Comparative Biochemistry and Physiology, Part D 6 (2011) 23–30

Contents lists available at ScienceDirect

Comparative Biochemistry and Physiology, Part D

journal homepage: www.elsevier.com/locate/cbpd

EST resources and establishment and validation of a 16 k cDNA microarray from Atlantic cod (*Gadus morhua*)☆

Rolf B. Edvardsen [a,*], Ketil Malde [a], Christian Mittelholzer [a,1], Geir Lasse Taranger [a], Frank Nilsen [a,b]

[a] Institute of Marine Research, PO Box 1870, Nordnes N-5817 Bergen, Norway
[b] Department of Biology, University of Bergen, N-5020 Bergen, Norway

ARTICLE INFO

Article history:
Received 4 March 2010
Received in revised form 21 June 2010
Accepted 22 June 2010
Available online 28 June 2010

Keywords:
Annotation
Aquaculture
cDNA libraries
Gene expression
Marine environment
MHC
Microarray

ABSTRACT

The Atlantic cod, *Gadus morhua*, is an important species both for traditional fishery and increasingly also in fish farming. The Atlantic cod is also under potential threat from various environmental changes such as pollution and climate change, but the biological impact of such changes are not well known, in particular when it comes to sublethal effects that can be difficult to assert. Modern molecular and genomic approaches have revolutionized biological research during the last decade, and offer new avenues to study biological functions and e.g. the impact of anthropogenic activities at different life-stages for a given organism. In order to develop genomic data and genomic tools for Atlantic cod we conducted a program were we constructed 20 cDNA libraries, and produced and analyzed 44006 expressed sequence tags (ESTs) from these. Several tissues are represented in the multiple cDNA libraries, that differ in either sexual maturation or immulogical stimulation. This approach allowed us to identify genes that are expressed in particular tissues, life-stages or in response to specific stimuli, and also gives us information about potential functions of the transcripts. The ESTs were used to construct a 16 k cDNA microarray to further investigate the cod transcriptome. Microarray analyses were preformed on pylorus, pituitary gland, spleen and testis of sexually maturing male cod. The four different tissues displayed tissue specific transcriptomes demonstrating that the cDNA array is working as expected and will prove to be a powerful tool in further experiments.

1. Introduction

The Atlantic cod, *Gadus morhua*, is a marine teleost belonging to the family Gadidae. The Atlantic cod consists of several different stocks where the Northeast Artic stock is the largest, and it has been an important species in the fisheries for a very long time. More recently, Atlantic cod has been regarded as one of the new promising species for marine aquaculture.

The introduction of cod into aquaculture has revealed several significant challenges compromising profitable and environmentally sustainable production of this species. This includes susceptibility to disease and stress, skeletal deformations (Fjelldal et al., 2009), larval mortality and control of sexual maturation (Taranger et al., 2006, 2010). Early sexual maturation and spawning of cod in the sea cages can also lead to release of fertilized eggs into the environment, thereby potentially affecting the genetic composition of wild cod stocks (Jorstad et al., 2008).

The development of new molecular and genomic tools and methods will play an important role in solving many of the problems that cod aquaculture is currently facing. Furthermore, the functional genomics approach offers great opportunities to study the interactions between genes and environment, which has particular importance in the face of increasing anthropogenic activity that may affect the living environment for Atlantic cod such as off-shore petroleum exploration, transport, as well as ocean climate changes. The genomic approaches will also be of great importance to investigate the implications of the resent findings of the separation into a range of genetically distinct local populations of Atlantic cod (e.g. Nielsen et al., 2009), e.g. along the Norwegian coast, as well as to understand implications of evolutionary effects of fisheries on important life-history traits such as size and age at sexual maturity (e.g. Olsen et al., 2005; Dieckmann and Heino, 2007) and the underlying mechanisms.

Resources such as ESTs, microarrays and whole genome sequencing projects exist for a number of fish species such as the zebrafish, fugu and stickleback, but a relatively small fraction of the data is from marine, cold-water species. With respect to farmed fish several recent sequencing efforts has produced large amounts of data, especially ESTs, for both Atlantic salmon (Adzhubei et al., 2007), Atlantic halibut

☆ This paper stems from a presentation at the Genomics in Aquaculture symposium held at Bodø on 5th–7th July 2009. This was the first international meeting devoted exclusively to this field and it was funded by the Research Council of Norway (grant 192126/S40).

* Corresponding author. Institute of Marine Research, Nordnesgaten 50, P.O. Box 1870 Nordnes, N-5817 Bergen, Norway. Tel.: +47 55 90 65 05; fax: +47 55 23 85 31.
E-mail address: rolfbe@imr.no (R.B. Edvardsen).
[1] Current address: University of Basel, Klingelbergstrasse 50/70, CH-4056 Basel, Switzerland.

1744-117X/$ – see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.cbd.2010.06.002

(Douglas et al., 2007) and other species including the Atlantic cod for which about 200,000 ESTs are available by February 2010.

There are at present five ray-finned fish genomes sequenced from model species and made public available (Ensembl. http://www.ensembl.org/index.html, National Center for Biotechnology Information. http://www.ncbi.nlm.nih.gov). This is an excellent resource for knowledge and a facilitator for functional genomics. While the more aquaculture important species are lagging behind in this aspect, several genome sequencing projects are ongoing (Quinn et al., 2008; Sarropoulou et al., 2008). The cod genome is currently also being sequenced primarily using the 454 sequencing technology for which the EST data presented here will be an important asset in the annotation and assembly process (The Cod Genome Project http://www.codgenome.no, Johansen et al., 2009). This project is well advanced and the 0.9 Gb genome is scheduled to be fully sequenced within 2010.

EST sequences have a large range of applications. Parts of the present cod EST dataset have been used for identification of genetic markers like microsatellites (e.g. (Westgaard et al., 2007; Delghandi et al., 2008a,b,c; Delghandi et al., 2009)) and single nucleotide polymorphs (SNPs) (Moen et al., 2008) and these genetic markers have been utilized to construct the first genetic map for the Atlantic cod (Moen et al., 2009). Even if new sequencing technologies now enable higher throughput and reduced costs classical Sanger sequences are valuable due to long read lengths and well know quality properties. Data on gene transcription obtained by microarrays can be widely applied depending on experimental set up and questions asked. An example is Whitehead and Crawford (2006) where they used a microarray approach to investigate tissue specific transcription between populations of *Fundulus heteroclitus*.

In the present study we have produced 44,006 ESTs from 20 cDNA libraries and given an analysis of the sequences. The cDNA clones were utilized to construct a 16 k microarray. The microarray was validated to confirm its reproducibility and accuracy on four different tissues from sexually maturing male cod.

2. Materials and methods

2.1. cDNA library construction and EST sequencing

The list of libraries constructed is listed in Table 1 (details of sampling are in supplementary Table S1). Total RNA for library construction was extracted by Trizol (Sigma) and by RNeasy Mini Kit (Qiagen) followed by an enrichment of polyA RNA by Poly(A) Pure™ or Poly(A) Purist™ (Ambion) or Oligotex (Qiagen)respectively. The cDNA libraries were constructed directly in the pBluscript SK + vector as described by the manufacturer (Stratagene). All clones were blue-white screened and white clones were picked randomly from all the different libraries for plasmid purification. Bacteria were grown overnight in 96 well dishes (Millipore) and plasmids were purified according to the recommendations from the manufacturer (Millipore). Clones were sequenced using vector primer m13r or m13f and BigDye chemistry (Applied Biosystems). Sequence data from this study has been submitted to GenBank with accession numbers GW841082 to GW863522.

2.2. Trimming, clustering, and assembly

The ESTs were first masked for vector using the UniVec database cross_match (P. Green, unpublished), and for *E. coli* contamination using RBR (Malde et al., 2006). The sequences were then masked for low quality, removing bases from each end of the EST until quality score exceeded 10, and sliding average quality (window size 20) exceeded 13. Finally, SeqClean (Lee et al., 2005) was used to trim the sequences of the masked parts, and to remove sequences that were shorter than 100 bases after trimming. The resulting sequences were then processed using the TGICL (Pertea et al., 2003) pipeline, which performs clustering with megaBLAST (Zhang et al., 2000) and contig assembly using CAP3 (Huang and Madan, 1999).

Table 1
Overview of the cDNA libraries produced, number of transcripts from each library and number of probes from these libraries represented on the microarray.

Library	Description	N of ESTs	%-of total	N of transcipts	Probes on array
CHO	Head kidney	2919	6.40	950	1824
CLE	Liver	1365	3.00	413	768
CHY	Pituitary	2552	5.60	1385	1824
CLU	Liver, immature	1920	4.20	1000	0
CEG	3–6 dpf	1912	4.20	821	0
CPY	Pylorus	2804	6.20	789	1920
CEM	Embryo, 1–9 dpf	6641	14.60	2504	2976
CTE	Testis	2647	5.80	1368	1632
CHJ	Brain	1695	3.70	872	1248
COV	Ovary	2348	5.20	1353	1728
CMI	Spleen	2190	4.80	1191	0
CSMI	Stimulated spleen	1644	3.60	928	1152
CTA	Column	2797	6.20	1365	576
CGG	Beard	776	1.70	567	0
CHU	Brain, immature	1632	3.60	1239	0
CSH	Stimulated head kidney	2016	4.40	1178	0
CTH	Thymus	1912	4.20	1417	0
COY	Eye	1152	2.50	751	0
CNN	Skin	1152	2.50	798	0
CLF	Larvae before feeding	1932	4.30	1153	0
Total		44,006	100	22,042	

The contigs and singletons were annotated using BLASTX (E-value of 10^{-7} and a word size of 4) against the UniRef90 protein database (Suzek et al., 2007), and the matches were used with the GOA database (Barrell et al., 2009) to assign GO (Gene Ontology) terms to each contig or singleton.

2.3. Fish samples for array validation

The four male fish (*G. morhua*, Gadidae) originated from four different families of Norwegian coastal cod strains from Tysfjord and the Porsanger fjord (Dahle et al., 2006), and were reared in a common garden setup from hatching at the IMR marine station at Austevoll. The fish were all from the 2004 year class and sampled in November 2006 when they all were approaching their first sexual maturation (expected to spawn in February 2007). The fish weighed from 1.8 to 2.2 kg, with a condition factor of about 1.2, and displayed no sign of infections or damage.

2.4. Microarray

2.4.1. Microarray construction

Probes were amplified from the individual cDNA clones by PCR using pBluescript-specific primers (fwd: ATAGGGCGAATTGGGTACCG and reverse: AAAGGGAACAAAACGTGGAGC). PCR reactions (100 µL) contained 20 µL 5× reaction buffer (Promega), 2 mM $MgCl_2$, 100 µM dNTPs, 0.15 µM of each primer and 2.5 U GoTaq® DNA polymerase (Promega). An initial 2 min denaturation was followed by 35 PCR cycles (94 °C for 30 s, 60 °C for 15 s and 2 min elongation at 72 °C) and a final 10 min elongation at 72 °C. The PCR products were purified using Millipore Multiscreen PCRµ96 Plate according to manufactures instructions. All purified probes were checked for size and purity by gel-electrophoresis (Invitrogen E-gel 96well 2% (GP)). The gel bands were inspected manually and all probes with corresponding wells showing up empty, displaying multiple bands and or with a band lower than 400 bp were flagged as "bad" for the microarray analysis. Probes in 50% DMSO were printed on Aminosilane coated slides (Corning® UltraGaps™) at 18 °C and 45–55% relative humidity using a BioRobotics, Micro Grid II arrayer (Genomic Solutions®) with Mikrospot 10 k split pins. Slides were dried in a desiccator cabinet for

24–48 h and DNA crosslinked at 350 mJ/cm^2 using a UV Stratalinker 2400, (Stratagene Inc.). The 16,000 spots were printed in 48 subarrays (each spot in duplicate on each subarray).

2.4.2. RNA isolation, cDNA labeling for microarray hybridization

RNA was isolated for individual animals using the RNAeasy Mini kit (Qiagen) according to the manufacture's recommendations. The RNA samples were frozen at −80 °C until analysis. One aliquot was used for RNA integrity and quantity measures using the Agilent 2100 Bioanalyzer and NanoDrop Spectrophotometer (OD 260/280 and 260/230 ratios), respectively. Another aliquot was used for cDNA synthesis and labeling using Fair Play® Microarray Labeling Kit (Stratagene) according to the manufacturer's instructions. 10 ug total RNA was used for cDNA generation. Samples were Cy5 labeled and a common reference standard (based on RNA from all tissues and stages used to construct the cDNA libraries) was labeled with Cy3. Labeling efficiency and quantity of labeled cDNA were determined using the NanoDrop Spectrophotometer. Slides were pre-hybridized in 20× SSC, 10% SDS and 1% BSA for about 45 min at 65 °C followed by washing twice in water and once in isopropanol. Slides were dried by centrifugation in a mini-centrifuge. Sample and reference were unified, and diluted in Tris buffer pH 8.0. After sample denaturation (100 °C, 2 min) hybridization was performed at 60 °C overnight with rotation using Agilent 2× hybridization buffer (250 μL) in Agilent hybridization chambers. The slides were put in 2× SSC/0.1% SDS at 65 °C to remove gasket slide and then washed for 5 min in 1× SSC at 65 °C, for 5 min in 0.2× SSC at RT, for 45 s in 0.05× SSC at RT, and centrifuged dry (mini-centrifuge).

2.4.3. Experimental design

RNA from each sample was prepared as described above, and hybridized to Cod 16 k IMR (Institute of Marine Research) microarrays. All samples were randomly labeled two batches, and hybridized in four batches. Slides were scanned directly after the washing procedure using an Agilent scanner at a resolution of 10 μm with default settings.

2.4.4. Preprocessing: filtering and normalization of microarray data

The scanned microarray images were analyzed using the GenePix Pro 6.0 software package and exported as image quantitation files (gpr- and jpg-files). The data files were quality controlled using R (R Development Core Team, 2005, http://www.r-project.org), and analyzed using J-Express Pro v.2.7 (Dysvik and Jonassen, 2001, http://www.molmine.com). Control probes, empty spots and probes marked with bad quality were removed from the analysis. Genes with more than 30% missing values were removed from the analysis and the remaining missing values were estimated using LSimpute Adaptive (Bo et al., 2004). Each array was normalized by Lowess (Cleveland and Devlin, 1988). Log 2 transformed ratios of foreground signals were used in the final gene expression matrix.

We provide MIAME-compliant description of the microarray study, available in the arrayexpress database (HYPERLINK "http://www.ebi.ac.uk/arrayexpress"www.ebi.ac.uk/arrayexpress).

2.5. SAM and GSEA

The search for differentially expressed genes was performed both on a single gene and gene set level. Two-class unpaired analysis in SAM (Tusher et al., 2001) as implemented in J-Express was used to look for differentially expressed genes on a gene by gene basis, while GSEA (Gene Set Enrichment Analysis, (Subramanian et al., 2005; Govoroun et al., 2006) was used to look for sets of genes sharing common characteristics that were differentially expressed between the classes examined. Gene sets were created using the Gene ontology (GO) (Rhee et al., 2008) cellular component, molecular function and biological function. This information was extracted from the OBO v.1 download dated May 7th 2009, found at http://www.geneontology.org. Probes were collapsed to contigs, before running GSEA. The largest value of all probes belonging to the same gene on an array was used as the value for that gene. Gene sets smaller than 15, and larger than 500, were excluded from the analysis. Two-way unpaired SAM was used to rank the genes. Significance of the gene set analysis was tested by permutating the scores over the samples (5000 iterations).

2.6. Northern blots

Total RNA from the liver, pituitary, pylorus and spleen (2.7 μg per lane) was mixed with Northern Max Formaldehyde Loading Dye (Ambion), denaturated (10 min, 80 °C), ethidium bromide added and samples run on a 1% denaturizing agarose gel (MOPS). Quality and quantity of RNA were evaluated under UV light before RNA was blotted onto Hybond-N nylon membrane (Amersham) using standard upward blotting technique in 10× SSC blotting buffer and crosslinked at 120 mJ/cm^2 using a UV Stratalinker (Stratagene).

PCR probes were produced as described for microarray probes and analyzed on agarose gel. The individual probes were cut out of the gel and purified using Ultrafree-DA (Millipore). PCR product (25 ng) was labeled with ^{32}P 3000 Ci/mmol (Perkin Elmer) using Rediprime (GE Healthcare) according to manufactures instructions. After denaturation (90 °C, 10 min) individual probe (2.8 ng/mL hybridization buffer) was hybridized to individual membranes at 68 °C overnight (Perfect Hyb hybridization buffer (Sigma), 7.5 mL per filter). The membranes were washed with 2× SSC/0.1% SDS (2×5 min RT), 1× SSC/0.1% SDS (1×15 min RT), 0.1× SSC/0.1% SDS (2×10 min 68 °C) followed by exposure on Kodak BioMax MS for 1 day.

3. Results and discussion

3.1. EST, number of transcripts and annotation

We have constructed 20 un-normalized cDNA libraries from polyA enriched RNA from a number of different tissues and stages of development (Table 1). We analyzed a total of 45,350 sequences, including 44,006 EST produced from these libraries and 1344 cod ESTs from other sources. Masking the sequences for quality, vector, and contamination removed 5876 sequences, and the remaining 39,474 sequences were assembled, resulting in 3539 contigs and 20,464 singletons. After this process, 610 sequences still had fragments showing similarity to sequences in UniVec, the longest match had a length of 58, and 557 matches were 25 bases or shorter. Of the 24,003 putative transcripts, 12,030 (2577 contigs and 9453 singletons) received a protein annotation. There was some redundancy in the protein annotations, so that a total of 8001 unique proteins were identified, and 2216 proteins were assigned to multiple contigs/singletons.

The chosen quality masking is quite aggressive and eliminates over 14% of the ESTs entirely, and masks almost 25% of the sequence data. The current settings were chosen after several experiments, including using even stricter masking (using a quality score of 20) and using only SeqClean's default masking parameters. Although less strict masking retains more sequences in the data set, it also results in smaller clusters and an inflated number of singletons. Conversely, a stricter masking reduces both the number of singletons and the redundancy in the protein annotations, but it also reduces the number of predicted proteins.

The BLAST hits were then used with the GOA GO (Gene ontology) associations to assign GO terms to each contig and singletons (Fig. 1). This shows that the ESTs cover a large range of biological processes and classes. The use of the data from the Gene Ontology project to associate gene data set with biological processes can be useful, but has some drawbacks which must be taken into account. One problem is connected to the imprecision and difference in annotation obtained which is related to the database being searched, another is the fact the many genes may be involved in more than one biological process. In addition, the best characterized processes will have a larger number of associated genes and can therefore often be overrepresented and give a bias towards these processes (Rhee et al., 2008).

3.2. Comparison of the sequences produced from the libraries

Several tissues are represented in multiple libraries that differ in either sexual maturation or immulogical stimulation. This approach allows us to identify genes that are present in particular stages or in response to specific stimuli, and also gives us an opportunity to investigate the function of particular transcripts.

Libraries from liver tissue were produces from both sexually maturing females (CLE) and immature females (CLU). We identified 14 different contigs that were uniquely found in liver tissue from sexually maturing

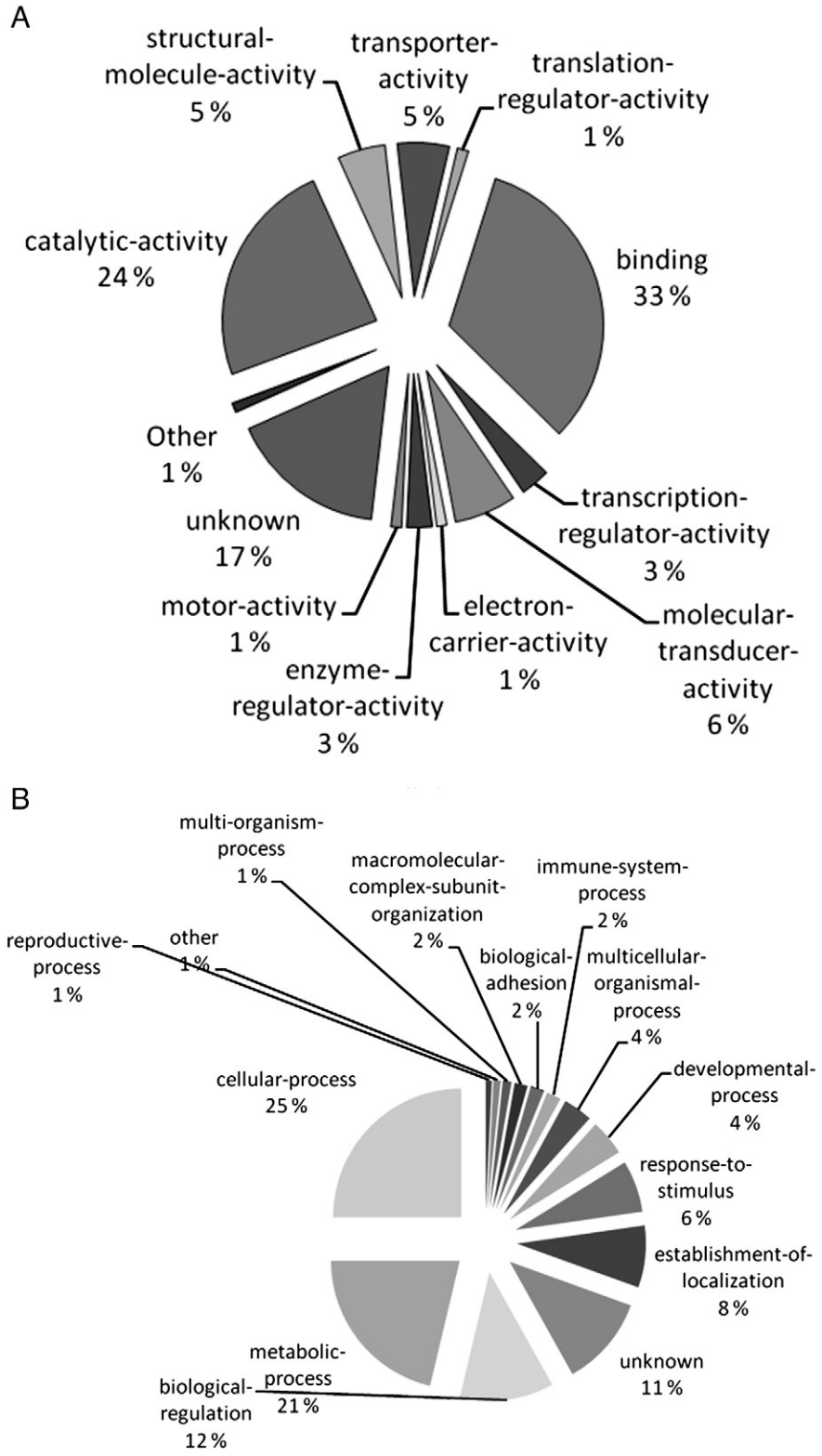

Fig. 1. GO 3 classes for all ESTs. Distribution of ESTs with homology to the UniRef90 database (E-value>1E-06). The genes are classified by A) Molecular function, B) Biological processes and C) Cellular component according to Gene Ontology classification.

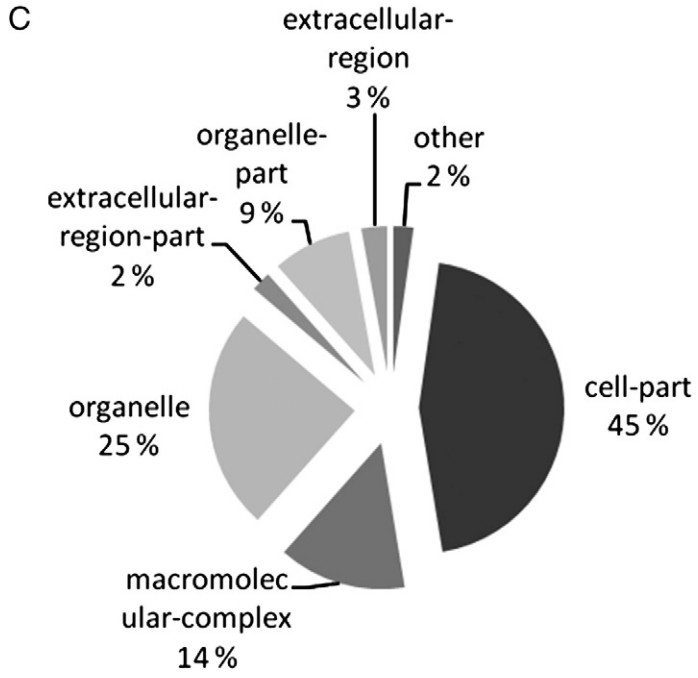

Fig. 1 (*continued*).

females. (Supplementary Table S2) All these encode transcripts associated with oocyte/egg production and some of them are highly abundant (e.g. vitellogenins, different types of egg shell proteins). These transcripts are not present in any other library and points to the importance of stage specific library to obtain sequences from genes transcribed in narrow windows. The highest abundant transcript in the liver library from immature specimens (CLU, Table 1) was a putative 14 kDa apolipoprotein (CL5Contig2) with 141 ESTs compared to 13 transcripts in the CLE-library. (Choudhury et al., 2009) identified a similar protein from rainbow trout and classified that as a homologue to the mammalian apolipoprotein A-II. However, our BLAST search only gave significant hits with sequences obtained from different species of teleosts. In the corresponding brain libraries (CHU for immature specimens and CHJ for mature ones, Supplementary Table S1) it was not possible to see the same trend with approaching sexual maturation causing transcripts to be up-regulated.

We also produced spleen and kidney libraries from normal (CMI and CHO, respectively) and immunologically stimulated specimens (CSMI and CSH). (see Supplementary Table S2). A key feature in the spleen and head kidney libraries was the presence of several highly abundant haemoglobin transcripts (i.e. haemoglobin subunits beta 1, beta 2 and alpha 1). According to the initial annotation there is no large proportion of putative immune-related transcripts in neither the spleen nor the head kidney libraries. However, a large obstacle in identifying immune-related genes in cod or any other fish species is relatively poorly characterized immune system at the molecular level compared to the situation in other vertebrates like mammals. The long evolutionary distance between teleost species and mammals also make room for significant differences in how different fish species may handle infectious diseases. MHC I is found in the present dataset but we have not identified any transcripts similar to MHC II. (Pilstrom et al., 2005) discussed the poor antibody response in cod and hypothesized that lack of MHC II could be one possible explanation, since MHC II so far not has been found in this species. Of the putative immune-related transcripts identified in the stimulated versus non stimulated spleen and head kidney most of them appears to be part of the innate immune system. Examples are CL236Contig1 which encodes a bactericidal permeability-increasing protein and CL276Contig1 which encodes a non-specific cytotoxic cell receptor protein 1.

3.3. Microarray

3.3.1. Microarray construction

The microarray was constructed as described in Section 2.5 on the basis of 15,648 clones from the cDNA libraries listed in Table 1. In addition to complement our random selection, 80 clones were handpicked from other libraries, 500 clones were contributed from Genome Atlantic in Canada and 120 from NIFES (National Institute of Nutrition and Seafood Research in Bergen, Norway). The total number of ETSs presented on the array is 16,348, 10,742 probes represent 3252 different contigs, while the remaining 6058 probes represent singletons. Of the contigs, 874 (1555 probes) have no annotation, while 3967 of the singletons lack annotation. The GO molecular function classification was used to visualize the annotated probes (Fig. 2).

In order to validate the microarray we sampled spleen, pituitary, testis and pylorus from 4 mature male Norwegian coastal cod (see Section 2.3). The samples were prepared and the experiment run as described in Section 2.4.2.

The Correspondence analysis (CA) plot is used to look for the greatest co-variance (between samples and genes) in the data. In the plot (Fig. S1, supplementary information) the samples from the same group are plotted together. There also seem to be certain genes that are correlated with each of the sample groups.

Furthermore, we also examined 4 differential expressed clones, one from each tissue, by Northern Blots as described in Section 2.6. The result showed a clear correspondence between the microarray data and the positive spots on the gel, see supplementary Fig. S2. In addition, three of the clones used for the blot were compared to the whole EST dataset, and a clear correlation to the microarray data and the Northern results can be seen.

3.3.2. Differential expression

The four tissues included in the present study have very different functions, and, hence they are very suitable for biological validation of the microarray. The SAM results clearly showed distinct transcription patterns for many genes easily distinguishing the four different tissues. To further illustrate the tissue specific transcription patterns a

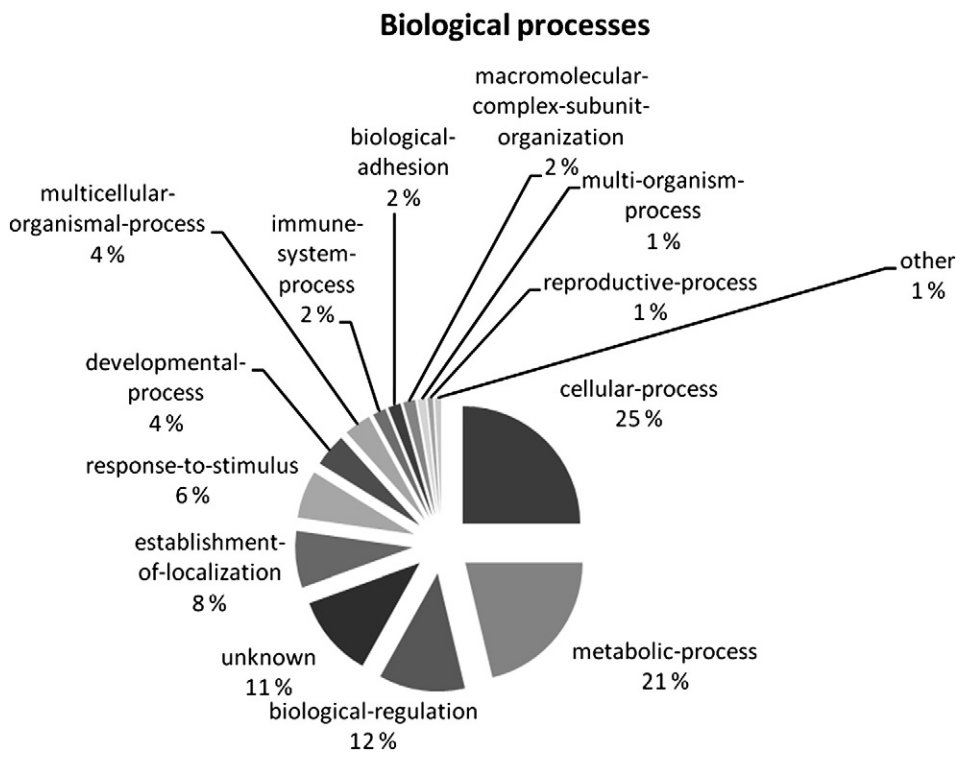

Fig. 2. GO biological process array. Classification of the annotated ESTs present in the microarray into Biological processes class by Gene ontology.

selection of probes showing significant variation between the 4 different tissues are displayed in supplementary Fig. S3.

The pylorus samples contain pancreatic tissue and these samples will also represent functions found in endocrine and exocrine pancreas cells. A key function for exocrine pancreatic cells is to transcribe genes encoding peptidases involved in digestion and store these as zymogens for release into the digestive tract. Different types of serine peptidase like trypsins, chymotrypsins and elastases are well known digestive enzymes being produced and stored in exocrine pancreatic cells. Our cDNA microarray contains probes for many serine peptidases. Of the 112 probes encoding trypsins or chymotrypsins none showed a distinct up-regulation in the pyloric samples compared to the other investigated tissues. There is also 38 probes representing elastases on the microarray and they showed a similar transcription pattern as trypsins and chymotrypsins. This clearly point to constitutive transcription of these genes and that the regulation is at the protein level. A transcript with significant identity to Peptidase D (peptidase M24B family) from various vertebrates was 8–10 folds up-regulated in the pyloric samples compared to the three other tissues. Peptidase D is transcribed in the brain in mouse and the kidney in rat and it is a signature peptidase from the pyloric sample in Atlantic cod (present study). Although we have not identified the cell type our putative peptidase D is transcribed in, the pyloric samples from cod are very different from brain or kidney tissue where this gene is active in rat or mice and points to a possible new or extended function for this gene. A microarray probe (CPY1782) encoding a putative chitinase was significantly up-regulated in the pyloric samples. Chitinases degrade chitin, chitotriose and chitobiose. In mammals, it has been suggested that chitinases are involved in host defense against nematodes and other pathogens, and it has been detected in spleen and cultured macrophages. However, since fish like cod eat crustaceans it is also possible that chitinases can be involved in digestion processes. The fact that this gene was upregulated in the pyloric samples may suggest an involvement in digestion rather than host defense.

The pituitary is an endocrine gland secreting several key hormones involved in a range of biological processes like growth, water and osmoregulation and reproduction. In general, many probes encoding hormone receptors and/or hormones were strongly up-regulated in the pituitary compared to the other examined tissues.

The testis is a specialized tissue producing male gametes. This is also clearly reflected in genes up-regulated in the present experiment. An example is the outer dense fiber of sperm tail protein 3 probe (CTE776) which has been identified in testis in several vertebrate species is also significantly up-regulated in the cod testis.

Approximately 50% of the array probes do not have any significant hits in the non redundant database in GenBank. The microarray experiment showed that a significant proportion of these were differential regulated in the examined tissues. In the testis, 10 of the 50 most differential regulated transcripts did not have a significant hit in GenBank. Three transcripts with no significant hit in GenBank have been included in supplementary Fig. S3. This makes microarray experiments a useful tool to link novel gene transcripts to know biological processes and an important step towards characterization of new genes and gene products.

3.3.3. Gene Set enrichment analysis

Gene Set Enrichment Analysis (GSEA) is a computational method that determines whether an a priori defined set of genes shows statistically significant, concordant differences between two biological states. This enables detection of biological processes, such as transcriptional programs and metabolic pathways distributed across a large network of genes and that are often hard to identify at the level of individual genes (Subramanian et al., 2005). By applying this method to our dataset we can visualize processes distinguishing and characterizing the four tissues from each other (Supplementary Table S3). We have compared each tissue against the three others as described in Section 2.5.

There are several gene sets showing up-regulation compared to the other tissues for both Testis, Pylorus, Pituitary and Spleen, while only the testis displayed clear data of the opposite, down-regulated processes. The down regulation observed in testis seems to a large extent to be related to transcript associated to the immune system, which seems reasonable since testis is a specialized tissue producing male gametes with a high turnover. A key feature for up-regulated transcripts in the pituitary seems to be related to signaling (e.g. receptor binding, steroid binding) due to the endocrine function of this particular tissue. Of gene sets up-regulated in spleen heme binding and iron binding are signatures of spleen activity. However, there was no enrichment for immune-related genes in this analysis. The GSEA clearly demonstrate that the constructed microarray provides biological meaningful results since the four different tissues examined are distinct.

4. Summary and conclusions

We have analyzed ESTs from 20 different *G. morhua* cDNA libraries covering 8000 unique transcripts and representing a large range of biological processes. A 16 k cDNA microarray was constructed on the basis of the ESTs. The microarray has been used to study the transcriptome of four different tissues in cod. The different tissues displayed tissue specific transcriptomes and the clones compared with northern blotting verify the results from the array. The results make biological sense and the microarray can be a valuable instrument for studying gene expression in this important fish species.

A major challenge in annotation of the EST-data from new species like the cod is the relative large proportion of significant matches to poorly annotated sequences from other fish species, resulting in a large fraction of transcripts being of unknown function. Similarity to a sequence does not necessarily give precise information about what biological process a particular gene or protein is involved in when the evolutionary distance increase. The amount of sequences in databases will continue to increase but will not necessary make it easier to perform accurate annotation based on database searches. This will be become even more important with the expected full sequencing of fish genomes in the near future.

Acknowledgments

We would like to thank Heidi Kongshaug, Stig Mæhle and Sara Ferreira for excellent assistance in preparing cDNA libraries, sequencing, microarray probe preparation, microarray experiments and northern blot. We take Sonal Patel and Audun Nerland for providing samples for the production of the CSMI and CTH cDNA libraries. The fish for the microarray experiment was provided by Knut Jørstad through the Biobank project. We acknowledge both NIFES and Sharen Bowman at Genome Atlantic for providing clones for the array. We would also like to thank Anne-Kristin Stavrum and the NMC (Norwegian Microarray Consortium) for valuable help on the microarray analysis, and the NMC node in Trondheim for printing the arrays. This research was supported by the FUGE program in the Norwegian Research Council.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.cbd.2010.06.002.

References

Adzhubei, A.A., Vlasova, A.V., Hagen-Larsen, H., Ruden, T.A., Laerdahl, J.K., Hoyheim, B., 2007. Annotated expressed sequence tags (ESTs) from pre-smolt Atlantic salmon (*Salmo salar*) in a searchable data resource. BMC Genomics 8, 209.

Barrell, D., Dimmer, E., Huntley, R.P., Binns, D., O'Donovan, C., Apweiler, R., 2009. The GOA database in 2009—an integrated Gene Ontology Annotation resource. Nucleic Acids Res. 37, D396–D403.

Bo, T.H., Dysvik, J., Jonassen, I., 2004. LSimpute: accurate estimation of missing values in microarray data with least squares methods. Nucleic Acids Res. 32.

Choudhury, M., Yamada, S., Komatsu, M., Kishimura, H., Ando, S., 2009. Homologue of mammalian apolipoprotein A-II in non-mammalian vertebrates. Acta Biochim. Biophys. Sin. 41, 370–378.

Cleveland, W.S., Devlin, S.J., 1988. Locally weighted regression — an approach to regression-analysis by local fitting. J. Am. Statist. Assoc. 83, 596–610.

Dahle, G., Jorstad, K.E., Rusaas, H.E., Ottera, H., 2006. Genetic characteristics of broodstock collected from four Norwegian coastal cod (*Gadus morhua*) populations. ICES J. Mar. Sci. 63, 209–215.

Delghandi, M., Stenvik, J., Nilsen, F., Wesmajervi, M.S., Fjalestad, K.T., Damsgard, B., 2008a. Identification and characterisation of nine new gene-associated microsatellite markers for Atlantic cod (*Gadus morhua* L.). Conserv. Genet. 9, 747–749.

Delghandi, M., Wesmajervi, M.S., Mennen, S., Nilsen, F., 2008b. Development of twenty sequence-tagged microsatellites for the Atlantic cod (*Gadus morhua* L.). Conserv. Genet. 9, 1395–1398.

Delghandi, M., Wesmajervi, M.S., Tafese, T., Nilsen, F., 2008c. Twenty-three novel microsatellite markers developed from Atlantic cod *Gadus morhua* L. expressed sequence tags. J. Fish Biol. 73, 444–449.

Delghandi, M., Wesmajervi, M.S., Mennen, S., Nilsen, F., 2009. New polymorphic dinucleotide microsatellite markers for Atlantic cod (*Gadus morhua* L.). Conserv. Genet. 10, 1037–1040.

Dieckmann, U., Heino, M., 2007. Probabilistic maturation reaction norms: their history, strengths, and limitations. Mar. Ecol. Prog. Ser. 335, 253–269.

Douglas, S.E., Knickle, L.C., Kimball, J., Reith, M.E., 2007. Comprehensive EST analysis of Atlantic halibut (Hippoglossus hippoglossus), a commercially relevant aquaculture species. BMC Genomics 8, 144.

Dysvik, B., Jonassen, I., 2001. J-Express: exploring gene expression data using Java. Bioinformatics 17, 369–370.

Fjelldal, P.G., van der Meeren, T., Jorstad, K.E., Hansen, T.J., 2009. A radiological study on vertebral deformities in cultured and wild Atlantic cod (*Gadus morhua*, L.). Aquaculture 289, 6–12.

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.

Huang, X., Madan, A., 1999. CAP3: a DNA sequence assembly program. Genome Res. 9, 868–877.

Johansen, S.D., Coucheron, D.H., Andreassen, M., Karlsen, B.O., Furmanek, T., Jorgensen, T.E., Emblem, A., Breines, R., Nordeide, J.T., Moum, T., Nederbragt, A.J., Stenseth, N.C., Jakobsen, K.S., 2009. Large-scale sequence analyses of Atlantic cod. Nat. Biotechnol. 25, 263–271.

Jorstad, K.E., Van Der Meeren, T., Paulsen, O.I., Thomsen, T., Thorsen, A., Svasand, T., 2008. "Escapes" of eggs from farmed cod spawning in net pens: recruitment to wild stocks. Rev. Fish Sci. 16, 285–295.

Lee, Y., Tsai, J., Sunkara, S., Karamycheva, S., Pertea, G., Sultana, R., Antonescu, V., Chan, A., Cheung, F., Quackenbush, J., 2005. The TIGR Gene Indices: clustering and assembling EST and known genes and integration with eukaryotic genomes. Nucleic Acids Res. 33, D71–D74.

Malde, K., Schneeberger, K., Coward, E., Jonassen, I., 2006. RBR: library-less repeat detection for ESTs. Bioinformatics 22, 2232–2236.

Moen, T., Hayes, B., Nilsen, F., Delghandi, M., Fjalestad, K.T., Fevolden, S.E., Berg, P.R., Lien, S., 2008. Identification and characterisation of novel SNP markers in Atlantic cod: evidence for directional selection. BMC Genetics 9.

Moen, T., Delghandi, M., Wesmajervi, M.S., Westgaard, J.I., Fjalestad, K.T., 2009. A SNP/microsatellite genetic linkage map of the Atlantic cod (*Gadus morhua*). Anim Genet 40, 993–996.

Nielsen, E.E., Hemmer-Hansen, J., Poulsen, N.A., Loeschcke, V., Moen, T., Johansen, T., Mittelholzer, C., Taranger, G.L., Ogden, R., Carvalho, G.R., 2009. Genomic signatures of local directional selection in a high gene flow marine organism; the Atlantic cod (*Gadus morhua*). BMC Evol. Biol. 9.

Olsen, E., Lilly, G.R., Heino, M., Morgan, M.J., Brattey, J., Dieckmann, U., 2005. Assessing changes in age and size at maturation in collapsing populations of Atlantic cod (*Gadus morhua*). Can. J. Fish Aquat. Sci. 62, 811–823.

Pertea, G., Huang, X.Q., Liang, F., Antonescu, V., Sultana, R., Karamycheva, S., Lee, Y., White, J., Cheung, F., Parvizi, B., Tsai, J., Quackenbush, J., 2003. TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. Bioinformatics 19, 651–652.

Pilstrom, L., Warr, G.W., Stromberg, S., 2005. Why is the antibody response of Atlantic cod so poor? The search for a genetic explanation. Fish Sci. 71, 961–971.

Quinn, N.L., Levenkova, N., Chow, W., Bouffard, P., Boroevich, K.A., Knight, J.R., Jarvie, T.P., Lubieniecki, K.P., Desany, B.A., Koop, B.F., Harkins, T.T., Davidson, W.S., 2008. Assessing the feasibility of GS FLX Pyrosequencing for sequencing the Atlantic salmon genome. BMC Genomics 9.

R Development Core Team, 2005. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria3-900051-07-0.

Rhee, S.Y., Wood, V., Dolinski, K., Draghici, S., 2008. Use and misuse of the gene ontology annotations. Nat. Rev. Genet. 9, 509–515.

Sarropoulou, E., Nousdili, D., Magoulas, A., Kotoulas, G., 2008. Linking the genomes of nonmodel teleosts through comparative genomics. Mar. Biotechnol. 10, 227–233.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., Mesirov, J.P., 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. U. S. A. 102, 15545–15550.

Suzek, B.E., Huang, H., McGarvey, P., Mazumder, R., Wu, C.H., 2007. UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics 23, 1282–1288.

Taranger, G.L., Aardal, L., Hansen, T., Kjesbu, O.S., 2006. Continuous light delays sexual maturation and increases growth of Atlantic cod (*Gadus morhua* L.) in sea cages. Ices J. Mar. Sci. 63, 365–375.

Taranger, G.L., Carrillo, M., Schulz, R.W., Fontaine, P., Zanuy, S., Felip, A., Weltzien, F.A., Dufour, S., Karlsen, O., Norberg, B., Andersson, E., Hausen, T., 2010. Control of puberty in farmed fish. Gen. Comp. Endocrinol. 165, 483–515.

Tusher, V.G., Tibshirani, R., Chu, G., 2001. Significance analysis of microarrays applied to the ionizing radiation response. Proc. Natl. Acad. Sci. U. S. A. 98, 5116–5121.

Westgaard, J.I., Tafese, T., Wesmajervi, M.S., Nilsen, F., Fjalestad, K.T., Damsgard, B., Delghandi, M., 2007. Development of ten new EST-derived microsatellites in Atlantic cod (*Gadus morhua* L.). Conserv. Genet. 8, 1503–1506.

Whitehead, A., Crawford, D.L., 2006. Neutral and adaptive variation in gene expression. Proc. Natl. Acad. Sci. U. S. A. 103, 5425–5430.

Zhang, Z., Schwartz, S., Wagner, L., Miller, W., 2000. A greedy algorithm for aligning DNA sequences. J. Comput. Biol. 7, 203–214.