Transcriptomic characterization of the larval stage in gilthead seabream (*Sparus aurata*) by 454 pyrosequencing.

Manuel Yúfera[1], Silke Halm[1], Sergi Beltran[2], Berta Fusté[3], Josep V. Planas[4], Gonzalo Martínez-Rodríguez[1]

Corresponding autor: email manuel.yufera@icman.csic.es
Running title: Transcriptomic of gilthead seabream larvae
Key words: 454 pyrosequencing, gilthead seabream, larvae, transcriptome

[1] Instituto de Ciencias Marinas de Andalucía (ICMAN-CSIC), Apartado Oficial 11510 Puerto Real, Cádiz, Spain.

[2] Unitat de Bioinformàtica, CCiT-UB, Universitat de Barcelona, 08028 Barcelona, Spain.

[3] Unitat de Genòmica, CCiT-UB, Universitat de Barcelona, 08028 Barcelona, Spain.

[4] Departament de Fisiologia i Immunologia, Facultat de Biologia, Universitat de Barcelona and Institut de Biomedicina de la Universitat de Barcelona (IBUB), 08028 Barcelona, Spain.

Abstract

Gilthead seabream (*Sparus aurata*) is a teleost belonging to the family Sparidae with a high economical relevance in the Mediterranean countries. Although genomic tools have been developed in this species in order to investigate its physiology at the molecular level and consequently its culture, genomic information on post-embryonic development is still scarce. In this study we have investigated the transcriptome of a marine teleost during the larval stage (from hatching to 60 days after hatching) by the use of 454-pyrosequencing technology. We obtained a total of 68,289 assembled contigs, representing putative transcripts, belonging to 54,606 different clusters. Comparison against all *S. aurata* expressed sequence tags (ESTs) from the NCBI database revealed that up to 34,722 contigs, belonging to about 61 % of gene clusters, are sequences previously not described. Contigs were annotated through an iterative Blast pipeline by comparison against databases such as NCBI RefSeq from *Danio rerio*, Swissprot or NCBI teleost ESTs. Our results indicate that we have enriched the number of annotated sequences for this species by more than 50 % compared with previously existing databases for the gilthead seabream. Gene Ontology analysis of these novel sequences revealed that there is a statistically significant number of transcripts with key roles in larval development, differentiation, morphology and growth. Finally, all information has been made available online through user-friendly interfaces such as GBrowse and a Blast server with a graphical frontend.

Keywords: *Sparus aurata*; 454-pyrosequencing; transcriptomic; post-embryonary stage.

Introduction

Gilthead seabream (*Sparus aurata*, Sparidae) is the most economically relevant marine teleost species produced by the aquaculture industry in the Mediterranean countries with a yearly average production ranging between 140,000 and 170,000 tm in the last four years (Apromar 2011). As a pioneer species of farmed marine fish, it has been the subject of scientific research for almost four decades. These studies have targeted all the different biological processes (reproduction, genetics, immunology, pathology, environmental stress, development, growth and nutrition) and stages of the life cycle (for review see Pavlidis and Mylonas 2011).

The larval stage is probably the most critical period in the production cycle of marine fishes. During the development to the juvenile stage, the larvae exhibit fascinating high growth rates with dramatic changes in anatomy, physiology and behaviour (Yúfera 2011; Yúfera et al. 2011). During this stage, the initial biomass at first feeding can increase from 2 to 3 orders of magnitude within a few weeks. Many key developmental events occur during the onset of feeding and the subsequent growing phase of actively eating larvae, making this period the most sensitive to rearing procedures. Morphological and physiological processes such as skeletal development and calcification, the differentiation and proliferation of new fibres in the myotomal musculature, achievement of complete structures in gut and gills, the progressive activation of pancreatic and intestinal digestive enzymes, swimming bladder inflation, completion of eye retina layers, development of lymphoid organs and acquisition of full immunological capacities, are dependent on complex mechanisms that regulate cell differentiation, organogenesis and the adequate functionality of organs and tissues (Alami-Durante et al. 2007; Campinho et al. 2010; Johnston 2006; Power et al. 2008; Yúfera and Darias 2007; Zapata et al. 2006). Any failure in these processes, highly dependent on the nutritional and environmental conditions as well as on the response capacity of the developing larvae to external stimuli, causes malformations, developmental delays, poor growth and massive mortalities (Barahona-Fernandez 1982; Johnston et al. 1998; Koumoundouros et al. 2009; Polo et al. 1991; Villeneuve et al. 2005; Yúfera et al. 1993). The availability of healthy and well-developed juveniles is the basis for a successful and efficient production in the fish farming industry. All developmental requisites during the larval period must be satisfied and

the potential problems properly solved to achieve a high juvenile quality. A deep knowledge of the genes regulating these developmental processes will allow a more exhaustive molecular approach for continuing unravelling the mechanisms of development from egg throughout juvenile stages, mainly in those aspects related to nutrition and growth, stress response and immunology.

During recent years, the use of molecular and genomic tools has contributed significantly to our understanding of key physiological processes in fish, including the larval phase (Mazurais et al. 2011). Collections of expressed sequenced tags (ESTs), generated by single-pass sequencing of cDNA libraries, are currently available for many freshwater and marine species of interest in aquaculture such as rainbow trout *Oncorhynchus mykiss* (Govoroun et al. 2006; Sánchez et al. 2009), Atlantic salmon *Salmo salar* (Adzhubei et al. 2007; Hagen-Larsen et al. 2005; Hayes et al. 2007), common carp *Cyprinus carpio* (González et al. 2007), catfish *Ictalurus punctatus* and *I. furcatus* (Li et al. 2007; Wang et al. 2010), Nile tilapia *Oreochromis niloticus* (Lee et al. 2010), cod *Gadus morhua* (Johansen et al. 2011), European seabass *Dicentrarchus labrax* (Chini et al. 2006; Louro et al. 2010), Asian seabass *Lates calcarifer* (Xia and Yue 2010), Atlantic halibut *Hippoglossus hippoglossus* (Douglas et al. 2007), Senegalese sole *Solea senegalensis* (Cerdà et al. 2008), turbot *Psetta maxima* (Pardo et al. 2008), and half-smooth tongue sole *Cynoglossus semilaevi* (Sha et al. 2010). For the gilthead seabream, an initial EST project yielded approximately 1,400 sequences, 810 of which were derived from a mixed juvenile and yolk sac larvae cDNA library and the other 584 sequences were derived from an adult liver cDNA library (Sarropoulou et al. 2005a). More recently, a large-scale EST sequencing project generated approximately 30,000 ESTs by Sanger sequencing of 14 normalized cDNA libraries from adult gilthead seabream tissues (Louro et al. 2010). However, no post-embryonic larvae or larval tissues were included for library construction and sequencing and, therefore, genes expressed preferentially throughout early development are most likely to be underrepresented in that particular dataset. Although it is clear that the genomic resources for this species have significantly improved over the last few years, there is still an important lack of information on genes expressed during early development. For this reason, further efforts to improve the contribution of larval sequences in existing EST collections for the gilthead seabream are needed.

In contrast to conventional Sanger sequencing of cDNA libraries, next generation sequencing technologies such as 454 pyrosequencing are able to offer an unprecedented, comprehensive view of the transcriptome. Over the last two years, several studies have reported on the use of pyrosequencing in teleost fish with various goals: to improve or establish transcriptome databases in the rainbow trout (Salem et al. 2010), eel *Anguilla anguilla* (Coppe et al. 2010), eelpout *Zoarces viviparus* (Kristiansson et al. 2009) and cod (Johansen et al. 2011); to characterize the immune response to bacterial pathogens in the Japanese seabass *Lateolabrax japonicus* (Xiang et al. 2010) and the yellow croaker *Pseudosciaena crocea* (Mu et al. 2010); to identify single nucleotide polymorphisms in catfish (Liu et al. 2011) and whitefish *Coregonus* spp (Renaut et al. 2010) and to elucidate the molecular basis behind different morphotypes and sympatric species (Elmer et al. 2010; Goetz et al. 2010; Jeukens et al. 2010). These studies have evidenced the usefulness of 454-pyrosequencing technology to improve the genomic resources of non-model teleost species. With the specific aim to enrich current EST databases for the gilthead seabream in order to comprehensively cover the larval developmental period and the transition to juvenile stages, we have used for the first time 454-pyrosequencing technology to significantly improve our knowledge of the transcriptome of this important marine species. Furthermore, we have implemented state-of-the-art informatics tools to provide easy access to all information.

Materials and Methods

Sample generation and RNA isolation

Gilthead seabream larvae hatched from eggs obtained from natural spawning and reared in 250-L tanks under 12 h light/12 h dark illumination cycle at 19.5 ± 1 ºC temperature and 35 g L^{-1} salinity following standard procedures (Polo et al. 1992). The larvae were fed on rotifers (*Brachionus rotundiformis* and *B. plicatilis*) from first feeding to 25 days after hatching (dah) and on fresh hatched *Artemia* nauplii and metanauplii from 15 and 25 dah, respectively. Commercial feeds were added from 50 dah onwards (Figure 1). Special care was taken in adjusting the amount of prey and dry food supplied the day before each sampling in order to allow that larval guts became empty during the night. Larvae were checked before the early sampling and those samples showing occasional guts with food content were discharged for the

analysis. Larvae for RNA extraction were periodically collected in pools from hatching until 60 dah (Figure 1) and preserved in RNAlater (Ambion//Life Technologies, Carlsbad CA, USA) at -20 ºC until further processing. Total RNA was extracted from approximately 20 mg of each larvae stage using the RNeasy Mini Kit with on column RNase-free DNase digestion (Qiagen, Hilden, Germany). RNA quantity was determined using the Eppendorf Biophotometer plus, and RNA quality assessed with the 2100 Bioanalyzer and the RNA 6000 Nano kit (Agilent, Santa Clara CA, USA). Only RNA with a RNA Integrity Number (RIN) higher than 9 was used for downstream applications. Equal amounts of RNA from larvae at 3, 4, 6, 8, 11, 25, 54 and 60 dah were pooled and reanalysed. Part of the remaining volume was split into two tubes, each containing 15 µg of total RNA at 600 ng/µL and with a RIN of 9.2.

cDNA synthesis and normalization

Full-length enriched double stranded cDNA was synthesized from 1.5 µg of pooled total RNA using MINT cDNA synthesis kit (Evrogen, Moscow, Russia) according to manufacturer's protocol, and was subsequently purified using the QIAquick PCR Purification Kit (Qiagen). The amplified cDNA was normalized using Trimmer kit (Evrogen) to minimize differences in representation of transcripts. The method involves denaturation-reassociation of cDNA, followed by a digestion with a Duplex-Specific Nuclease (DSN) enzyme (Shagin et al. 2002; Zhulidov et al. 2004). The enzymatic degradation occurs primarily on the highly abundant double-stranded cDNA fraction. The single-stranded cDNA fraction was then amplified twice by sequential PCR reactions according to the manufacturer's protocol. Normalized cDNA was purified using the QIAquick PCR Purification Kit (Qiagen).

454 pyrosequencing

Samples of 5 µg of normalized cDNA were used to generate a 454 library. cDNA was fractionated into small, 300 to 800 base pair (bp) fragments and the specific A and B adaptors were ligated to both the 3' and 5' ends of the fragments. The A and B adaptors were used for purification, amplification, and sequencing steps. One sequencing run was performed on the GS-FLX using Titanium chemistry (Margulies et al. 2005). 454 pyrosequencing is based on sequencing-by-synthesis, addition of one

(or more) nucleotide(s) complementary to the template strand results in a chemiluminescent signal recorded by the CCD camera within the instrument. The signal strength is proportional to the number of nucleotides incorporated in a single nucleotide flow. All reagents and protocols used were from Roche 454 Life Sciences, USA.

Transcriptome assembly

Raw sequencing data was processed with 454's gsRunProcessor 2.0.0.12 using default settings and SFF files were submitted to the NCBI Sequence Read Archive (Submission SRA038178.1). Those reads that passed all manufacturer's quality filters were processed with SeqClean software (command: seqclean 454Reads.fasta -v Remove -c 8 -l 50 -x 95 -y 18 -M -L , http://compbio.dfci.harvard.edu/tgi/software/) to remove poly-A tails plus primers and adapters used in the normalization procedure. The resulting reads were assembled as a 454 EST project with Mira v3rc3 (command: mira --project= 454ReadsClean --job=denovo,est,normal,454 454_SETTINGS -LR:mxti=no -CL:qc=no:cpat=no, Chevreux et al. 1999). The resulting contigs were screened for contaminants (against 94,681 *Artemia* and *Brachionus* NCBI nucleotide sequences, UniVec and *E. coli*) and low complexity sequences with SeqClean (settings: -l 1 -x 95 -M -A); the 62 contigs (15 with length > 99 bp) with low complexity and the 22 matching *Artemia* or *Brachionus* sequences are shown in Supplementary File 1. Contigs were grouped in clusters with BlastClust v2.2.22 (settings: -S 95 -L 0.6 -b F, Altschul et al. 1990) and were reciprocally compared with MegaBlast v2.2.22 (default settings, Morgulis et al. 2008) against the 67,670 *S. aurata* ESTs from NCBI. The 67,670 ESTs from NCBI were also clustered with BlastClust using the same settings as above.

Transcriptome annotation

Those sequences longer than 99bp were annotated in a 3 steps iterative Blast (Altschul et al. 1990; Camacho et al. 2008) approach (Figure 2) aiming to obtain highly informative annotations while reducing computational time. In each step, BLAST v2.2.22 (Camacho et al. 2008) was used locally to compare the sequences against a certain database (BLASTx (settings: -e 1e-5) against NCBI's *D. rerio* RefSeq, BLASTx (settings: -e 1e-3) against SwissProt, and tBLASTx (settings: -e 1e-5) against selected fish sequences from NCBI's nr/nt and EST databases); the XML

results generated were fed into Blast2GO pipe v2.3.5 with default settings to generate DAT files compatible with the GUI version, which was then used to recover the best Blast hit descriptor for each sequence. The results were parsed, and sequences without descriptor, or with a non-informative one, were used in the next Blast step; the ones with an informative descriptor were kept and considered to be well annotated. At the end, a single file was constructed containing the sequences kept in the first two steps and the sequences from the last step.

Functional annotation.

The 44,977 Blast annotated sequences (Table 3) were mapped to Gene Ontology (GO) (Ashburner et al. 2000) terms using the Blast2GO program v2.3.5. (Conesa et al. 2005; Götz et al. 2008). Specific GO terms were selected from the pool of mapped GO terms for each sequence applying an annotation score with a GO weight of 5 and an annotation cutoff of 55. All subsequent analyses were conducted on gene clusters, using one representative sequence of each cluster. Level 2 GO pies for Biological Process, Molecular Function and Cellular Component were drawn using sequence filters of 10. Fisher´s exact test was applied to select all GO terms that were significantly overrepresented in the novel seabream sequences using all 454 seabream annotated gene clusters (i.e. 54,606, Table 2) as reference group. The false discovery rate (FDR) was set at 0.01.

Results

We performed the 454-pyrosequencing of pooled larval samples comprising the first 60 dha. This developmental period covers specific feeding-related anatomical and physiological milestones such as the opening of the mouth, the beginning of external feeding and the digestive functionality, the intestine maturation and development of gastric functionality (Figure 1). A normalized library was constructed from pooled larval total RNA and sequenced in two half-plate regions of a pico titer plate, yielding a total of 309.3 million bases (Mb) distributed in 869,077 reads that passed all filters from 454's default processing pipeline (Table 1). Sequences were further processed with SeqClean (The gene index project; Computational biology and functional genomics laboratory) in order to remove poly-A tails plus primers and adapters used for normalization, resulting in 867,856 clean reads adding up to 288.1 Mb. Those

reads were assembled as a 454 EST project using Mira into 68,289 contigs (consensus sequences), each contig representing a putative transcript. Finally, all contigs were grouped with BlastClust into 54,606 clusters, each cluster representing a putative gene product.

Discovery of novel gilthead seabream transcripts

To assess the comprehensiveness of the newly sequenced transcriptome, the 67,670 gilthead seabream ESTs available at the moment at NCBI were compared against the 68,289 contigs described here using MegaBlast. Since 50,696 of the previously described ESTs (Table 2) had at least one hit among the newly sequenced ESTs, we conclude that we re-sequenced to some extent 74.92 % of all available gilthead seabream NCBI ESTs. Noteworthy, the new sequences covered on average 78.2 % of the NCBI ESTs length according to the best Blast hit. According to our analysis with BlastClust, all NCBI ESTs belong to 50,965 different clusters; 36,488 of those clusters (71.6 %) contained at least one EST that had been fully or partially re-sequenced in our study.

To identify sequences not previously described we used MegaBlast to query the newly sequenced contigs against the 67,670 NCBI ESTs. Noteworthy, only 49.2 % of all gilthead seabream trancriptome sequences described here were found among the already available NCBI ESTs and, according to the best Blast hit, 71.0 % of their length was covered on average. Therefore, we estimate that 34,722 (50.8 %) of the 68,289 gilthead seabream ESTs presented in this study had not been previously described. These novel ESTs belong to 33,408 different clusters (61.2 % of the total).

Annotation and Gene Ontology analysis

The gilthead seabream transcriptome was annotated by submitting the 67,391 sequences ≥100 bp to an iterative Blast search pipeline (Figure 2). A total of 22,633 sequences were annotated when blasted against the *D. rerio* RefSeq protein database, and another 2,024 when blasted against the SwissProt database. All remaining, non annotated sequences were then blasted against teleost NCBI and EST databases, resulting in another 20,320 sequences being annotated. Sequences annotated when blasted against the SwissProt database were also included in this last step. Altogether,

the iterative Blast search resulted in a total of 44,977 annotated sequences that were assigned with the name of the best Blast hit (table 3).

To obtain further information on the sequenced transcripts, Blast results were analyzed with Blast2GO to retrieve Gene Ontology (GO) information. 21,073 sequences were mapped to GO terms, 15,818 of which were finally annotated with specific GO terms (Table 3). GO terms abundance (level 2) was analysed on gene clusters. Terms describing biological processes were most abundant for cellular process (21.90 %), metabolic process (15.61 %), developmental process (12.51 %) and multicellular organismal process (12.28 %) among annotated gene clusters (Figure 3A). Child GO terms for developmental processes were highest for anatomical structure development (5.02 %), multicellular organismal development (4.62 %) and cellular developmental process (2.63 %). GO terms describing molecular functions were highest for binding (45.14 %) and catalytic activity (35.75 %) (Figure 3B). Finally, GO terms describing cellular components were highest for cell, organelle and macromolecular complex (Figure 3C). These results suggest that gilthead seabream larvae undergo important metabolic activities associated to a fast growth and tissue and organ remodelling.

The result of the Fisher´s exact test showed that a high number of newly-described gilthead seabream genes are involved in various developmental processes (Table 4 and Supplementary File 2) such as the formation of the neural system, sensory organs, muscular system, circulatory system, epithelia and other organs and tissues. Other well-represented processes were those involved in morphogenesis and growth. Moreover, biological processes involved in reproduction, behaviour as well as in the response to abiotic and endogenous stimulus were also overrepresented (Table 4 and Supplementary File 2).

Among GO terms describing molecular functions, the following terms were overrepresented in the novel gilthead seabream 454 genes: transducer activity, receptor activity, transporter activity, kinase activity, channel activity, ion transport and DNA binding. GO terms describing cellular components were significantly overrepresented for several issues related to membranes, cell parts, cell envelope, nucleus and ion channel complex (Supplementary File 2). Furthermore, many of these

novel genes belong to gene families with established roles in embryonic and developmental processes, such as homeobox proteins, nuclear receptors, sox genes and forkhead box proteins and genes involved in the Wnt- and BMP-pathways (Table 5 and Supplementary File 3).

Tools for the analysis of the gilthead seabream transcriptome

Several online tools (https://bioinfdata.ccit.ub.edu/apps/) have been implemented to provide easy access to the data generated in this study. The Generic Genome Browser (GBrowse, Stein et al. 2002) is used as a framework to visualize and interrogate the data. GBrowse is an interactive track-based application in which the user can decide which information wants to see at any given moment. For example, for every contig the user can download and see its consensus sequence and all the reads that were assembled to construct it. Features such as the GC content, the 6 frame aminoacidic translation or restriction enzyme cutting regions can also be easily visualized. In addition, the assigned name of each contig, Blast results, and cluster groups have been uploaded. Therefore, the user has various search options, such as for a contig or cluster ID, or for certain gene name or description, to obtain a graphical representation of all matching results before inspecting them individually. Noteworthy, we have also built in a mySQL interface that allows complex queries on most of the fields so that results can be downloaded as text tables. Finally, all contigs can be blasted online through our Blast server with graphical interface, yielding results directly linked to the GBrowse application.

Discussion

To date, only limited genomic and transcriptomic studies on fish embryos and yolk-sac larvae have been conducted in a few teleost fish species: in zebrafish, *Danio rerio*, (Packham et al. 2009; Ton et al. 2002), in gilthead seabream (Ferrareso et al. 2008; Sarropoulou et al. 2005b), in cod (Drivenes et al. 2012; Johansen et al. 2011) and in the marine killifish, *Fundulus heteroclitus*, (Bozinovic et al. 2011). To our knowledge, only the EST collections generated for halibut (Douglas et al. 2007) and Senegalese sole (Infante et al. 2008) included samples of yolk sac, pre- and post eye-migration larvae. The present study, therefore, represents the first transcriptome-wide study covering the complete larval period that has been performed in fish using

454 pyrosequencing technology, from the opening of the mouth and first feeding to the start of the metamorphosis into the juvenile stage.

The use of 454 pyrosequencing is allowing a notable improvement of the genomic resources for teleosts (Coppe et al. 2010; Elmer et al. 2010; Goetz et al. 2010; Jeukens et al. 2010; Johansen et al. 2011; Kristiansson et al. 2009; Liu et al. 2011; Mu et al. 2010; Renaut et al. 2010; Salem et al. 2010; Xiang et al. 2010) and other aquatic animals with economical interest (Huan et al. 2012; Milan et al. 2011). In our study, we have obtained more than 850,000 good quality reads that assembled into 68,289 contigs. The EST sequencing and annotation strategy implemented in the present study has notably enriched the existing transcriptome dataset for the gilthead seabream.

When blasting all assembled seabream 454 sequences against all available *S. aurata* ESTs from NCBI, 33,567 sequences were found to have at least one hit. In other words, more than 50 % of the transcripts identified in the present study were not identified by previous Sanger EST sequencing efforts and, therefore, are not represented in the NCBI database. In contrast, approximately 75 % of *S. aurata* published ESTs mapped to seabream 454 sequences (Table 3), indicating that in the present study we have identified a large proportion of transcripts already present in the NCBI database, therefore validating the approach.

The novel sequences derived from a pool of seabream larvae sampled throughout early developmental stages include an extensive number of genes related to developmental processes. Transcripts with a relevant role in the postembryonic development of practically all tissues and organs in seabream larvae appear well represented in our survey. Interestingly, many of the new sequences are related to the ontogeny of the brain and neuron system, as well as to the development of the sensory organs (Table 4). The nervous system in fish develops during the whole larval period (Nieuwenhuys 2011) and, as in other vertebrates, the development of functional neural circuits is associated to improvement in functionally of organs and tissues with key roles in the early life stages (Gibson and Ma 2011). Particularly important is its involvement in the enhancement of movement capacity and swimming behaviour (McLean et al. 2007), in the development of the olfactory epithelium and proliferation

of olfactory crypt cells and neuromasts (Camacho et al. 2010; Hansen and Zielinski 2005; Sandulescu et al. 2011) and in the development of the eye, retina layers and photosensitive chromophores related to both visual and non-visual functions (Blackshaw et al. 2004; Lamb et al. 2007). In fact, many of the novel sequences belonging to development-related gene families are associated to the above-mentioned processes (Table 5 and Supplementary File 3). In this sense, it is interesting to remark that we found several homeobox genes related to brain development and jaw/bone formation such as the developing brain homeobox (bdx1a) and some distal-less homeobox proteins (dlx2a, dlx4b, dlx5a, dlx6a). We also found sox genes (sox3, Sox5, sox9a, sox11, sox12, sox14a) involved in brain and nervous system development, cell fate and chondrogenesis.

The changes of gene expression profile within the post-embryonic larval period have been examined in European seabass larvae using heterologous (Darias et al. 2008; Mazurais et al. 2011) and custom (Ferrareso et al. 2010) cDNA microarrays, and in Atlantic halibut larvae using a custom microarray (Douglas et al. 2008; Murray et al. 2010). The limited number of sequences derived from early embryonic tissues in the microarray platforms used in these studies underscores the importance of having large genome resources to advance our understanding of the molecular mechanisms driving developmental processes in fish larvae. Thereby, the new sequence resource on the early stages of gilthead seabream that we provide here offers boundless opportunities for genomic research on physiological aspects and represents an excellent tool for understanding those factors and mechanisms that are conditioning the developmental processes and consequently the appropriate development and growth of this species. The gilthead seabream transcriptome will also provide a valuable tool for other sparid species that are being cultivated in temperate and subtropical waters worldwide.

Importantly, we have implemented a framework to easily visualize, interrogate and download the gilthead seabream transcriptome. This kind of bioinformatics tools, often not provided with genome-wide studies, imply that no particular skills are required to profit from the results. Therefore, we believe that the data presented here will help to accelerate research on fish growth and development which, ultimately, will be beneficial for aquaculture development, control and optimization.

Conclusions

The present study is the first using 454 pyrosequencing technology in the gilthead seabream, and one of the few in marine teleosts aiming to increase transcriptomic resources in this species. We have obtained a transcriptome dataset of 293 Mb with a total number of 68,289 contigs. Our work has increased significantly the EST resources available for gilthead seabream and, together with the bioinformatic tools that we are providing, will be particularly useful in advancing in developmental and morphogenesis studies as well as for improving current microarray designs.

Acknowledgments

This research has been funded by the Spanish Ministry of Science and Innovation MICINN + FEDER/ERDF, Consolider Ingenio 2010 Program (Project Aquagenomics, CSD2007-0002). This study also benefited from participation in LARVANET - COST action FA0801 (EU RTD framework Programme). We would like to thank V. Gordo, P. Altube, C. Ligero and A. Godoy for their help in implementing GBrowse and Blast servers.

References

Adzhubei AA, Vlasova AV, Hagen-Larsen H, Ruden TA, Laerdahl JK, Høyheim B (2007) Annotated Expressed Sequence Tags (ESTs) from pre-smolt Atlantic salmon (*Salmo salar*) in a searchable data resource. BMC Genomics 8:209

Alami-Durante H, Olive N, Rouel M (2007) Early thermal history significantly affects the seasonal hyperplastic process occurring in the myotomal white muscle of *Dicentrarchus labrax* juveniles. Cell Tissue Res 327:553-570

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403-410

Apromar (2011) La acuicultura marina de peces en España. Fondo Europeo de Pesca, Ministerio de Medio ambiente y Medio Rural y Marino, Spain. 76 pp

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis

A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25(1):25-29

Barahona-Fernandez MH (1982) Body deformation in hatchery-reared European seabass *Dicentrarchus labrax* (L.). Types, prevalence and effect of fish survival. J Fish Biol 21:239-249

Blackshaw S, Harpavat S, Trimarchi J, Cai L, Huang H, Kuo WP, Weber G, Lee K, Fraioli RE, Cho SH, Yung R, Asch E, Ohno-Machado L, Wong WH, Cepko CL (2004) Genomic Analysis of Mouse Retinal Development. PLoS Biol 2(9):e247

Bozinovic G, Sit TL, Hinton DE, Oleksiak MF (2011) Gene expression throughout a vertebrate's embryogenesis. BMC Genomics 12:132

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL (2008) BLAST+: architecture and applications. BMC Bioinformatics 10:421

Camacho S, Ostos-Garrido MV, Domezain A, Carmona R (2010) Study of the olfactory epithelium in the developing sturgeon. Characterization of the crypt cells. Chem senses 35:147-156

Campinho MA, Galay-Burgos M, Sweeney GE, Power DM (2010) Coordination of deiodinase and thyroid hormone receptor expression during the larval to juvenile transition in sea bream (*Sparus aurata*, Linnaeus). Gen Comp Endocrinol 165:181–194

Cerdà J, Mercadé J, Lozano JJ, Manchado M, Tingaud-Sequeira A, Astola A, Infante C, Halm S, Viñas J, Castellana B, Asensio E, Cañavate P, Martínez-Rodríguez G, Piferrer F, Planas J, Prat F, Yúfera M, Durany O, Subirada F, Rosell E, Maes T (2008) Genomic resources for a commercial flatfish, the Senegalese sole (*Solea senegalensis*): EST sequencing, oligo microarray design, and development of the bioinformatic platform Soleamold. BMC Genomics 9:508

Chevreux B, Wetter T, Suhai S (1999) Genome Sequence Assembly Using Trace Signals and Additional Sequence Information. Computer Science and Biology: Proceedings of the German Conference on Bioinformatics (GCB) 99. pp 45-56

Chini V, Rimoldi S, Terova G, Saroglia M, Rossi F, Bernardini G, Gornati R (2006) EST-based identification of genes expressed in the liver of adult seabass (*Dicentrarchus labrax*, L.). Gene 376:102–106

Computational biology and functional genomics laboratory: Seqclean (http://compbio.dfci.harvard.edu/tgi/software/)

Conesa A, Götz S, García-Gómez J, Terol J, Taron M, Robles M (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinfiormatics 21:3674-3676

Coppe A, Pujolar JM, Maes GE, Larsen PF, Hansen MM, Bernatchez L, Zane L, Bortoluzzi S (2010) Sequencing, de novo annotation and analysis of the first *Anguilla Anguilla* transcriptome: EeelBase opens new perspectives for the study of the critically endangered European eel. BMC Genomics 11:635

Darias MJ, Zambonino-Infante JL, Hugot K, Cahu CL, Mazurais D (2008) Gene expression pattern during the larval development of the European seabass (*Dicentrarchus labrax*) by microarray analysis. Mar Biotechnol 10:416-428

Douglas SE, Knickle LC, Kimball J, Reith ME (2007) Comprenhensive EST analysis of Atlantic halibut (*Hippoglossus hippoglossus*), a commercially relevant aquaculture species. BMC Genomics 8:144

Douglas SE, Knickle LC, Williams J, Flight RM, Reith ME (2008) A first generation Atlantic halibut *Hippoglossus hippoglossus* (L.) microarray: application to developmental studies. J Fish Biol 72:2391-2406

Drivenes Ø, Taranger GL, Edvardsen RB (2012) Gene expression profiling of Atlantic cod (*Gadus morhua*) embryogenesis using microarray. Mar Biotechnol. DOI: 10.1007/s10126-011-9399-y

Elmer KR, Fan S, Gunter HM, Jones JC, Boekhoff S, Kuraku S, Meyer A (2010) Rapid evolution and selection inferred from the transcriptomes of sympatric crater lake cichlid fishes. Mol Ecol 19 (Suppl. 1):197-211

Ferrareso S, Vitulo N, Mininni AM, Romualdi C, Cardazzo B, Negrisolo E, Reinhardt R, Canario AVM, Patarnello T, Bargelloni L (2008) Development and validation of a gene expression oligo microarray for the gilthead sea bream (*Sparus aurata*). BMC Genomics 9:580

Ferrareso S, Milan M, Pellizzari C, Vitulo N, Reinhardt R, Canario AVM, Patarnello T, Bargelloni L (2010) Development of an oligo DNA microarray for the European sea bass and its application to expression profiling of jaw deformity. BMC Genomics 11:354

Gibson DA, Ma L (2011) Developmental regulation of axon branching in the vertebrate nervous system. Development 138:183-195

Goetz FW, Rosauer D, Sitar S, Goetz G, Simchick C, Roberts S, Johnson R, Murphy C, Bronte CR, MacKenzie S (2010) A genetic basis for the phenotypic

differentiation between siscowet and lean lake trout (*Salvelinus namaycush*). Mol Ecol 19:176-196

González SF, Chatziandreou N, Nielsen ME, Li W, Rogers J, Taylor R, Santos Y, Cossins A (2007) Cutaneous immune responses in the common carp detected using transcript analysis. Mol Immunol 44:1664-1679

Götz S, García-Gómez JM, Terol J, Williams TD, Nueda MJ, Robles M, Talón M, Dopazo J, Conesa, A (2008) High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic Acids Res 36(10):3420-3435

Govoroun M, Le Gac F, Guiguen Y (2006) Generation of a large scale repertoire of Expressed Sequence Tags (ESTs) from normalised rainbow trout cDNA libraries. BMC Genomics 7:196

Hagen-Larsen H, Laerdahl JK, Panitz F, Adzhubei A, Høyheim B (2005) An EST-based approach for identifying genes expressed in the intestine and gills of pre-smolt Atlantic salmon (*Salmo salar*). BMC Genomics 6:171

Hansen A, Zielinski BS (2005) Diversity in the olfactory epithelium of bony fishes: development, lamellar arrangement, sensory neuron cell types and transduction components. J Neurocytol 34:183-208

Hayes B, Laerdahl JK, Lien S, Moen T, Berg P, Hindar K, Davidson WS, Koop BF, Adzhubei A, Høyheim B (2007) An extensive resource of single nucleotide polymorphism markers Associated with Atlantic salmo (*Salmo salar*) expressed sequences. Aquaculture 265:82-90

Huan P, Wang H, Liu B. (2012) Transcriptomic analysis of the clam Meretrix meretrix on different larval stages. Mar Biotechnol. DOI: 10.1007/s10126-011-9389-0

Infante C, Asensio E, Cañavate JP, Manchado M (2008) Molecular characterization and expression analysis of five different elongation factors I alpha genes in the flatfish Senegalese sole (*Solea senegalensis* Kaup): Differential gene expression and thyroid hormones dependence during morphogenesis. BMC Molecular Biology 9:19

Jeukens J, Renaut S, St-Cyr J, Nolte AW, Bernatchez L (2010) The transcriptomic of sympatric dwarf and normal lake whitefish (*Coregonus clupeaformis* spp., Salmonidae) divergence as revealed by next-generation sequencing. Mol Ecol 19:5389-5403

Johansen SD, Karlsen BO, Furmanek T, Andreassen M, Jørgensen TE, Bizuayehu TT, Breines R, Emblem Å, Kettunen P, Luukko K, Edvardsen RB, Nordeide JT, Coucheron DH, Moum T (2011) RNA deep sequencing of Atlantic cod transcriptome. Comp Biochem Physiol part D 6:18-22

Johnston IA, Cole NJ, Abercrombie M, Vieira VLA (1998) Embryonic temperature modulates muscle growth characteristics in larval and juvenile herring. J Exp Biol 201:623–646

Johnston IA (2006) Environment and plasticity of myogenesis in teleost fish. J Exp Biol 209:2249-2264

Koumoundouros G, Ashton C, Sfakianakis DG, Divanach P, Kentouri M, Anthwal N, Stickland NC (2009) Thermally-induced phenotypic plasticity of swimming performance in European sea bass *Dicentrarchus labrax* juveniles. J Fish Biol 76:1309-1322

Kristiansson E, Asker N, Förlin L, Larsson DGJ (2009) Characterization of the *Zoarces viviparous* liver transcriptome using massively parallel pyrosequencing. BMC Genomics 10:345

Lamb TD, Collin SP, Pugh EN Jr (2007) Evolution of the vertebrate eye: opsins, photoreceptors, retina and eye cup. Nat Rev Neurosci 8:960-976

Lee BY, Howe AE, Conte MA, D'Cotta H, Pepey E, Baroiller JF, di Palma F, Carleton KL, Kocher TD (2010) An EST resource for tilapia based on 17 normalized libraries and assembly of 116,899 sequence tags. BMC Genomics 11:278

Li P, Peatman E, Wang S, Feng J, He C, Baoprasertkul P, Xu P, Kucuktas H, Nandi S, Somridhivej B, Serapion J, Simmons M, Turan C, Liu L, Muir W, Dunham R, Brady Y, Grizzle J, Liu Z (2007) Towards the ictalurid catfish transcriptome: generation and analysis of 31,215 catfish ESTs. BMC Genomics 8:177

Liu S, Zhou Z, Lu J, Sun F, Wang S, Liu H, Jiang Y, Kucuktas H, Kaltenboeck L, Peatman E, Liu Z (2011) Generation of genome-scale gene-associated SNPs in catfish for the construction of a high-density SNP array. BMC Genomics 12:53

Louro B, Passos ALS, Souche EL, Tsigenopoulos C, Beck A, Lagnel J, Bonhomme F, Cancela L, Cerdà J, Clark MS, Lubzens E, Magoulas A, Planas JV, Volckaert FAM, Reinhardt R, Canario AVM (2010) Gilthead sea bream (*Sparus auratus*) and European sea bass (*Dicentrarchus labrax*) expressed sequence tags:

characterization, tissue-specific expression and gene markers. Mar Genomics 3:179-191

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, Dewell SD, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzik JP, Jando SC, Alenquer ML, Jarvie TP, Jirage KB, Kim JB, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, Mckenna NP, Myers EW, Nickerson E, Nobile JR, Plant R. Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Snirivasan M, Tartaro KR, Tomasz A, Gogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, Rothberg JM (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437(7057):376-380

Mazurais D, Darias M, Zambonino-Infante JL, Cahu CL (2011) Transcriptomics for understanding marine fish larval development. Can J Zool 89:599-611

Milan M, Coppe A, Reinhardt R, Cancela LM, Leite RB, Saavedra C, Ciofi C, Chelazzi G, Patarnello T, Bortoluzzi S, Bargelloni L (2011) Transcriptome sequence and microarray development for the Manila clam, *Rupitapes philippinarum*: genomic tools for environmental monitoring. BMC Genomics 12:234

Morgulis A, Coulouris G, Raytselis Y, Madden TL, Agarwala R, Schäffer AA (2008) Database indexing for production MegaBLAST searches. Bioinformatics 15:1757-1764

Mu Y, Ding F, Cui P, Ao J, Hu S, Chen X (2010) Transcriptome and expression profiling analysis revealed changes of multiple signaling pathways involved in immunity in the large yellow croaker during *Aeromonas hydrophila* infection. BMC Genomics 11:506

McLean D, Fan J, Higashima SI, Hale M, Fetcho JR (2007) A topographic map of recruitment in spinal cord. Nature 446:71-75

Murray HM, Lall SP, Rajaselvam R, Boutilier LA, Flight RM, Blanchard B, Colombo S, Mohindra V, Yúfera M, Douglas SE (2010) Effect of early introduction of microencapsulated diet to larval Atlantic halibut, *Hippoglossus hippoglossus*, L. assessed by microarray analysis. Mar Biotechnol 12:214-229

Nieuwenhuys R (2011) The development and general morphology of telescephalon of acnitopterygian fishes: synopsis, documentation and commentary. Brain Struct Funct 215:141-157

Packham IM, Grey C, Heath PR, Hellewell PG, Ingham PW, Crossman DC, Milo M, Chico TJA (2009) Microarray profiling reveals CXCR4a is downregulated by blood flow in vivo and mediates collateral formation in zebrafish embryos. Physiol Genomics 38:319-327

Pardo BG, Fernández C, Millán A, Bouza C, Vázquez-López A, Vera M, Alvarez-Dios JA, Calaza M, Gómez-Tato A, Vázquez M, Cabaleiro S, Magariños B, Lemos ML, Leiro JM, Martínez P (2008) Expressed sequence tags (ESTs) from immune tissues of turbot (*Scophthalmus maximus*) challenged by pathogens. BMC Vet Res 4:37

Pavlidis M, Mylonas CC (eds.) (2011) Sparidae. Biology and aquaculture of gilthead seabream and other species, Wiley-Blackwell, Oxford. UK. 390 pp

Polo A, Yúfera M, Pascual E (1991) Effect of temperature on egg and larval development of *Sparus aurata* L. Aquaculture 92:367-375

Polo A, Yúfera M, Pascual E (1992) Feeding and growth of gilthead seabream (*Sparus aurata* L.) larvae in relation to the size of the rotifer strain used as food. Aquaculture 103:45-54

Power DM, Einarsdóttir IE, Pittman K, Sweeney GE, Hildahl J, Campinho MA, Silva N, Sæle Ø, Galay-Burgos M, Smáradóttir H, Björnsson BT (2008) The molecular and endocrine basis of flatfish metamorphosis. Rev Fish Sci 16(S1):95-111

Renaut S, Nolte AW, Bernatchez L (2010) Mining transcriptome sequences towards identifying adaptive single nucleotide polymorphisms in lake whitefish species pairs (*Coregonus* spp. Salmonidae). Mol Ecol 19:115-131

Salem M, Rexroad III CE, Wang J, Thorgaard GH, Yao J (2010) Characterization of the rainbow trout transcriptome using Sanger and 454-pyrosequencing approaches. BMC Genomics 11:564

Sánchez CC, Smith TPL, Wiedman RT, Vallejo RL, Salem M, Yao J, Redroad III CE (2009) Single nucleotide polymorphism discovery in rainbow trout by deep sequencing of a reduce representation library. BMC Genomics 10:559

Sandulescu C, Teow RY, Hale ME, Zhang C (2011) Onset and dynamic expression of S100 proteins in the olfactory organ and the lateral line system in zebrafish development. Brain Res 1383:120-127

Sarropoulou E, Power DM, Magoulas A, Geisler R, Kotoulas G (2005a) Comparative analysis and characterization of expressed sequence tags in gilthead sea bream (*Sparus aurata*) liver and embryos. Aquaculture 243:69-81

Sarropoulou E, Kotoulas G, Power DM, Geisler R (2005b) Gene expression profiling of gilthead seabream during early development and detection of stress-related genes by the application of cDNA microarray technology. Physiol Genomics 23:182-191

Sha Z, Wang S, Zhuang Z, Wang Q, Wang Q, Li P, Ding H, Wang N, Liu Z, Chen S (2010) Generation and analysis of 10 000 ESRs from the half-smooth tongue sole *Cynoglossus semilaevis* and identification of microsatellite and SNP markers. J Fish Biol 76:1190-1204

Shagin DA, Rebrikov DV, Kozhemyako VB, Altshuler IM, Shcheglov AS, Zhulidov PA, Bogdanova EA, Staroverov DB, Rasskazov VA, Lukyanov S (2002) A novel method for SNP detection using a new duplex-specific nuclease from crab hepatopancreas. Genome Res 12(12):1935-42

Stein LD, Mungall C, Shu SQ, Caudy M, Mangone M, Day A, Nickerson E, Stajich JE, Harris TW, Arva A, Lewis S (2002) The generic genome browser: A building block for a model organism system database. Genome Res 12:1599-1610

Ton C, Stamatiou D, Dzau VJ, Liew CC (2002) Construction of a zebrafish cDNA microarray: gene expression profiling of the zebrafish during development. Biochem Biophys Res Comm 296:1134-1142

Villeneuve L, Gisbert E, Zambonino-Infante JL, Quazuguel P, Cahu CL (2005) Effect of nature of dietary lipids on European sea bass morphogenesis: implication of retinoid receptors. British J Nutr 94:877-884

Wang S, Peatman E, Abernathy J, Waldbieser G, Lindquist E, Richardson P, Lucas S, Wang M, Li P, Thimmapuram J, Liu L, Vullaganti D, Kucuktas H, Murdock C, Small BC, Wilson M, Liu H, Jiang Y, Lee Y, Chen F, Lu J, Wang W, P Xu, Somridhivej B, Baoprasertkul P, Quilang J, Sha Z, Bao B, Wang Y, Wang Q, Takano T, Nandi S, Liu S, Wong L, Kaltenboeck L, Quiniou S, Bengten E, Miller N, Trant J, Rokhsar D, Liu Z (2010) Assembly of 500,000 inter-specific catfish expressed sequence tags and large scale gene-associated marker development for whole genome association studies. Genome Biol 11:R8

Xia JH, Yue GH (2010) Identification and analysis of immune-related transcriptome in Asian seabass *Lates calcarifer*. BMC Genomics 11:356

Xiang LX, He D, Dong WR, Zhang YW, Shao JZ (2010) Deep sequencing-based transcriptome profiling analysis of bacteria-challenged *Lateolabrax japonicus* reveals insight into the immune-relevant genes in marine fish. BMC Genomics 11:472

Yúfera M, Darias MJ (2007) The onset of feeding in marine fish larvae. Aquaculture 268:53-63

Yúfera M, Pascual E, Polo A, Sarasquete MC (1993) Effect of starvation on the feeding ability of the gilthead seabream (*Sparus aurata* L.) larvae at first feeding. J Exp Mar Biol Ecol 169:252-272

Yúfera M (2011) Feeding behaviour in larval fish. In: Larval Fish Nutrition, Holt GJ (ed) Ames, Iowa USA: Wiley-Blackwell, pp 285-305

Yúfera M, Conceição LEC, Battaglene S, Fushimi H, Kotani T (2011) Early Development and Metabolism. In: Sparidae. Biology and aquaculture of gilthead seabream and other species, Pavlidis M, Mylonas CC (eds) Oxford UK: Wiley-Blackwell, pp 133-168

Zapata A, Diez B, Cejalvo T, Gutiérrez-de Frías T, Cortés A (2006) Ontogeny of the immune system of fish. Fish Shellfish Immunol 20:126-136

Zhulidov PA, Bogdanova EA, Shcheglov AS, Vagner LL, Khaspekov GL, Kozhemyako VB, Matz MV, Meleshkevitch E, Moroz LL, Lukyanov SA, Shagin DA (2004) Simple cDNA normalization using Kamchatka crab duplex-specific nuclease. Nucleic Acids Res 32(3):e37

Figure captions

Fig. 1 Gilthead seabream early growth and main developmental events at 20 ºC of water temperature. Triangles indicate timepoints when larvae were sampled, in days after hatching (dha)

Fig. 2 Annotation strategy and Gene Ontology analysis. The gilthead seabream transcriptome was annotated by submitting the 67,391 sequences ≥100 bp to an iterative Blast search pipeline. Annotated sequences were analyzed with Blast2go to retrieve Gene Ontology (GO) information

Fig. 3 Gene ontology terms (level 2) of 454 gilthead seabream gene clusters (%). A: Biological process; B: molecular function; C: Cellular components. For developmental process, level 3 categories are also depicted

Additional files

Supplementary File 1 List of contigs with low complexity sequence or matching Artemia or Brachionus NCBI nucleotidic sequences. Contigs with length < 100 bp are shown in red. LC means Low Complexity. Details column indicates contig's start and end position of string matching with the sequence shown.

Supplementary File 2 Complete list of GO terms for “Biological Process” (Section 1), “Molecular Function” (Section 2) and “Cellular Component” (Section 3) significantly overrepresented (false discovery rate FDR <0.05) in novel gilthead seabream gene clusters when compared to all seabream 454 gene clusters. In each section, column A contains the GO identification number, column B the GO term, and column C the FDR

Supplementary File 3. Full list of novel gilthead seabream transcripts involved in differentiation and development. Each section contains genes of the following gene families/regulatory pathways: Section 1: forkhead box proteins; Section 2: homeobox proteins; Section 3: sox proteins; Section 4: nuclear receptors; Section 5: genes involved in the Wnt-pathway; Section 6: genes involved in the Bmp-pathway. Columns contain the following information: Column A: sequence ID; column B: abbreviated Blast description; column C: gene symbol; column D: accession number of best Blast hit; Column E: sequence length; Column F: E-value of best Blast hit; column G: full description of best Blast hit

Table 1 Summary statistics of 454 sequencing and assembly

454 Sequencing	
Total number of wells	2,229,669
Before Seqclean processing	
Number of reads	869,077
Number of bases in reads (Mb)	309.3
Average length of reads (bp)	355.9
After Seqclean processing	
Number of cleaned reads	867,856
Number of bases in cleaned reads (Mb)	288.1
Average length of cleaned reads (bp)	332.0
Mira assembly	
Number of contigs	68,289
Total number of bases in contigs (Mb)	40.7
Number of contigs >500bp	36,470
Average length of contigs (bp)	596

Table 2 Reciprocal Megablast between *S. aurata*'s NCBI EST collection (Saurata NCBI) and the *S. aurata* contigs obtained by 454 sequencing (Saurata 454). First row shows the results obtained by using Saurata NCBI as query and Saurata 454 as database. Second row shows the results obtanied by using Saurata 454 as query and Saurata NCBI as database

	Total number of queries (corresponding number of clusters)	Number of queries with ≥1 hit (corresponding number of clusters)	Percentage of queries with ≥ 1 hit (corresponding percentage of clusters)	Percentage mean query coverage	Percentage mean maximal identity	Mean E-value
Saurata NCBI query vs Saurata 454 database	67670 (50965)	50696 (36488)	74.92 (71.59)	78.2	97.5	2.7304E-09
Saurata_454 query vs Saurata NCBI database	68289 (54606)	33567 (21198)	49.15 (38.82)	71.0	96.8	3.37209E-09

Table 3 Blast, mapping and annotation results for contigs ≥ 100 bp

Total number of contigs ≥ 100 bp submitted to Blast searches	67,391
Number of contigs with/without a Blast hit	44,977/22,414
Number of contigs with/without GO mapping	21,073/23,904
Number of contigs with/without GO annotation	15,818/5,255

Table 4 Representative GO terms for 'Biological Process' significantly overrepresented (FDR <0.05) in novel seabream EST sequences when compared to all seabream 454 sequences. The complete list can be found in Supplementary File 2.

GO Term	Name	FDR
GO:0007399	nervous system development	2.80E-45
GO:0007275	multicellular organismal development	4.90E-44
GO:0048731	system development	4.90E-44
GO:0022008	neurogenesis	5.30E-36
GO:0032502	developmental process	8.50E-33
GO:0048856	anatomical structure development	2.00E-31
GO:0000902	cell morphogenesis	3.40E-27
GO:0048513	organ development	1.80E-25
GO:0007409	axonogenesis	4.50E-23
GO:0007420	brain development	4.40E-17
GO:0048598	embryonic morphogenesis	7.10E-14
GO:0007411	axon guidance	5.40E-13
GO:0007423	sensory organ development	6.80E-13
GO:0009790	embryonic development	9.90E-13
GO:0050793	regulation of developmental process	6.30E-10
GO:0001654	eye development	5.50E-09
GO:0003002	regionalization	6.50E-09
GO:0048568	embryonic organ development	1.10E-07
GO:0032774	RNA biosynthetic process	1.90E-07
GO:0009888	tissue development	2.30E-07
GO:0035270	endocrine system development	8.00E-07
GO:0001558	regulation of cell growth	9.80E-07
GO:0021510	spinal cord development	3.60E-06
GO:0001944	vasculature development	6.40E-06
GO:0040007	growth	7.10E-06
GO:0031016	pancreas development	3.20E-05
GO:0060429	epithelium development	3.30E-05
GO:0034645	cellular macromolecule biosynthetic process	6.30E-05
GO:0001568	blood vessel development	1.00E-04
GO:0048732	gland development	1.30E-04
GO:0006941	striated muscle contraction	1.60E-04
GO:0003009	skeletal muscle contraction	1.60E-04
GO:0048066	developmental pigmentation	4.10E-04
GO:0008283	cell proliferation	6.00E-04
GO:0048736	appendage development	8.60E-04
GO:0007276	gamete generation	8.90E-04
GO:0007369	gastrulation	1.40E-03
GO:0009855	determination of bilateral symmetry	1.60E-03
GO:0007398	ectoderm development	1.70E-03

GO:0035239	tube morphogenesis	1.70E-03
GO:0090066	regulation of anatomical structure size	1.80E-03
GO:0048925	lateral line system development	2.00E-03
GO:0001704	formation of primary germ layer	2.00E-03
GO:0043062	extracellular structure organization	2.30E-03
GO:0048880	sensory system development	2.60E-03
GO:0043583	ear development	3.30E-03
GO:0048882	lateral line development	3.40E-03
GO:0050931	pigment cell differentiation	3.50E-03
GO:0035050	embryonic heart tube development	7.10E-03
GO:0002520	immune system development	9.60E-03
GO:0048534	hemopoietic or lymphoid organ development	9.60E-03
GO:0009913	epidermal cell differentiation	1.20E-02
GO:0021761	limbic system development	1.20E-02
GO:0042471	ear morphogenesis	1.80E-02
GO:0033333	fin development	1.90E-02
GO:0006468	protein amino acid phosphorylation	2.00E-02
GO:0048592	eye morphogenesis	2.70E-02
GO:0009952	anterior/posterior pattern formation	3.40E-02
GO:0009953	dorsal/ventral pattern formation	3.60E-02
GO:0001525	angiogenesis	3.90E-02

Table 5 Selected transcripts with a relevant role in the postembryonic development of gilthead seabream larvae belonging to forkhead box protein, sox protein, homeobox protein and nuclear receptors gene families. Many of the new sequences are related to the ontogeny of the brain and neuron system, as well as to the development of the sensory organs. The full list including Wnt- and Bmp-pathways can be found in Supplementary File 3.

Sequence Name	Gene Name	Hit Acc. No.	Sequence length	Blast
forkhead box proteins				
forkhead box protein p4	foxp4	XP_685353	659	7.40E-73
forkhead box protein p2	foxp2	NP_001025253	585	1.96E-67
forkhead box protein q1a	foxq1a	NP_001077284	824	1.72E-26
forkhead box protein c1a	foxc1a	NP_571803	330	1.43E-21
forkhead box protein o3a	foxo3a	NP_001009988	556	2.99E-06
Sox genes				
Transcription factor Sox 3	sox3	NP_001001811	1162	4.12E-147
Transcription factor Sox 9	sox9a	NP_571718	729	2.38E-92
Transcription factor Sox12	sox12	AY277960	806	4.71E-87
Transcription factor Sox 11b	sox11b	NP_571412	479	1.56E-44
Transcription factor Sox5	sox5	AY277973	537	1.43E-43
Transcription factor Sox 11a	sox11a	NP_571411	601	3.84E-34
Transcription factor Sox14a	sox14a	AY277955	151	5.50E-08
homeobox proteins				
homeobox protein dlx5a	dlx5a	NP_571381	961	1.00E-124
homeobox protein hox-c4a	hoxc4a	NP_571197	1275	2.11E-115
gs homeobox 2	gsx2	NP_001124196	984	7.23E-102
distal-less homeobox gene 2a-like	Loc100329608	XP_002663348	1072	8.24E-86
lim homeobox protein lhx9 isoform1	lhx9	NP_001017710	683	1.13E-83
homeobox protein dlx6a	dlx6a	NP_571398	849	2.82E-80
homeo box c12a	hoxc12a	NP_001104229	754	7.79E-72
homeobox even-skipped homolog protein 2	evx2	NP_571307	458	1.67E-69
iroquois-class homeodomain protein irx-5	irx5	NP_001038692	506	6.50E-66
lim homeobox protein lhx1	lhx1	NP_571291	355	3.66E-65
pituitary homeobox 1	pitx1	NP_001035436	484	4.85E-59
pou class transcription factor 1	pou1	NP_571236	595	8.21E-53
zinc fingers and homeoboxes 3	zhx3	NP_942108	568	1.42E-52
homeobox protein hmx2	hmx2	NP_001108570	425	1.86E-51
pre-b-cell leukemia homeobox 1	pbx1	NP_001077322	556	1.94E-51
paired box gene 7b	pax7b	NP_001139621	509	1.00E-42
pre-b-cell leukemia homeobox 4	pbx4	NP_571522	495	5.12E-41
homeobox protein orthopedia b	otpb	NP_571175	1270	5.82E-41
iroquois homeobox protein 3a	irx3a	NP_571342	574	2.58E-40
sine oculis homeobox homolog 1a	six1a	NP_001009904	806	3.03E-40
nuclear receptors				
nuclear receptor subfamily 4 group a member 3	nr4a3	NP_001166100	788	2.44E-88
nuclear receptor subfamily 0 group B member 2	nr0b2a	XP_001921455	1225	7.94E-88
nuclear receptor subfamily 5 group a member 5 isoform 1	nr5a5a	NP_999944	970	2.28E-76
peroxisome proliferator-activated receptor alpha b	pparab	NP_001096037	441	1.41E-62
estrogen receptor 2a	esr2a/nr3a2-b	NP_851297	429	6.19E-57
nuclear receptor subfamily 4 group a member 1	nr4a1	NP_001002173	761	3.98E-55
nuclear receptor subfamily 5 group member 5 isoform 2	nr5a5b	NP_001070740	748	6.03E-53
nuclear receptor subfamily 1 group d member 2b	nr1d2b	NP_571140	966	3.31E-51
hepatocyte nuclear factor 4-alpha	hnf4a/nr2a1	NP_919349	493	7.05E-43
nuclear receptor subfamily 4 group a member 2	nr4a2	NP_001106956	692	3.19E-41
RAR-related orphan receptor A, paralog a	roraa/nr1f1-b	NP_001103637	430	9.15E-40
RAR-related orphan receptor c a	rorca/nr1f3-a	NP_001076288	283	3.49E-36
peroxisome proliferator-activated receptor gamma	pparg/nr1c3	NP_571542	254	7.31E-34
estrogen-related receptor gamma	esrrga/nr3b3	NP_998119	669	1.42E-22

hepatocyte nuclear factor 4-gamma	hnf4g/nr2a2	NP_001068579	449	1.90E-22
nuclear receptor subfamily 2 group f member 1-a	nr2f1a/couptfalpha-A	NP_571255	414	6.60E-21
estrogen-related receptor beta	esrrb/nr3b2	XP_001333980	736	2.24E-19
retinoic acid receptor alpha-b	rarab/nr1b1-b-B	NP_571474	575	4.29E-19
RAR-related orphan receptor c b	rorcb/nr1f3-b	XP_690743	853	4.01E-18
coup transcription factor 2	nr2f2/couptfbeta	NP_571258	406	4.14E-17

FIGURE 1

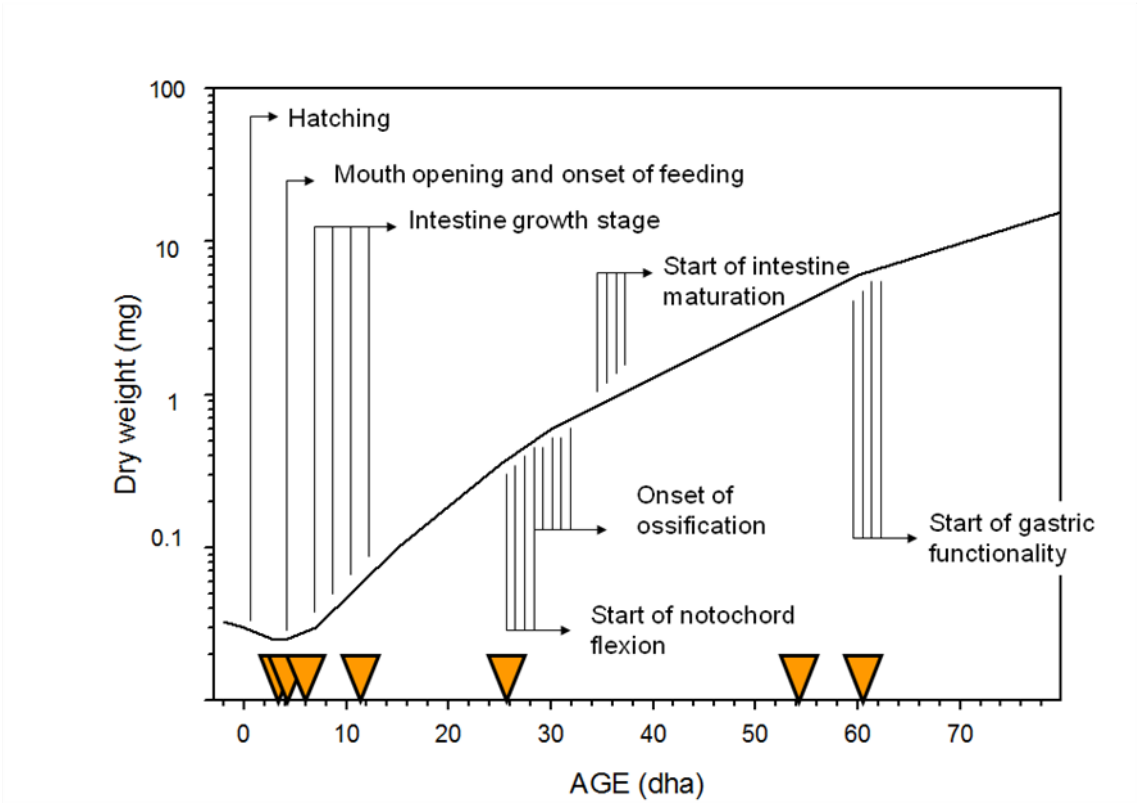
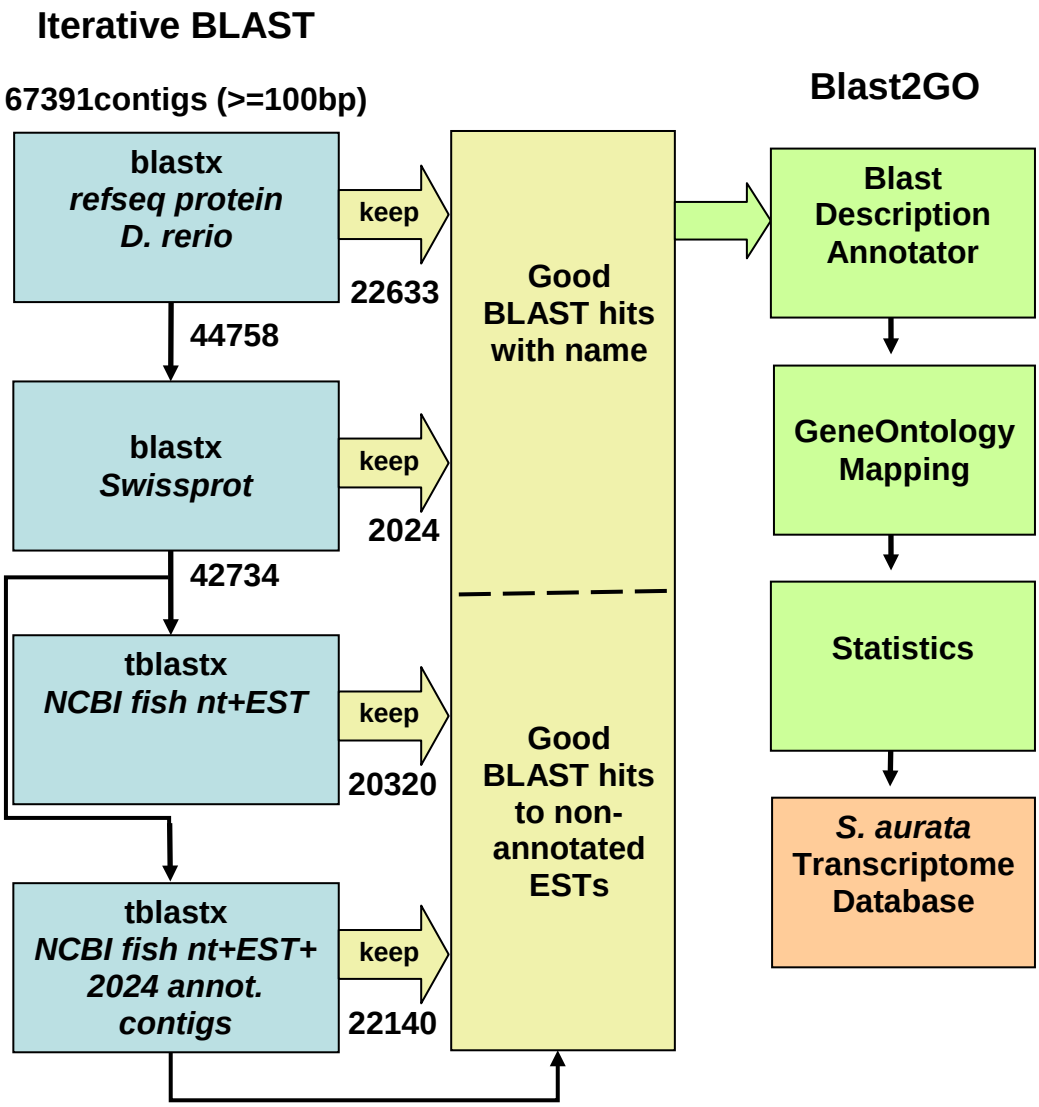
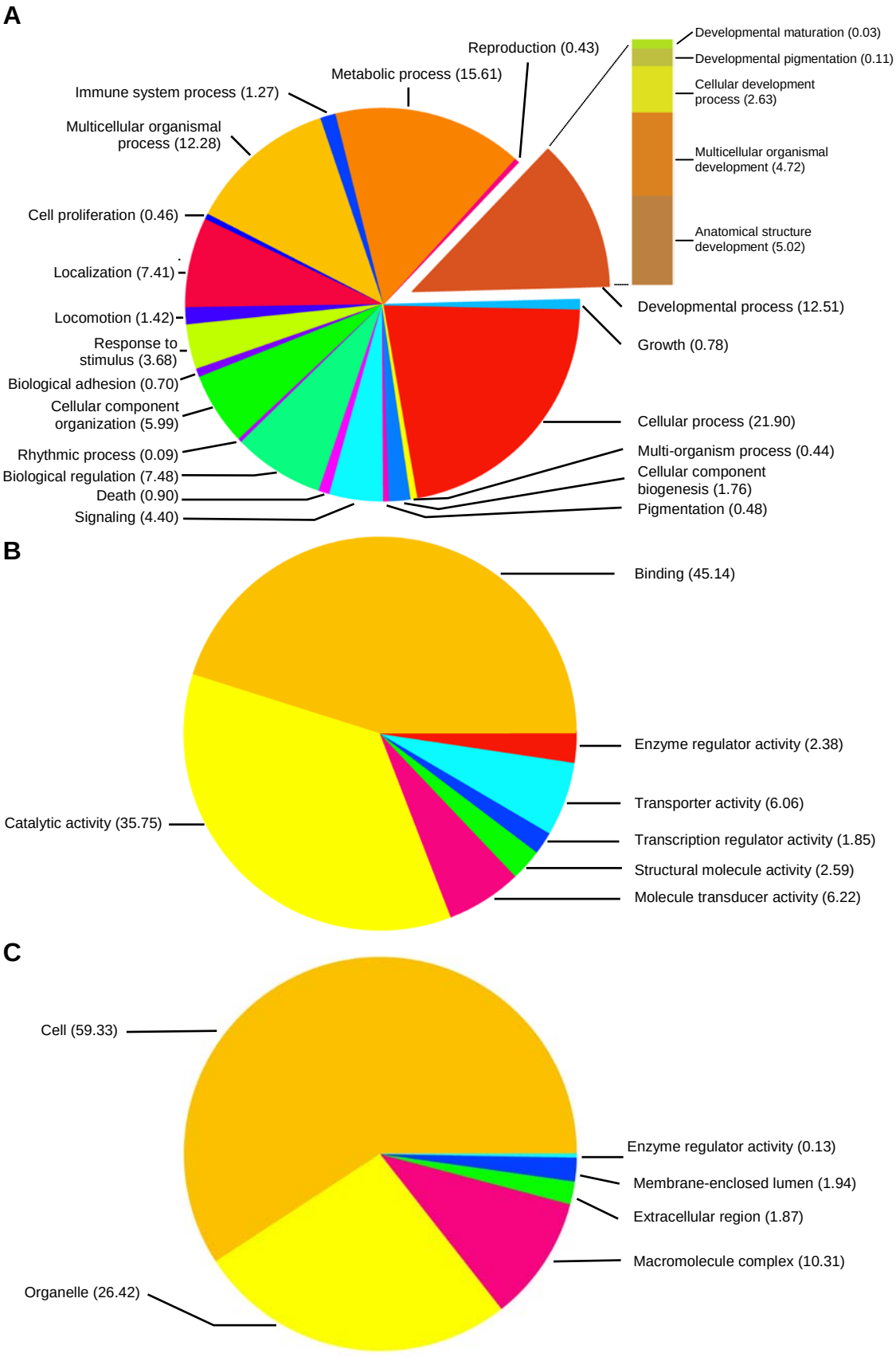

100
10
1
0.1
Dry weight (mg)
Hatching
Mouth opening and onset of feeding
Intestine growth stage
Start of intestine maturation
Onset of ossification
Start of gastric functionality
Start of notochord flexion
0
10
20
30
40
50
60
70
AGE (dha)

FIGURE 2


Iterative BLAST
67391contigs (>=100bp)
blastx
refseq protein
D. rerio
keep
22633
44758
blastx
Swissprot
keep
2024
42734
tblastx
NCBI fish nt+EST
keep
20320
tblastx
NCBI fish nt+EST+
2024 annot.
contigs
keep
22140
Good
BLAST hits
with name
Good
BLAST hits
to non-
annotated
ESTs
Blast2GO
Blast
Description
Annotator
GeneOntology
Mapping
Statistics
S. aurata
Transcriptome
Database


A
Reproduction (0.43)
Metabolic process (15.61)
Immune system process (1.27)
Multicellular organismal process (12.28)
Cell proliferation (0.46)
Localization (7.41)
Locomotion (1.42)
Response to stimulus (3.68)
Biological adhesion (0.70)
Cellular component organization (5.99)
Rhythmic process (0.09)
Biological regulation (7.48)
Death (0.90)
Signaling (4.40)
Developmental maturation (0.03)
Developmental pigmentation (0.11)
Cellular development process (2.63)
Multicellular organismal development (4.72)
Anatomical structure development (5.02)
Developmental process (12.51)
Growth (0.78)
Cellular process (21.90)
Multi-organism process (0.44)
Cellular component biogenesis (1.76)
Pigmentation (0.48)
B
Binding (45.14)
Enzyme regulator activity (2.38)
Transporter activity (6.06)
Transcription regulator activity (1.85)
Structural molecule activity (2.59)
Molecule transducer activity (6.22)
Catalytic activity (35.75)
C
Cell (59.33)
Enzyme regulator activity (0.13)
Membrane-enclosed lumen (1.94)
Extracellular region (1.87)
Macromolecule complex (10.31)
Organelle (26.42)