gene annotation was enhanced by incorporating transcriptomic data from different sources including cDNA from 10 different individuals generated by the research project "Consortium BFTbySNP" as well as BTF specific data retrieved from the NCBI public database. These data were incorporated into the gene prediction pipeline to guide and support the identification of candidate gene.

The resulting gene annotations have been assessed by comparison of predicted protein sequence with proteins from other species, showing a high rate of similarity with those of other fishes (97% of predicted proteins mapped to Teleostei Uniref90 reference clusters), supporting the good quality of our gene annotations.

All the 41,508 predicted BFT proteins were subjected to functional annotation and 63% of the candidate sequences (26,151 protein) were associated to functions assigned by accurate homology-based approaches according to the standard catalogue of Gene Ontology (GO), covering, with different proportions, the three ontology aspects: biological process, molecular function and cellular component (with a total of 13,915 different GO terms). Moreover, sequence analysis tools were adopted to complement functional annotation with protein features (secretory signal peptides, mitochondrial-targeting peptides and/or transmembrane domains) and annotations of GO cellular component terms.

The presence of candidate genes and markers for sex-related traits was investigated in the BFT genome by searching for sequence similarity with candidate sex-determining genes characterized in previous studies in T. orientalis and other bony fishs (as zebrafish, cod, medaka, Patagonian pejerrey, fugu, rainbow trout, turbot, Yellowtail). Only 3 out of the 35 candidate genes and markers did not find a match on the assembled BFT genome. All other sequences were located each in different scaffolds, since the candidate sequences determining sexual traits are widely spread this result does not support the identification of a sex-determining region in the BFT genome. The identified sequences represent a baseline for further analysis targeting differential expression of those candidate regions in BFT individuals of known sex (not available in this study).

DETAILED REPORT OF AWARDED ACTIVITIES

b) Sampling Task b1 – Sampling of strata 12d, 13c, 15/15c

According to the awarded activities, the Partnership "NextGeneration BFT" was in charge of the sampling of biological samples and otoliths from fish of several strata of juvenile and medium-sized BFT from subareas of the Central Mediterranean and Western Mediterranean, according to the Table 1 reported below.

Table 1. Summary of the sampling strata awarded to the Partnership "NextGeneration BFT"

ICCAT GBYP BIOLOGICAL SAMPLING STRATA TABLE						**SAMPLING AWARDED TO NEXT GENERATION BFT Number of specimens**				
SAMPLING STRATUM CODE	SAMPLING AREA	AREA CODE	SUBAREA	SAMPLING SUBAREA CODE	GEAR CODE	YOY (AGE 0)	JUVENILES	MEDIUM	LARGE	**TOTAL**
12d	Central Mediterranean	CMED	Sicily (East Sicily and Ionian Sea)	ESI	LL		50	50		**100**
13c	Central Mediterranean	CMED	Adriatic Sea	AS	UNCL			50		**50**
15/15c	Western Mediterranean	WMED	Tyrrhenian Sea (South)	TY	HL/LL		50	50		**100**
					TOTAL		**100**	**150**		**250**

Sampling accomplished

Scientists and technicians of the UNIBO-BiGeA and Com.Bio.Ma. have established agreements with local authorities, local fishermen and industries to carry out the needed sampling activities within national jurisdictions. All sampling materials and labelled tubes have been provided to "NextGeneration BFT" partners and collaborators right after the communication of the assigned sampling strata, i.e. in the second week of July 2017.

Labelling of samples was performed in accordance with the GBYP DB requirements. We decided to start a new numbering of samples. The reason was that these sampling activities are performed by the consortium "NextGeneration BFT" (NGBFT), with respect to the single institutes as UNIBO-BiGeA and Com.Bio.Ma that were partners of a different consortium in the previous GBYP Phases. Example of labelling is "NGBFT-Area-Stage-#Specimen-Tissue type/replicate", i.e. "NGBFT-ESI-J-01/Ma". Such labelling has been communicated to the manager of the GBYP BFT database in due time.

Sampling was carried out from August 2017 to January 2018 (Table 2).

Table 2. Summary of the sampling periods according to strata.

ICCAT GBYP BIOLOGICAL SAMPLING STRATA						SAMPLING PERIOD BY 31/01/2018 (starting date - ending date)		
SAMPLING STRATUM CODE	SAMPLING AREA	AREA CODE	SUBAREA	SAMPLING SUBAREA CODE	GEAR CODE	YOY (AGE 0)	JUVENILES	MEDIUM
12d	Central Mediterranean	CMED	Sicily (East Sicily and Ionian Sea)	ESI	LL	08/12/2017 - 31/01/2018	12/11/2017 - 08/12/2017	27/10/2017 - 26/11/2017
13c	Central Mediterranean	CMED	Adriatic Sea	AS	UNCL			07/08/2017 - 04/11/2017
15/15c	Western Mediterranean	WMED	Tyrrhenian Sea (South)	TY	HL/LL		07/07/2017 - 15/10/2017	03/10/2017 - 06/01/2018

Overall, 248 BFT individual samples were collected according to sampling protocols (Table 3). Since the expected sample size of the BFT juvenile strata (12d-J: N = 50 and 15_15c-J: N = 50) was not reached by the effective sampling (12d-J: N = 13 and 15_15c-J: N = 15), we targeted the abundant YOY in the areas. This change of sampling strategy was agreed upon discussion with the GBYP Coordinator by e-mailing. Sampling effort on YOY allowed the collection of 85 individuals from the 12d - East Sicily and Ionian Sea that were summed up to 28 Juveniles (strata 12d and 15_15c) and 135 Medium BFTs (strata 12d, 13c and 15_15c). Total and individual data have been submitted to GBYP Coordinator and DataBase Manager.

Table 3. Summary of the number of BFT individuals sampled according to strata.

ICCAT GBYP BIOLOGICAL SAMPLING STRATA						SAMPLING ACCOMPLISHED BY 31/01/2018 - Number of specimens			
SAMPLING STRATUM CODE	SAMPLING AREA	AREA CODE	SUBAREA	SAMPLING SUBAREA CODE	GEAR CODE	YOY (AGE 0)	JUVENILES	MEDIUM	**TOTAL**
12d	Central Mediterranean	CMED	Sicily (East Sicily and Ionian Sea)	ESI	LL	**85**	**13**	**50**	**148**
13c	Central Mediterranean	CMED	Adriatic Sea	AS	UNCL			**50**	**50**
15/15c	Western Mediterranean	WMED	Tyrrhenian Sea (South)	TY	HL/LL		**15**	**35**	**50**
					TOTAL	**85**	**28**	**135**	**248**

The sampling of juveniles was only partially accomplished because the scarcity of commercial catches of this stage. On the contrary, the great abundance of young-of-the-year individuals registered in the East Sicily and Ionian area led the opportunity to collect a high number of individuals at this stage (YOY). According to the mail correspondence occurred with the previous GBYP Coordinator in December 2018 and because the scarcity of YOY collected from other strata and other consortia, we added to the sampling strata the 12d Central Mediterranean - Sicily (East Sicily and Ionian Sea) – YOY. A number of YOYs was collected up to fill the amount of financial resources contracted with ICCAT GBYP.

The map reported in Figure 1 illustrated the sampling locations with sample sizes according to size classes.

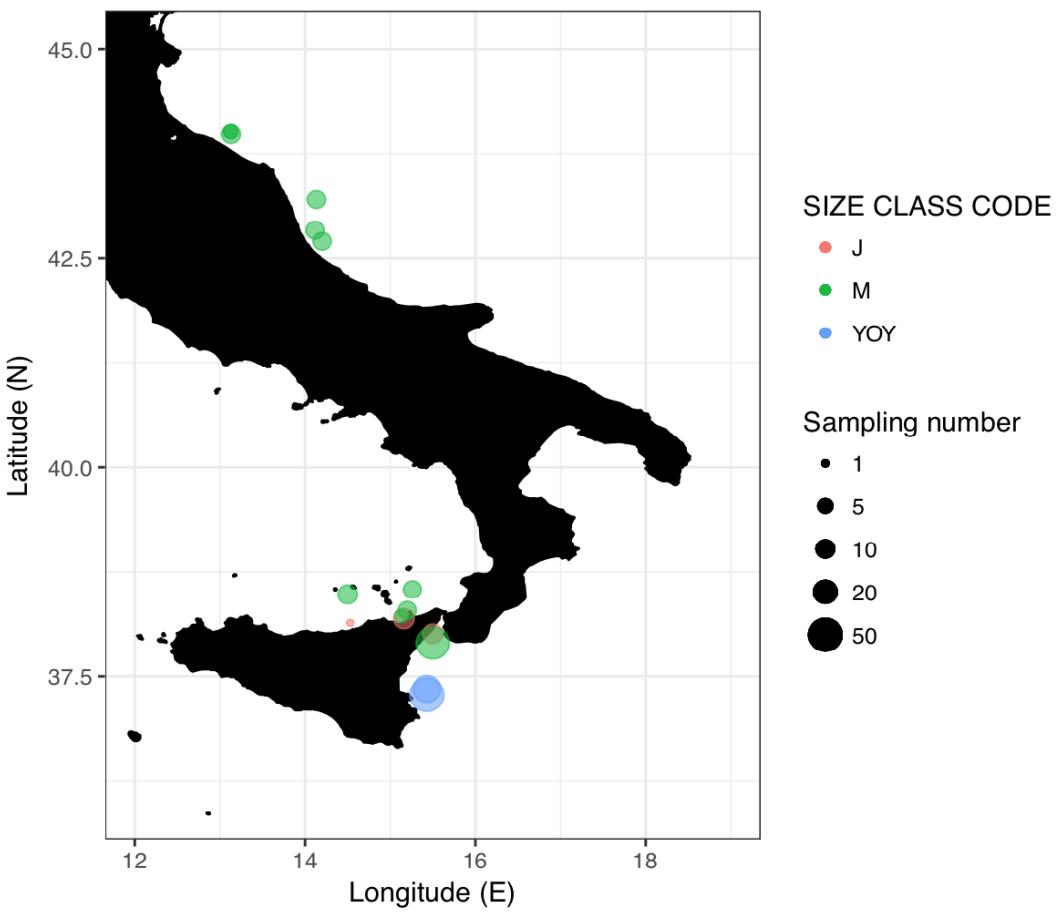

Figure 1. Map of the sampling locations with sample sizes according to size class (J: Juveniles; M: Medium; YOY: Young-Of-the-Year).

Detailed sampling data for all strata have been submitted to the GBYP Database Manager by the 31/01/2018. Sampled tissues will be submitted within 16/02/2018 to the GBYP Database Manager.

d) Analyses (Genetic analyses – Additional genetic analyses)

The Partnership "NextGeneration BFT" was in charge of "genetic analyses of transcriptomic and genomic data exploiting previous available data for defining the genomic variability of the species". The main objective of the activity awarded was to capitalize on available genomic resources of the Atlantic Bluefin Tuna (BFT) to be potentially exploited in several relevant biological ways (e.g the understanding of the link between standing genetic variation and performance and production traits like growth rate, spawning behaviour). The BFT genomic resources could be ultimately translated to improve the current management and exploitation of the species, with a special focus on rearing conditions in fattening cages, eventual reproduction in captivity and broodstock management. Within the proposed genetic tasks special attention will be paid to investigating in the BFT genome the presence of candidate genes for sex-related traits.

The awarded activities were organized in main tasks and subtask as detailed below:

GBYPPhase7-Task1 – Genome assembly: to produce a high-quality BFT genome assembly using ad-hoc pipelines. This task aims at generating a high-quality genome assembly of Atlantic Bluefin Tuna (BFT), integrating available sequencing data produced during previous research projects.

GBYPPhase7-Task2 – Genome annotation: to identify the locations of genes and of the coding regions in the BFT genome and determine their function. Task 2 was further divided in sub-tasks A, and B:

1. **Subtask2a**: Gene structural annotation, consisting in the identification of complete putative gene loci along the genomic sequence.
2. **Subtask2b**: Gene functional annotation, consisting in the determination of the functions of identified protein-coding genes.

Introduction

Increasing evidence are building up about the potential of genomic approaches, which are an increasingly important aspect of understanding evolutionary processes, such as physiological and morphological adaptation to the diverse habitats within the marine environment (Kelley et al, 2016). In parallel advances in genome technology now facilitate genome assembly for comparative and population genomics. Moreover, genomics tools can allow tackling several challenges to ensure sustainability and profitability of tuna fisheries management (Aquaculture Genomics et al, 2017; Kumar and Kocour, 2017; Li and Wang, 2017). The complete understanding of the gene networks underlying the most relevant life history traits as growth, reproduction, and disease resistance can

provide the knowledge-base for the development of welfare-centred and sustainable management strategies (Trumbic et al, 2015).

In order to achieve these goals, a set of genomic tools and resources must be made available including reference genome sequences and their annotations, genome-wide polymorphic markers, high- density and high-resolution linkage maps, and transcriptome resources (Aquaculture Genomics et al, 2017).

However, for non-model species, numerous computational challenges remain for the de novo assembly of genomes (El-Metwally et al, 2013). Genome assembly is also hindered by the higher level of polymorphisms found in some marine vertebrate species, and especially for fish, by the occurrence of the whole-genome duplication in teleosts (Taylor et al, 2003).

Several examples are supporting the usefulness of assessing genomic and functional variability in exploited marine species, starting from the emblematic cod, that represents the first annotated genome sequence for a commercially important fish species, revealing complex thermal adaptation and an unusual genome architecture (Star et al, 2011).

So far, draft genomes have been published or are being sequenced in over 24 aquatic and aquaculture species (Yue and Wang 2017). Tuna genome data (draft assemblies) are known for three species: *Thunnus orientalis* (Nakamura et al, 2013), *Thunnus albacares* (Malmstrøm et al, 2017) and *Thunnus maccoyii* (unpublished data, McWilliam et al, 2016). Genomic data have also recently been used in *T. orientalis* to construct Genetic Linkage Maps (Uchino et al, 2016), and to develop oligonucleotide DNA microarrays (Yasuike et al, 2016).

Until recently, publicly available genomic resources for *T. thynnus* have been relatively scarce, most being derived from a single EST (expressed sequence tag) sequencing project by Chini (Chini et al, 2008). Construction of oligonucleotide DNA microarrays from *T. thynnus* cDNA library has provided new data to study the reproductive physiology of bluefin tunas and various aspects of its ecology and genomics with strong applications in aquaculture (Gardner et al, 2012; Trumbic et al, 2015).

The understanding of sex-control and sex-determination and key genes involved in the reproductive molecular pathway may represent useful information for future marker-assisted selection in tunas (Bar et al, 2016). Marker-assisted breeding programs, based on genomic and mitochondrial DNA analysis, are already in place for PBFT (Sadawa and Agawa, 2016).

The identification of sex-loci has several practical applications. For instance, when phenotypic sex is difficult or impossible to obtain if specimens are not large enough for a visual assessment of gonads (i.e. early larval stages) or if fish are in non-spawning condition and preferentially kept alive. Relevant to notice is the possibility to have sex information from tagged specimens released in the

wild, as this could provide an additional parameter to be integrated in the tagging data analyses. Moreover, it allows the assignment of the sex phenotype for specimens for which this is impossible to obtain by other means, such as archaeological bone material or historical specimens of unknown sex (Star et al, 2016 and references therein).

Recently, sex-associated loci have been identified, localized and mapped in the genome of important commercial fishes such as turbot, cod and Pacific Bluefin tuna (Figueras et al, 2016; Star et al, 2016; Uchino et al, 2016).

In general, genome-wide sequencing has been performed using one single specimen. Likewise, transcriptome studies are in general based on a restricted number of individuals (n=5-10; Kumar and Kocour, 2017; Li and Wang 2017; Ribeiro et al, 2017; Yue and Wang 2017 and references therein).

Since ICCAT GBYP Phase2, considerable effort and resources have been employed by GBYP partners to develop ex novo multiple genomic resources for the Atlantic Bluefin Tuna (ABFT). All data analyses have been steered by the need to quantify genetic differentiation between Western and Eastern stocks and assess population structure within the Eastern stock. Multiple SNP panel have been derived to address these tasks and the evaluation of their potential for stock traceability and mix stocks analysis is ongoing (Puncher et al, 2015; Puncher et al, 2018; Rodríguez-Ezpeleta et al, 2015; Rodríguez-Ezpeleta et al, 2016; Rodríguez-Ezpeleta et al, 2017; GBYP Biological studies Final Reports of Phase4, Phase5 and Phase6 available at GBYP Web Page.

Little attention has been dedicated so far to the analysis of the newly generated data in terms of functional variation among the stocks and within the Mediterranean samples.

In this work, sequencing data already available (genome sequencing data from 5 individuals and cDNA sequencing data from 10 individuals) were capitalized to obtain a high-quality BFT genome assembly and its gene annotation (both structural and functional), without any need to analyse additional specimens. The BFT genomic resources could be ultimately translated to improve the current exploitation of the species, since developing genomic resources for Atlantic Bluefin tuna is of crucial importance for the continuous modernization of techniques used for management of the species as well as other tuna species (Bernatchez et al, 2017). The ways by which these resources will benefit conservation and scientific community are multiple. Annotation and comparison of our reference against other tuna or Teleost genomes will allow better functional characterization of tuna genes linked to conservation and survival relevant life-history traits. Our genomic reference is of use for the mapping and positioning of current and future genomic markers. Genome-wide SNP

data, and especially outlier loci, have broad applications in future studies, allowing for local analyses and the validation of population structure at micro-scale levels.

Materials and Methods

DNA extraction and sequencing

The main genomic DNA sequencing data used to generate the BFT assembly was obtained in the framework of the previous research project "Consortium BFTbySNP" and GBYP-phase2. In particular, these data come from four individuals (two from Western Mediterranean - Age0 and two from Gulf of Mexico – Age0). Genomic libraries were sequenced with 2 × 75-bp paired-end module on 3.25 lanes of an Illumina HiSeq2000 platform, producing, in total, 833 million reads.
Additional genomic resources were generated in the context of an academic UNIBO-founded project and integrated to those described above. In particular, a genomic library has been produced from a YOY individual collected outside the GBYP sampling effort. A mate pair library with insert size of 5kbp has been prepared and sequenced on Illumina HiSeq2000 with the 100bp Pair-End module. Before performing assembly, all libraries were quality trimmed and deprived of TrueSeq adapters.
Both fragment and mate-pair libraries are summarized in Table 4.

Transcriptome data

In order to improve gene structural annotation, we exploited different sources of BTF transcriptomic data. For this purpose, we integrated cDNA sequencing data produced by the "Consortium BFTbySNP" and GBYP-phase2 (data for ten individuals) with additional BTF data taken from the NCBI public database of Short Read Archive (SRA). In particular, the following RNA-seq libraries were used:

- cDNA sequencing from 10 individuals from 3 supposed major spawning basins (3 from Western Mediterranean – Age0, 3 from Eastern Mediterranean - Adult and 4 from Gulf of Mexico – Age0). From a full plate run of the Roche GS-FLX DNA Sequencer with Titanium chemistry, we obtained 1,182,738 cDNA sequences, assigned to each one of 10 sequenced individuals ranging from 17,402 to 264,784 reads per individual. After trimming according to quality value and adapter removal from Roche-454 and SMARTer libraries construction, 1,021,388 reads were retained. Mitochondrial transcripts were removed by mapping trimmed cDNA reads against the complete BFT mitochondrial genomes retrieved from NCBI (accessions: GU256522 and AY302574), excluding 7.8% of available sequences. Filtering for

ribosomal transcripts was performed by local BLAST search and 954,599 filtered reads were retrieved for downstream analyses.

- Two mixed-tissue normalised BFT cDNA libraries produced (using Roche-454 platform) in the context of a different study (Trumbic et al, 2015) and obtained from a pool of three individuals (male and female, immature) and different tissues (gill, gonad, heart, intestine, head kidney, liver, spleen, whole blood, red and white muscle, skin scrape). For these two libraries, a total of 976,904 raw reads were downloaded from the Short Read Archive (SRA) (accessions: ERX555873, ERX555874, BioProject: PRJEB7253).
- One Illumina paired-end cDNA library (comprising 29,910,574 paired-end short reads) produced in a different study ("ABFT SNP" - Grant ID PIEF-GA-2010-IEF, Research Executive Agency - EU) and obtained from a pool of 14 bluefin tuna larvae at different stages of development and downloaded from the SRA (accession: SRX669996, BioProject: PRJNA252566);
- cDNAs (10,163 sequences) expressed in ovary, testis and liver of BFT (Chini et al, 2008), downloaded from the NCBI Expressed Sequence Tags database (accessions: EC091633–EC93160, EG629962–EG631176, EC917676–EC919417, EG999340–EG999999, EH000001–EH000505, EH667253–EH668984, EL610526–EL611807, EC421414–EC422414, and EH379568–EH380065).

Table 4. Genomic and transcriptomic sequencing libraries used for the BFT genome assembly and annotation.

Library ID	Technology	Type	Number of reads	After trimming/filtering
TunaPE01	HiSeq2000	PE	241,907,854	227,851,376
TunaPE02	HiSeq2000	PE	188,000,414	185,077,468
TunaPE03	HiSeq2000	PE	179,541,104	172,600,438
TunaPE04	HiSeq2000	PE	223,825,544	214,678,542
TunaMP01	HiSeq2000	MP	328,979,630	287,182,594
TunaRNASEQ	Roche-454	SE	1,182,738	954,599
ERX555873	Roche-454	SE	707,344	707,344
ERX555873	Roche-454	SE	269,560	269,560
SRX669996	Illumina	PE	29,910,574	29,910,574
TunaEST	Sanger	SE	10,163	10,163

Genome Assembly

Genome assembly of the BFT was performed adopting an ad-hoc, iterative procedure using different assembly tools and parameters. This was performed on a 48-processor workstation equipped with Intel Xeon E5-2670 2.3GHz processors and 512GB of RAM.

In the first instance, using the four PE fragment libraries reported in Table 1, several preliminary assemblies were generated using different *de Bruijn graph*-based software programs, including SOAPdenovo (Li *et al.*, 2010), ABySS (Simpson *et al.*, 2009), Gossammer (Conway *et al.*, 2012) and Velvet (Zerbino and Birney, 2008). Different runs of each assembler were performed using default parameters and varying the *k-mer* size (the main parameter for *de Bruijn* graph construction step) in the range of 27-64bp. The different assemblies produced were therefore evaluated using the QUAST genome quality assessment tool (Gurevich *et al.*, 2013), assuming an estimated genome size of 780Mbps (Hardie and Hebert, 2004). Assembly statistics were all computed by firstly filtering-out contigs shorter than 1000bp. After comparative analysis of NG50 statistics, the assembly generated using the ABySS assembler with k-mer size 57 was chosen as the best one according to NG50 statistics.

The optimal assembler and k-mer size were then selected to generate BFT assembly scaffolds. In particular, an additional run of ABySS was performed using (i) the four PE fragment libraries to generate contig level assembly and (ii) the MP library to perform assembly scaffolding. The actual scaffolds were computed from assembled contigs using the internal ABySS scaffolding module. The scaffold-level assembly generated was finally filtered according to scaffold/contig length (sequences below 1000bp were filtered-out and not considered for downstream analyses).

As a final step, in order to reduce the number of gaps introduced after scaffolding, the four fragment libraries were re-aligned back to the assembly. Then, the GapFIller software (Boetzer and Pirovano, 2012) was used to locally refine the assembly in the attempt of closing gaps.

The final assembly was validated for completeness using CEGMA (Parra *et al.*, 2007) with default parameters against the 248 genes in the Core Eukaryotic Gene (CEG) dataset, and BUSCO (Simão *et al.,* 2015), using the Actinopterygii lineage dataset comprising 4584 genes.

Genome structural annotation

The final gene inventory of the BFT was obtained using the AUGUSTUS gene-finding software (Stanke *et al.*, 2008) with a BFT-specific gene model. In particular, instead of using one of the different pre-defined gene models (available only for a limited set of species such as human or zebrafish), in order to improve the gene annotation quality, we decided to re-train AUGUSTUS with

a BFT-specific gene statistical model obtained directly from our genome assembly (produced in the context of the GBYPPhase7-Task1).

Hence, our gene-finding pipeline consisted in two phases: (i) the identification of a reliable set of gene loci to be used as training set for the generation of the BFT-specific gene model; (ii) the actual prediction of putative protein-coding genes using AUGUSTUS with the newly generated BFT model. As a first, preliminary step, a BFT repeat library was obtained by running the RepeatModeler software (Smit and Hubley, 2008-2015) with default parameters on complete scaffolds of the genomic sequence. The repeat library was then used to mask repetitive sequences on the genome by means of RepeatMasker (Smit et al., 2013-2015). The repeat-masked genomic sequence was then used in all subsequent gene prediction steps.

Next, we identified on the BFT genome sequence a set of loci that have been subsequently adopted as training set for the AUGUSTUS training module. To accomplish this task, four different ab-initio gene finders, including SNAP (Korf, 2004), GENSCAN (Burge and Karlin, 1998), GlimmerHMM (Majoros *et al.*, 2004) and AUGUSTUS itself were run on the complete BFT genome sequence. Each software run produced a genome-wide prediction of complete gene structures, including genomic locations for coding sequences, introns, start and stop codons. The outputs of the different programs were then analysed and cross-compared, retaining only those genomic loci that were supported by predictions from at least two different programs. After this step, we obtained a preliminary set of gene loci.

To reduce false positive predictions, we performed an additional filtering procedure at the level of protein sequence. Hence, all the predicted amino acid sequences were computed for each preliminary gene and compared to proteins from the five different teleost fishes available at the Ensembl databank (Aken *et al.*, 2016). To obtain a non-redundant protein collection, complete proteomes from the five teleosts (zebrafish, medaka, greenpuffer, stickleback and fugu) were preliminary cross-compared and clustered according to sequence similarity. Only representative sequences were retained for each protein cluster. Table 5 reports a summary of the five protein datasets selected.

Table 5. Summary of protein datasets from the five teleost fishes at the Ensembl databank.

Organism	**Number of non-redundant proteins**
Danio rerio (zebrafish)	43,931
Oryzias latipes (medaka)	24,545
Tetraodon nigroviridis (greenpuffer)	23,067
Gasterosteus aculeatus (stickleback)	27,557
Takifugu rubripes (fugu)	47,789
Total	**166,889**

Comparisons between ab-initio predicted proteins and the teleost dataset were performed using the BLASTP program (Altschul *et al.*, 1997) with an e-value threshold of 1e-5. Predicted proteins were retained if a significant match was found with more than 25% sequence identity and covering more than the 60% of both the query and the target teleost protein sequence. The resulting set of predicted proteins were finally homology-reduced at 25% sequence identity and 90% sequence coverage (a necessary step to improve the training procedure) and corresponding gene structures used to train the BFT gene model by the AUGUSTUS training module.

After training, the final gene structures were predicted on the BFT genome running AUGUSTUS in prediction mode with the newly generated BFT gene model. The prediction was performed requiring only complete gene structures in output (option --gene-model=complete), also integrating transcriptomic data as additional support for the prediction.

In particular, in order to generate hints for AUGUSTUS, we adopted an iterative procedure comprising two steps.

Firstly, RNA-seq/EST sequences were spliced-aligned to the genome reference sequence. Different aligners were used according to the library type: for Roche-454 reads and for EST sequences, we used the BLAT aligner (Kent, 2002); Illumina PE RNA-seq reads were instead aligned using the TopHat2 program (Kim *et al.*, 2013). From alignment results, we identified spliced reads that spanned an exon-exon boundary and hence delineated a complete intron, leading to a preliminary set of intron hints.

However, when spliced alignment is performed reads will often not be aligned correctly, particularly when a short part of a read is aligned in one exon and the larger part in another exon. To overcome this problem and increase the number of reads properly mapped (and hence the number of hints provided to AUGUSTUS), starting from preliminary spliced alignments, we selected those reads that were properly mapped and we used them to generate exon-exon junction sequences by concatenating regions flanking each identified intron.

The complete set of reads was then re-aligned to exon-exon junction sequences, using BLAT for 454 or EST sequences and Bowtie2 (Langmead and Salzberg, 2012) for Illumina reads. From reads mapping to exon-exon junctions an additional set of hints was generated.

In this way, each read was mapped twice: once directly to the reference genome and once to the exon-exon junction sequences obtained from the first alignments. Adopting this strategy significantly improved the number and quality of AUGUSTUS hints.

The same procedure was performed for each set of transcriptomic data considered. All hints were finally collected into a single file which was provided to AUGUSTUS as support for gene prediction.

Gene functional annotation

Biological functions were assigned to predicted BFT protein sequences using two complementary approaches: (i) homology-based information transfer using platforms such as Blast2GO v5.0.3 (Conesa *et al.*, 2005; Götz *et al.*, 2008) and the Bologna Annotation Resource v3 (BAR3; Profiti *et al.*, 2017); (ii) a pipeline integrating several predictive tools for improving Gene Ontology (GO) cellular component annotation.

Blast2GO annotates input protein sequences using BLAST searches to find one or more similar sequences from which GO annotations are retrieved. We ran Blast2GO against the NCBI non-redundant database using a BLAST e-value threshold of 1e-3. The Blast2GO taxonomic filter was enforced to only select BLAST hits from Vertebrata, ensuring coherent functional annotations. Moreover, additional annotations were obtained by running InterProScan (Jones *et al.*, 2014) (using the designated Blast2GO feature) to search against public databases including PFam (Finn *et al.*, 2016), PANTHER (Thomas *et al.*, 2003), SMART (Letunic and Bork, 2017) and ProDom (Bru *et al.*, 2005). Additional GO terms derived from InterProScan results were finally merged with primary Blast2GO annotations.

BAR3 (Profiti *et al.*, 2017) is an annotation platform based on a pre-computed graph-based clustering of the UniProtKB database. More specifically, in BAR3, all the UniProtKB sequences are firstly compared pairwise, identifying pairs of proteins fulfilling strict similarity criteria (sequence identity >= 40% and pairwise alignment coverage >= 90%). A graph is then built connecting identified protein pairs and clusters of similar sequences are identified by isolating connected components of the graph. Each cluster is annotated by computing over-represented GO annotations through a Bonferroni-corrected Fisher's exact test. New sequences enter a cluster after satisfying similarity constraints (with at least one protein in the cluster) and are annotated by similarity using statistically-validated cluster annotations. The current version of BAR3 has been built using the UniProtKB release 2016_05 and includes 28,869,663 sequences in 1,361,773 clusters (Profiti *et al.*, 2017). BAR3 annotations for predicted BTF protein sequences were obtained and merged with Blast2GO/InterProScan GO annotations.

As a complement of the above-mentioned homology-based approaches, sequence analysis tools were applied to annotate protein features and sub-cellular localization of proteins for which transfer by homology failed. Tools adopted include: DeepSig (Savojardo *et al.*, 2018), to detect and annotate secretory signal peptides, TPpred3 (Savojardo *et al.*, 2015), to predict mitochondrial-targeting peptides, MemPype (Pierleoni *et al.*, 2011), to predict alpha-helical transmembrane domains and membrane protein localization, and BaCelLo (Pierleoni *et al.*, 2006) to discriminate sub-cellular

localization of globular proteins. Output of different sequence analysis programs were analysed and processed using an in-house pipeline to provide coherent GO cellular component and features annotations.

Identification of sex-linked loci

Sex-related genes characterized in previous studies in *Thunnus orientalis* (Agawa *et al.*, 2015; Uchino *et al.*, 2016) and in other teleost fishes (Figueras *et al.*, 2016; Star *et al.*, 2016; Purcell *et al.,* 2018) were downloaded and aligned to the BFT genome assembly using the exonerate software v.2.2.0 (Slater & Birney, 2005) with option --model coding2genome. Top-scoring hits were identified by setting exonerate option --bestn 1. Other DNA sequences known to be involved in sex differentiation (Agawa *et al.*, 2015; Star *et al.*, 2016) were mapped on the BFT genome using blastn.

Results and Discussion

Genome assembly results

In Table 6 are summarized genome assembly statistics. The resulting assembly has a total size of 770,762,883 (very close to the estimated genome size of 780Mbp) and it comprises 184,739 contigs and 103,646 scaffolds, with contig-level N50 and NG50 of 5,226 and 4,436, respectively, and scaffold-level N50 and NG50 of 43,769 and 42,677, respectively. The genome GC content is 39.65%.

Table 6. Genome assembly statistics assuming an estimated genome size of 780Mbp.

Assembly level	Assembler	Total length (bp)	# of Sequences	N50 (bp)	NG50 (bp)	GC content (%)
Scaffold	ABySS	770,762,883	103,646	43,769	42,677	39.65
Contig	ABySS	690,414,488	184,739	5,226	4,436	39.65

Currently available genome data (draft assemblies) for *Thunnus* species report comparable values of N50 statistic:

Thunnus orientalis (Nakamura et al, 2013)→ N50 scaffold 136,950 bp

Thunnus albacares (Malmstrøm et al, 2016,2017) → N50 scaffold 46,871 bp

Thunnus maccoyii (unpublished data, McWilliam et al, 2016) → N50>174,000 bp

CEGMA and BUSCO validation results are reported in Tables 7 and 8, respectively. CEGMA evaluation showed 51.6% of gene completeness for the assembly (75.8% if also partial matches are considered), while BUSCO analysis using the Actinopterygii dataset (4584 genes) reported 59.8%

and 81.9% of complete and partial matches, respectively. Furthermore, BUSCO reported a low core gene duplication rate (2.8%) and missing rate of 18.1%.

Table 7. Results of the CEGMA analysis on the BFT genome assembly.

CEGMA report	**# of genes (%)**
Complete	128 (51.6)
Partial	188 (75.8)

Table 8. Results of the BUSCO analysis on the BFT genome assembly.

BUSCO report	**# of genes (%)**
Complete	2742 (59.8)
Complete and single-copy	2613 (57.0)
Complete and duplicated	129 (2.8)
Fragmented	1013 (22.1)
Missing	829 (18.1)
Total groups searched	4584 (100)

Genome annotation results

The generation of the BFT-specific gene model has been performed using 5162 non-redundant training gene structures obtained as detailed in the Methods section. In Table 9 are reported details of the individual contributions from the different ab-initio gene finders adopted as well as the composition of the final training dataset.

Table 9. Generation of the BFT-specific gene model: ab-initio gene prediction on the BFT genome.

Program	**Gene model**	**Predicted genes**	**Genes retained**[a]	**Non-redundant genes**[b]
AUGUSTUS	Human	24626	3098	1537
GlimmerHMM	Zebrafish	93406	3457	2697
Genscan	Human	10369	864	386
SNAP	Human	20836	650	542
Total	**-**	**149237**	**8069**	**5162**

[a] Number of genes retained after comparison with the teleost protein dataset

[b] Number of genes retained after homology reduction

Using AUGUSTUS with the newly generated gene model as well as with the support of hints generated from transcriptomic data, we ended up with 41,508 protein-coding genes predicted on the BFT genome assembly (Table 10). The longest and shortest predicted proteins comprise 4,994 and 66 residues, respectively, while the average predicted protein length is of 306 residues (see Figure 2).

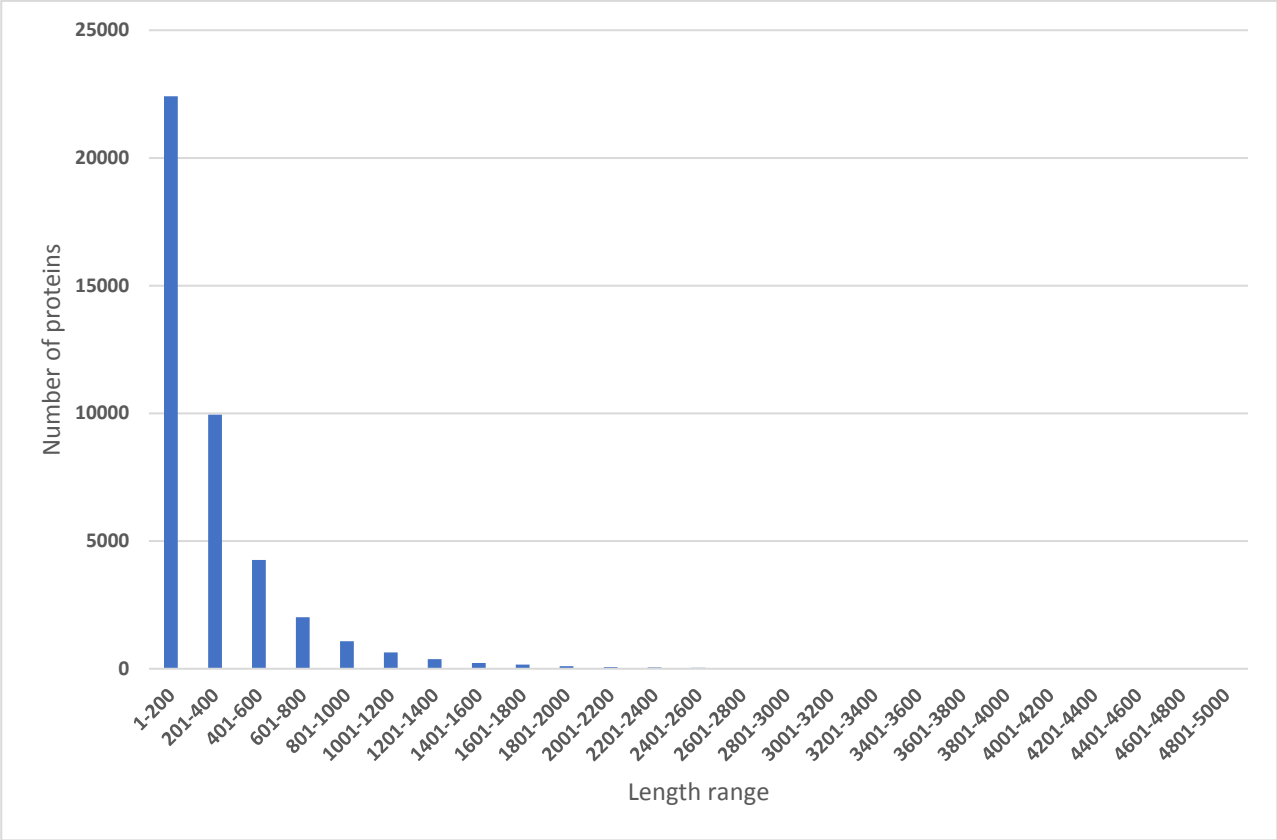

Figure 2. Length distribution of predicted protein sequence

Interestingly, 28,785 out of 41,508 predicted genes (about the 69%) were supported from at least one hint generated from transcriptomic data.

To further assess the quality of our results, all predicted protein sequences were mapped using BLASTP (with e-value threshold set to 0.001) to the Uniref90 reference clusters (Suzek *et al.*, 2015). 35,696 out of 41,508 proteins (86%) had a BLASTP match with at least one Uniref90 cluster. Among these, 35,339 (more than 99% of all matches) of top-scoring hits were to eukaryotic clusters, while 34,625 (97% of all matches) proteins mapped to Teleostei protein sequences. This result reveals a very low rate of contamination in our genome as well as a good quality of gene predictions.

Table 10. Final gene prediction on the BFT genome using the BFT-specific gene model and transcriptomic data support.

Genes	# of genes	Hinted
Total predicted	41,508	28,785
Mapped to at least one Uniref90 cluster	35,696	24,987
Mapped to Eukaryota Uniref90 clusters	35,339	24,030
Mapped to Teleostei Uniref90 clusters	34,625	24,584

In Figure 3 is shown the taxonomic distribution of BLASTP best hits (only taxa with more than 500 matches are reported). The taxa with the highest number of matches was the *Lates calcarifer* (barramundi).

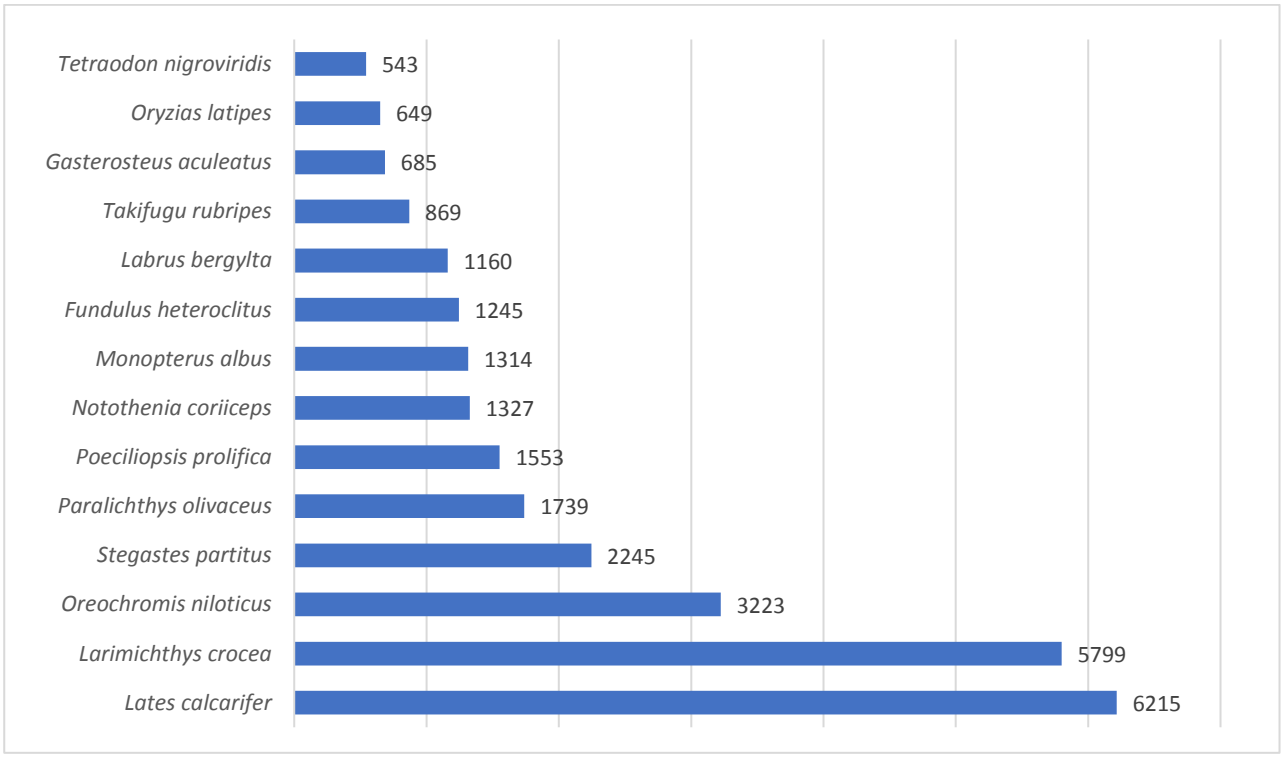

Figure 3. Distribution of BLASTP best matches of predicted proteins on the Uniref90 database. Only taxa with more than 500 matches are shown.

Functional annotation of BFT proteins

All the 41,508 predicted BFT proteins were subjected to functional annotation using Blast2GO and BAR3 platforms (seed Methods section for details). Table 11 reports a summary of annotated GO terms on the 41,805 BFT proteins. Overall, 26,151 protein sequences (corresponding to the 63% of all BFT proteins) were annotated with 13,915 different GO terms in the three ontology aspects (i.e. biological process, molecular function and cellular component). The 67% of all annotated GO terms

are in biological process sub-ontology (9,409 terms out of 13,915) whereas 23% are molecular function (3,192 terms out of 13,915) and 10% in cellular components (1,314 terms out of 13,915). This corresponds to the typical situation of many genome annotation projects and reflects the different abundances of terms defined in the three sub-ontologies.

Table 11. Summary of GO annotations on the BFT protein sequences obtained using homology-based approaches.

	Blast2GO+InterProScan		BAR3		Total	
GO aspect	**# of terms**	**# of proteins**	**# of terms**	**# of proteins**	**# of terms**	**# of proteins**
Biological Process	4,173	18,070	8,504	3,731	9,409	19,004
Molecular Function	1,984	20,587	2,458	3,549	3,192	21,177
Cellular Component	807	14,869	1,149	3,849	1,314	16,196
Total	6,964	25,810	12,111	4,124	13,915	26,151

To increase the coverage of functional annotation, the 25,312 BFT proteins missing annotation in the GO cellular component ontology were further processed using tools to predict protein features and sub-cellular localization (see section Methods for details). Figure 4 reports the distribution of predicted GO cellular component terms while Figure 5 provide summary of annotated protein features.

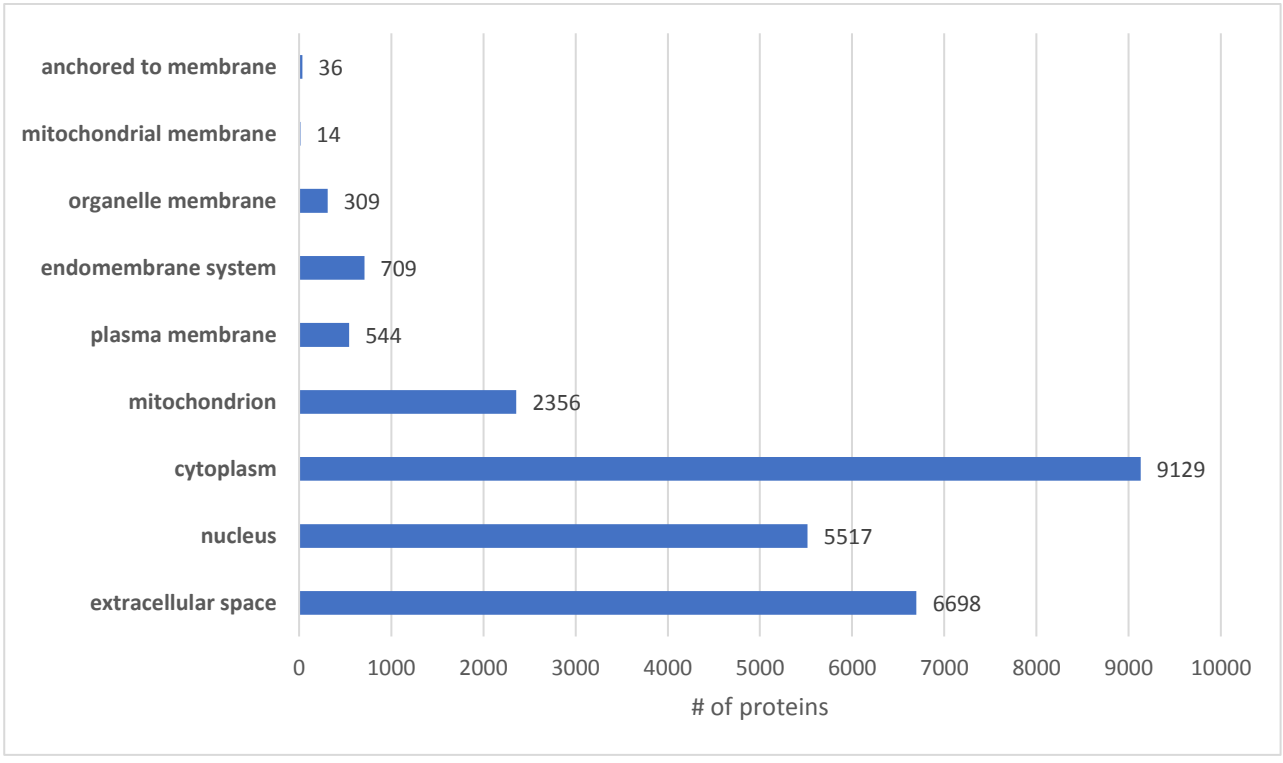

Figure 4. Distribution of GO cellular component terms predicted by sequence analysis tools on BFT protein sequences missing homology-based annotation.

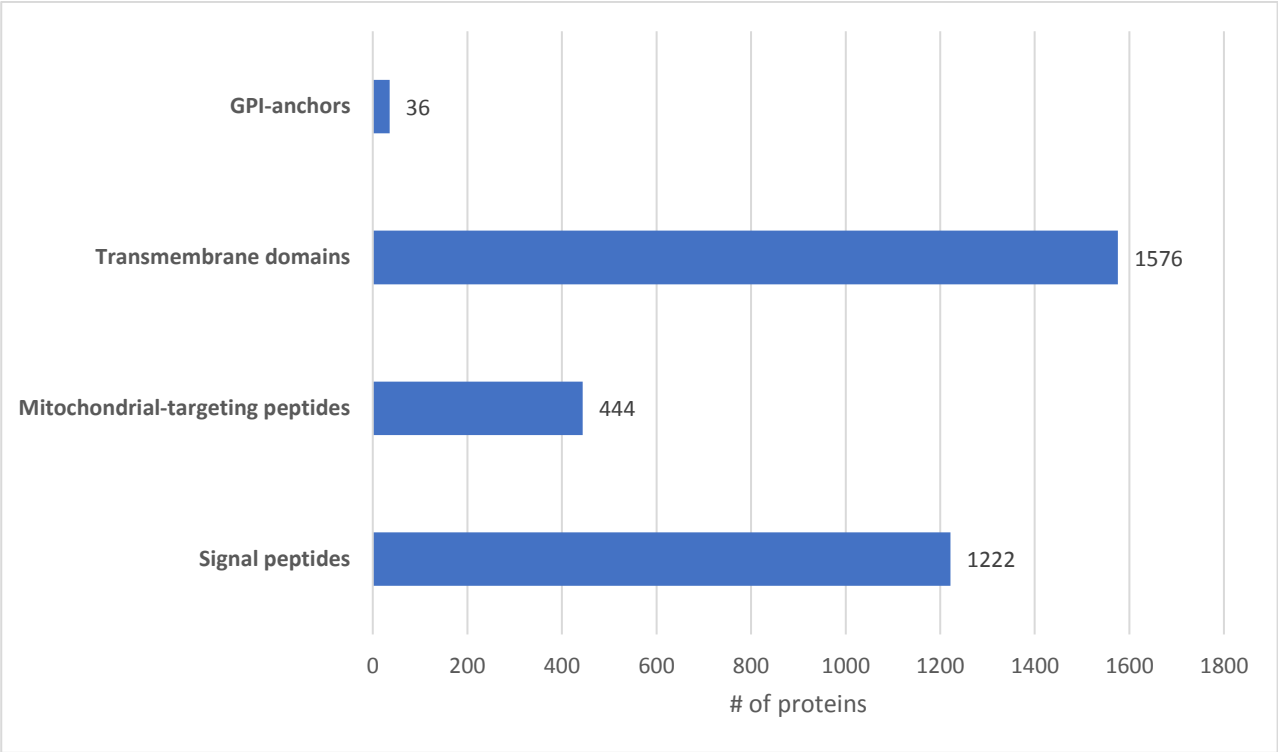

Figure 5. Distribution of protein features predicted by sequence analysis tools on BFT protein sequences missing homology-based annotation.

For sake of completeness, we report in Figures 6, 7 and 8, the distribution of the 20 most abundant GO terms assigned in the biological processes, molecular function and cellular component aspects, respectively.

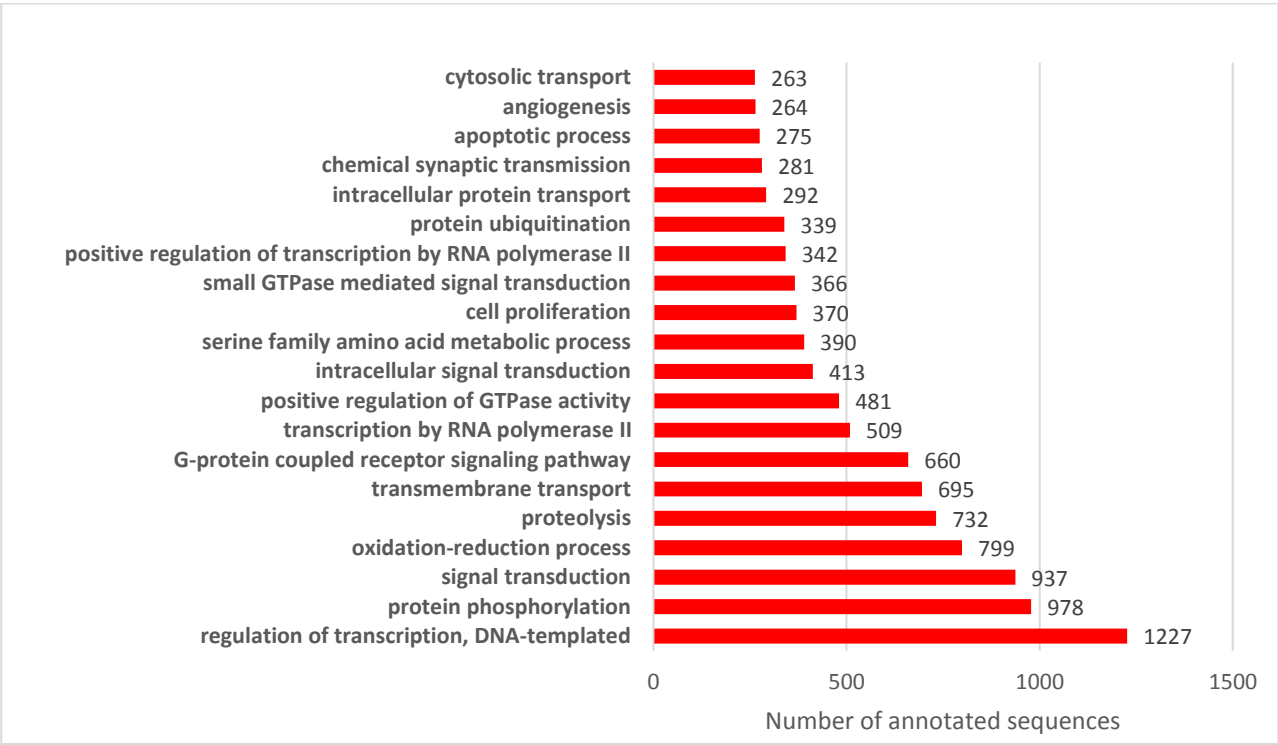

Figure 6. Distribution of the 20 most abundant GO biological processes annotated on BFT protein sequences using homology-based approaches.

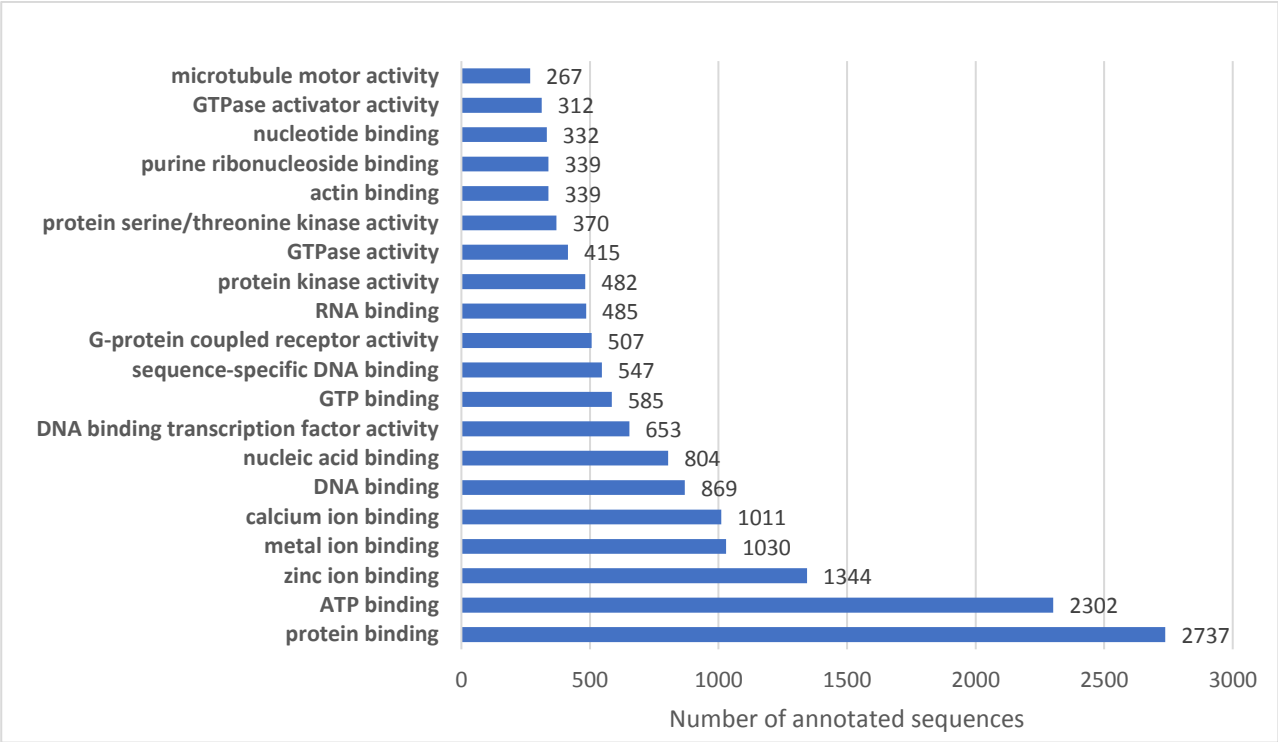

Figure 7. Distribution of the 20 most abundant GO molecular functions annotated on BFT protein sequences using homology-based approaches.

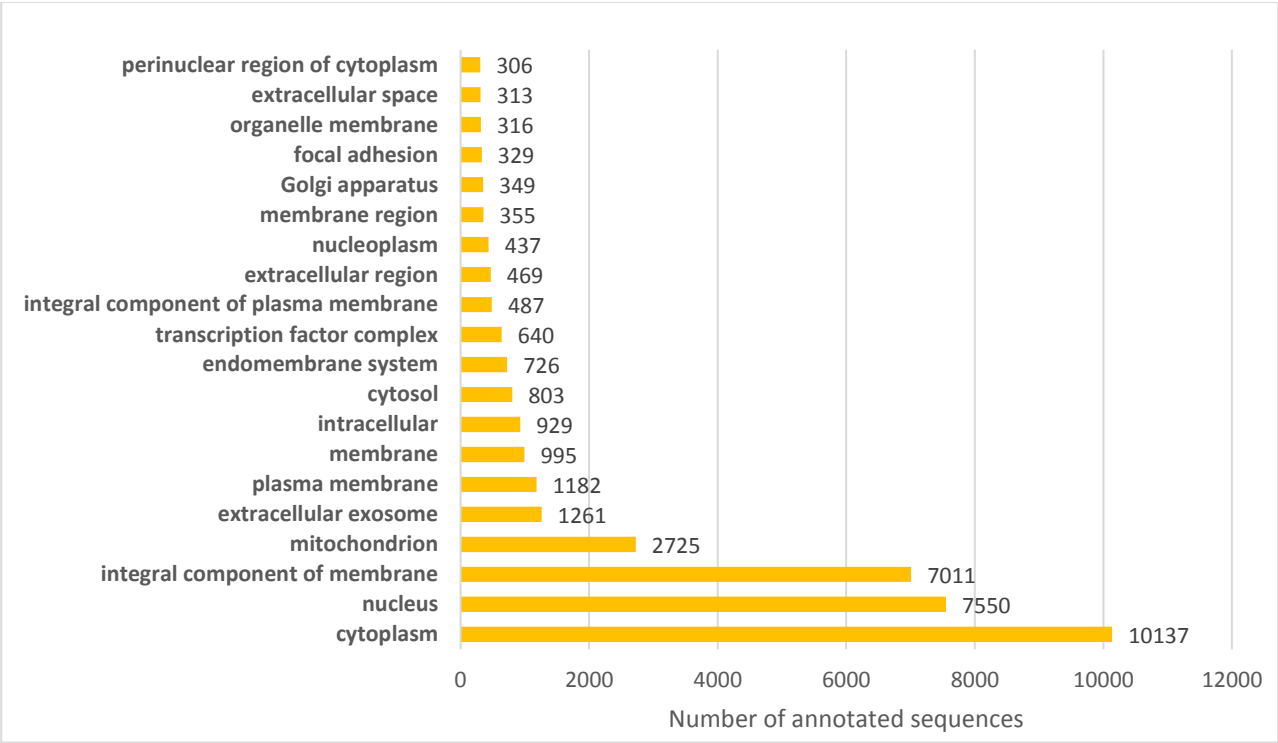

Figure 8. Distribution of the 20 most abundant GO cellular components annotated on BFT protein sequences either by homology-based approaches or predictive tools.

In summary, our results provide, for the first time, a genome-wide annotation of protein functions on the BFT genome sequence. Moreover, functional annotation coverage is in line with what

expected starting from available genomic and transcriptomic data. Further analyses will be carried-out in the future to improve the quality and coverage of annotations and to provide a centralized web resource to make data accessible to the scientific community.

Characterization of sex-linked loci on the BFT genomic sequence

The presence of candidate genes and markers for sex-related traits was investigated in the BFT genome by searching for sequence similarity with candidate sex-determining genes characterized in previous studies in *T. orientalis* and other bony fishes (as zebrafish, cod, medaka, Patagonian pejerrey, fugu, rainbow trout, turbot, Yellowtail). Only 3 out of the 35 candidate genes and markers did not find a match on the assembled BFT genome (Tables 12 and 13). All other sequences were located each in different scaffolds, not supporting the identification of a well-defined sex-determining region in the BFT genome. However, our results provide a first preliminary identification of putative regions prone to be further investigated using data from BFT individuals of known sex.

Table 12. Results of the alignment of known sex-related genes from various species on the BFT genome sequence.

				Alignment result					
Gene name	**Accession**	**Species**	**Reference**	**Query range**	**Scaffold**	**Start**	**Stop**	**Strand**	**Score**
akap11	XM_011473624.2	*O.latipes*	A	473-5531	2349	15173	30052	+	5528
amh	JN802292.1	*G.morhua*	A	67-439	928	55854	56214	-	163
amhr2	NM_001280009.1	*T.rubripes*	A	1005-1566	1708	23214	25277	+	548
amhy	HM153803.1	*O.hatcheri*	A	36-1179	928	55821	57307	-	875
ar	FJ268742.1	*G.morhua*	A	439-2146	519	132904	140538	-	1757
cyp19a	DQ402370.1	*G.morhua*	A	196-925	11341	5496	6679	-	900
cyp19b	JN802291.1	*G.morhua*	A	133-1480	14002	5296	7336	-	1750
dmrt	AJ506094.1	*G.morhua*	A	2-158	368	90573	90729	+	245
dmrt2a	JN802284.1	*G.morhua*	A	379-970	12922	3544	5232	-	564
dmrt3	JN802285.1	*G.morhua*	A	1-313	68550	598	910	+	528
dmrt4	JN802286.1	*G.morhua*	A	1-487	638	90414	90894	+	419
dmrt5	JN802287.1	*G.morhua*	A	1-1360	1108	33325	35168	-	1781
dmy	NM_001104680.1	*O.latipes*	A	246-432	1108	34838	35015	-	264
esr1	JX178935.1	*G.morhua*	A	255-1353	1096	40805	47495	+	1365
esr2a	JX178936.1	*G.morhua*	A	573-2175	3110	4707	17511	+	1812
esr2b	JK993476.1	*G.morhua*	A	35-182	3110	5457	5604	+	221
foxl2	NM_001104888.1	*O.latipes*	A	210-1155	14753	6115	7060	-	1661
gsdf	KC204828.1	*G.morhua*	A	358-433	2555	50416	50491	+	104
sdY	NM_001281416.1	*O.mykiss*	A			No match			
sox3	AB775143.1	*O.dancena*	A	163-1111	1946	29569	30517	+	1577
sox9a	JN802288.1	*G.morhua*	A	1-700	3029	5447	6380	-	1139
sox9b	JN802289.1	*G.morhua*	A	1-934	1495	88337	90017	+	1331
vasa	HM451456.1	*G.morhua*	A	702-1956	10807	6958	9623	+	1607
cyp19a1a	NM_131642.2	*D.rerio*	B	789-1576	14002	5296	6356	-	803
LHX9	NM_001037243.2	*D.rerio*	B	282-1662	2027	11548	19416	+	1960
SOX8A	NM_001284432.1	*D.rerio*	B	73-898	38311	718	2129	+	787
SOX17	NM_131287.2	*D.rerio*	C	153-627	3148	42661	43525	+	638
hsd17b3	AY551081.1	*D.rerio*	C			No match			
hsd17b1	BC163707.1	*D.rerio*	C	295-778	8512	57	824	-	404
sf1	BC163938.1	*D.rerio*	C	638-1457	62844	2	1530	-	1271
dmrt1	AY157561.1	*D.rerio*	C			No match			

A: Star *et al.*, 2016; B: Figueras *et al.*, 2016; C: Purcell *et al.,* 2018

Table 13. Results of the alignment of known sex-related markers on the BFT genome sequence.

		Alignment result						
Accession	**Species**	**Query range**	**Scaffold**	**Start**	**Stop**	**Strand**	**E-value**	**Score**
AB911114.1	*T.orientalis*	13-68	86382	1132	1187	-	1e-20	104
BADN01109032	*T.orientalis*	354-1564	10560	1564	1877	+	0	2217
BADN01109032	*T.orientalis*	1779-2043	10560	186	450	-	2e-134	484
BADN01109032	*T.orientalis*	1-109	10560	2122	2230	-	9e-48	196
BADN01109032	*T.orientalis*	2115-2210	10560	19	114	-	3e-37	161
EC421822.1	*T.thynnus*	250-533	4678	15917	16183	+	4e-62	243

Conclusions

- A high-quality BFT genome assembly has been generated capitalizing on available genomic data from previous GBYP-phases and on a genomic meta-pair library obtained from an academic UNIBO-founded project.
- The resulting genome assembly reported very good contiguity statistics (scaffold-level N50 and NG50 of 43,769 and 42,677, respectively) as well as a good level of completeness.
- Genome-wide annotation of protein-coding genes has been performed using a BFT-specific gene statistical model which have been incorporated into the AUGUSTUS gene-finding pipeline. Overall, 41,508 protein-coding genes were identified.
- The quality of the current gene annotation was enhanced by incorporating transcriptomic data from different sources including cDNA from 10 different individuals generated by us as well as BTF RNA-seq and EST data retrieved from public databases. These data were incorporated into the AUGUSTUS gene prediction pipeline to guide and support the identification of gene structures.
- The quality of gene annotations has been assessed by comparison of predicted protein sequence with proteins from other species, showing a high rate of similarity with genomes of other fishes (97% of predicted proteins mapped to Teleostei Uniref90 reference clusters). This result suggests a good quality of our gene annotations as well as a low level of contamination in the primary genomic data.
- All the 41,508 predicted BFT proteins were subjected to functional annotation and 63% of the candidate sequences (26,151 proteins) were associated to functions assigned by accurate homology-based approaches according to the standard catalogue of Gene Ontology (GO) covering with different proportions the three ontology aspects: i.e. biological process, molecular function and cellular component (with a total of 13,915 different GO terms).

Moreover, sequence analysis tools were adopted to complement the assignment of GO cellular component terms and to provide annotations of different types of protein features.

- The presence of candidate genes and markers for sex-related traits was investigated in the BFT genome by searching for sequence similarity with candidate sex-determining genes characterized in previous studies in *T. orientalis* and other bony fishs (as zebrafish, cod, medaka, Patagonian pejerrey, fugu, rainbow trout, turbot, Yellowtail). Only 3 out of the 35 candidate genes and markers did not find a match on the assembled BFT genome. All other sequences were located each in different scaffolds, not supporting the identification of a well-defined sex-determining region in the BFT genome. The identified genomic loci represent a baseline for further analysis targeting differential expression of those candidate region in BFT individuals of known sex (not available in this study).

REFERENCES

Agawa, Y., Iwaki, M., Komiya, T., Honryo, T., Tamura, K., Okada, T., Yagishita, N., et al. 2015. Identification of male sex-linked DNA sequence of the cultured Pacific bluefin tuna Thunnus orientalis. Fisheries Science, 81: 113-121.

Aken, B. L. *et al.* (2016) The Ensembl gene annotation system. *Database: The Journal of Biological Databases and Curation*, baw093.

Altschul, S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res*, **25**(17):3389–3402.

Aquaculture Genomics, G., Breeding, W., Abdelrahman, H., ElHady, M., Alcivar-Warren, A., Allen, S., Al-Tobasei, R., et al. 2017. Aquaculture genomics, genetics and breeding in the United States: current status, challenges, and priorities for future research. BMC Genomics, 18: 191.

Bar, I., Cummins, S., and Elizur, A. 2016. Transcriptome analysis reveals differentially expressed genes associated with germ cell and gonad development in the Southern bluefin tuna (*Thunnus maccoyii*). BMC Genomics, 17: 217.

Bernatchez L, Wellenreuther M, Araneda C, *et al.* (2017) Harnessing the Power of Genomics to Secure the Future of Seafood. *Trends in Ecology & Evolution*, 32(9) 665-680, https://doi.org/10.1016/j.tree.2017.06.010.

Boetzer, M. and Pirovano, W. (2012) Toward almost closed genomes with GapFiller. *Genome Biol.*, **13**(6):R56.

Boetzer, M. *et al*. (2011) Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics*, 27(4):578-9.

Bru, C. et al. (2005) The ProDom database of protein domain families: more emphasis on 3D. Nucleic Acid Research, 33: D212-D215.

Burge, C. and Karlin, S. (1997) Prediction of complete gene structures in human genomic DNA. *J Mol Biol.,* **268**(1):78-94.

Burley, S.K. *et al.* (2017) Protein Data Bank (PDB): The Single Global Macromolecular Structure Archive. *Methods Mol Biol.*, 1607:627-641.

Chini, V., Cattaneo, A. G., Rossi, F., Bernardini, G., Terova, G., Saroglia, M., and Gornati, R. 2008. Genes expressed in Blue Fin Tuna (Thunnus thynnus) liver and gonads. Gene, 410: 207-213.

Conesa, A. *et al.* (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, 21(18):3674-6.

Conway, T. *et al.* (2012) Gossamer—a resource-efficient de novo assembler. *Bioinformatics*, **28**(14):1937-8.

El‐Metwally, S., Hamza, T., Zakaria, M. & Helmy, M. (2013). Next-generation sequence assembly: four stages of data processing and computational challenges. PLoS Comput. Biol. 9, e1003345

Figueras, A., Robledo, D., Corvelo, A., Hermida, M., Pereiro, P., Rubiolo, J. A., Gomez-Garrido, J., et al. 2016. Whole genome sequencing of turbot (Scophthalmus maximus; Pleuronectiformes): a fish adapted to demersal life. DNA Res, 23: 181-192.

Finn, R.D. *et al.* (2016) The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.*, 44(D1):D279-85.

Gardner, L. D., Jayasundara, N., Castilho, P. C., and Block, B. 2012. Microarray gene expression profiles from mature gonad tissues of Atlantic bluefin tuna, Thunnus thynnus in the Gulf of Mexico. BMC Genomics, 13: 530.

Gurevich, A. *et al.* (2013) QUAST: quality assessment tool for genome assemblies. *Bioinformatics,* **29**(8): 1072-1075.

Götz, S. *et al.* (2008) High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acid Research*, 36: 3420-3435.

Hardie, D.C. and Hebert, P.D.N. (2004) Genome-size evolution in fishes. *Canadian Journal of Fisheries and Aquatic Sciences*, **61**:1636-1646.

Jones, P. *et al.* (2014) InterProScan 5: genome-scale protein function classification. *Bioinformatics*, 30(9): 1236-1240.

Kanehisa, M. *et al.* (2017) KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.*, 45(D1):D353-D361.

Kelley, J. L., Brown, A. P., Therkildsen, N. O., and Foote, A. D. 2016. The life aquatic: advances in marine vertebrate genomics. Nat Rev Genet, 17: 523-534.

Korf, I. (2004) Gene finding in novel genomes. *BMC Bioinformatics*, **5**: 59.

Kent, W. J. (2002). BLAT—The BLAST-Like Alignment Tool. *Genome Research*, 12(4), 656–664.

Kim, D. *et al.* (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology*, *14*(4), R36.

Kumar, G., and Kocour, M. 2017. Applications of next-generation sequencing in fisheries research: A review. Fisheries Research, 186: 11-22. doi:10.1016/j.fishres.2016.07.021

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4), 357–359.

Letunic, I. and Bork, P. (2017) 20 years of the SMART protein domain annotation resource. *Nucleic Acid Research*, 46(D1): D493-D496.

Li, R. *et al.* (2010) De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res.*, **20**(2):265-72.

Li, Y.-H., and Wang, H.-P. 2017. Advances of genotyping-by-sequencing in fisheries and aquaculture. Reviews in Fish Biology and Fisheries. doi:10.1007/s11160-017-9473-2

Majoros, W.H. *et al.* (2004) TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics*, **20**(16):2878-9.

Malmstrøm M, Matschiner M, Tørresen OK, Jakobsen KS, Jentoft S. (2017) Whole genome sequencing data and de novo draft assemblies for 66 teleost species. Scientific Data 4, Article number: 160132 doi:10.1038/sdata.2016.132

McWilliam, S., Grewe, P., Bunch, R., and Barendse, W. 2016. A draft genome assembly of southern bluefin tuna *Thunnus maccoyii*. eprint arXiv:1607.03955 available at https://arxiv.org/pdf/1607.03955.pdf

Nakamura, Y., Mori, K., Saitoh, K., Oshima, K., Mekuchi, M., Sugaya, T., Shigenobu, Y., et al. 2013. Evolutionary changes of multiple visual pigment genes in the complete genome of Pacific bluefin tuna. Proceedings of the National Academy of Sciences, 110: 11061-11066.

Parra, G. *et al.* (2007) CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics*, **23**(9):1061-1067.

Pierleoni, A. *et al.* (2006) BaCelLo: a balanced subcellular localization predictor. *Bioinformatics*. 22(14):e408-16.

Pierleoni, A. *et al.* (2011) MemPype: a pipeline for the annotation of eukaryotic membrane proteins. *Nucleic Acids Res.*, 39(Web Server issue):W375-80.

Profiti, G. *et al.* (2017) The Bologna Annotation Resource (BAR 3.0): improving protein functional annotation. *Nucleic Acids Res.* Epub:2017 Apr 27.

Puncher NG, Cariani A, Maes EG, Van Houdt J, Herten K, Albaina A, Cannas R, Rodríguez-Ezpeleta N, Arrizabalaga H, Addis P, Cau A, Goñi N, Fraile I, Laconcha Santamaria U, Tinti F (2015) Population structure and genetic management unit delineation in the bluefin tuna using a Genotyping-By-Sequencing approach. ICCAT SCRS/2015/048

Puncher, Gregory Neils; Cariani, Alessia; Maes, Gregory; Van Houdt, Jeroen; Herten, Koen; Cannas, Rita; Rodriguez-Ezpeleta, Naiara; Albaina, Aitor; Estonba, Andone; Lutcavage, Molly; Hanke, Alex; Rooker, Jay; Franks, James; Quattro, Joseph; Basilone, Gualtiero; Fraile, Igaratza; Laconcha, Urtzi; Goñi, Nicolas; Kimoto, Ai; Macías, A. David; Alemany, Francisco; Deguara, Simeon; Zgozi, Salem; Garibaldi, Fulvio; Oray, Isik; Karakulak, F. Saadet; Abid, Noureddine; Santos, Miguel; Addis, Piero; Arrizabalaga, Haritz; Tinti, Fausto (2018). Spatial dynamics and

mixing of bluefin tuna in the Atlantic Ocean and Mediterranean Sea revealed using next generation sequencing. Accepted by Molecular Ecology Resources MER-17-0349.R1.

Ribeiro, A. M., A. D. Foote, A. Kupczok, B. Frazao, M. T. Limborg, R. Pineiro, S. Abalde, S. Rocha & R. R. da Fonseca, 2017. Marine genomics: News and views. Mar Genomics 31:1-8 doi:10.1016/j.margen.2016.09.002.

Rodríguez-Ezpeleta N, Arrizabalaga H, Puncher GN, Alemany F, Deguara S, Rooker J, Lutcavage M, Oray I, Macías D, Valastro M, Irigoien X (2015) Inference of genetic population structure of Atlantic Bluefin tuna using RAD-seq derived SNP markers. SCRS/P/2015/007

Rodríguez-Ezpeleta Naiara, Díaz-Arce Natalia, Alemany Francisco, Deguara Simeon, Franks Jim, Rooker R Jay, Lutcavage Molly, Quattro Joe, Oray Isik, Macías David, Valastro Massimiliano, Irigoien Xabier, Arrizabalaga Haritz (2016) A genetic traceability tool for differentiation of Atlantic Bluefin Tuna (Thunnus thynnus) spawning grounds. SCRS/P/2016/032

Rodríguez-Ezpeleta Naiara, Díaz-Arce Natalia, Addis Piero, Abid Noureddine, Alemany Francisco, Deguara Simeon, Fraile Igaratza, Franks Jim, Hanke Alex, Itoh Tomoyuki, Karakulak Saadet, Kimoto Ai, Lauretta Matt, Lino G. Pedro, Lutcavage Molly, Macías David, Ngom Sow Fambaye, Notestad Leif, Oray Isik, Pascual Pedro, Quattro Joe, Richardson D. David, Jay R. Rooker, Valastro Massimiliano, Varela Jose Luis, Walter John, Irigoien Xabier, Arrizabalaga Haritz (2017). Genetic assignment of atlantic bluefin tuna feeding aggregations to spawning grounds. SCRS/2017/027.

Savojardo, C. *et al.* (2015) TPpred3 detects and discriminates mitochondrial and chloroplastic targeting peptides in eukaryotic proteins. *Bioinformatics*. 31(20):3269-75

Savojardo, C. *et al.* (2018) DeepSig: deep learning improves signal peptide detection in proteins. *Bioinformatics 2018*, Published online: 21 December 2017.

Sawada, Y., and Agawa, Y. 2016. Chapter 13 - Genetics in Tuna Aquaculture. In Advances in Tuna Aquaculture, pp. 323-332. Academic Press, San Diego.

Simão, F.A. *et al.* (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, **31**(19): 3210-3212.

Simpson, J.T. *et al.* (2009) ABySS: a parallel assembler for short read sequence data. *Genome Res.*, **19**(6):1117-23.

Slater, G. S. & Birney, E. (2005) Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*, **6**, 31.

Smit, A. and Hubley, R. (2008–2015) RepeatModeler Open-1.0. Available at http://www.repeatmasker.org.

Smit, A.F.A. *et al.* (2013–2015) RepeatMasker Open-4.0. Available at http://www.epeatmasker.org.

Stanke, M. and Morgenstern, B. (2005) AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res.*, 33(Web Server issue):W465-7.

Stanke, M. *et al.* (2008) Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics*, **24**(5):637-44.

Star B, Nederbragt AJ, Jentoft S, Grimholt U, Malmstrøm M, Gregers TF, Rounge TB, Paulsen J, Solbakken MH, Sharma A, Wetten OF. (2011) The genome sequence of Atlantic cod reveals a unique immune system. Nature 477, 207–210 (08 September 2011) doi:10.1038/nature10342

Star, B., Torresen, O. K., Nederbragt, A. J., Jakobsen, K. S., Pampoulie, C., and Jentoft, S. 2016. Genomic characterization of the Atlantic cod sex-locus. Sci Rep, 6: 31235.

Suzek, B. E. *et al.* (2015) UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, **31**(6):926–932.

Taylor, J. S., Braasch, I., Frickey, T., Meyer, A. & Van de Peer, Y. (2003). Genome duplication, a trait shared by 22,000 species of ray-finned fish. Genome Res. 13, 382–390.

Thomas, P.D, *et al.* (2003) PANTHER: a browsable database of gene products organized by biological function, using curated protein family and subfamily classification. *Nucleic Acid Research*, 31(1): 334-341.

Trumbic, Z., Bekaert, M., Taggart, J. B., Bron, J. E., Gharbi, K., and Mladineo, I. 2015. Development and validation of a mixed-tissue oligonucleotide DNA microarray for Atlantic bluefin tuna, Thunnus thynnus (Linnaeus, 1758). BMC Genomics, 16: 1007.

Uchino, T., Y. Nakamura, M. Sekino, W. Kai, A. Fujiwara, M. Yasuike, T. Sugaya, H. Fukuda, M. Sano & T. Sakamoto, 2016. Constructing Genetic Linkage Maps Using the Whole Genome Sequence of Pacific Bluefin Tuna (Thunnus orientalis) and a Comparison of Chromosome Structure among Teleost Species. Advances in Bioscience and Biotechnology 07(02):85-122 doi:10.4236/abb.2016.72010.

Vicedomini, R. *et al.* (2013) GAM-NGS: genomic assemblies merger for next generation sequencing. *BMC Bioinformatics*, 14 Suppl. 7:S6.

Yasuike, M., Fujiwara, A., Nakamura, Y., Iwasaki, Y., Nishiki, I., Sugaya, T., Shimizu, A., et al. 2016. A functional genomics tool for the Pacific bluefin tuna: Development of a 44K oligonucleotide microarray from whole-genome sequencing data for global transcriptome analysis. Gene, 576: 603-609.

Yue, G. H., and Wang, L. 2017. Current status of genome sequencing and its applications in aquaculture.

Zerbino, D. R. and Birney, E. (2008) Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, **18**(5):821–829.