Characterization of rainbow trout gonad, brain and gill deep cDNA repertoires using a Roche 454-Titanium sequencing approach

Aurélie Le Cam\textsuperscript{a,1}, Julien Bobe\textsuperscript{a,4}, Olivier Bouchez\textsuperscript{b,2}, Cédric Cabau\textsuperscript{b,3}, Olivier Kah\textsuperscript{c,4}, Christophe Klopp\textsuperscript{b,3}, Jean-Jacques Lareyre\textsuperscript{a,4}, Isabelle Le Guen\textsuperscript{a,4}, Jérôme Lluch\textsuperscript{d,2}, Jérôme Montfort\textsuperscript{a,3}, François Moreews\textsuperscript{b,3}, Barbara Nicol\textsuperscript{a,4}, Patrick Prunet\textsuperscript{a,4}, Pierre-Yves Rescan\textsuperscript{a,4}, Arianna Servili\textsuperscript{a,4}, and Yann Guiguen\textsuperscript{a,\S,5}

\begin{itemize}
\item \textsuperscript{a} INRA, UR1037 LPGP Fish Physiology and Genomics, F-35000 Rennes, France.
\item \textsuperscript{b} INRA, Sigenae UR875 BIA, F-31326 Castanet-Tolosan Cedex, France.
\item \textsuperscript{c} UMR CNRS 6026, IFR 140, Université de Rennes 1, France.
\item \textsuperscript{d} INRA, Plateforme Génomique, Génopole Toulouse / Midi-Pyrénées, Auzeville, F-31326 Castanet-Tolosan, France.
\item \S Corresponding author: Tel.: 00 33 2 02 23 48 50 09, Fax: 00 33 2 02 23 48 50 20
\item E-mail address: yann.guiguen@rennes.inra.fr
\end{itemize}

Keywords: transcriptome; next-generation sequencing (NGS); fish; salmonids; rainbow trout.

\textsuperscript{1} A.L.C. produced the double-stranded cDNA, supervised the construction and sequencing of the 454 libraries, and participated in the writing of the manuscript.
25 O.B. and J.L. aided in the construction of the 454 libraries and performed the
26 sequencing reactions.
27 J.M., C.C., F.M. and C.K. developed the rainbow trout contig browser and performed
28 the bioinformatic analyses.
30 repertoires, performed QPCR analysis and drafted parts of the manuscript.
31 Y.G. designed and supervised the study, participated in data analysis and drafted the
32 final manuscript.
33
34 Abbreviations: ACTH, adrenocorticotrophic hormone; amh, anti-müllerian hormone
35 gene; AMP, adenosine monophosphate; bHLH, basic-helix-loop-helix; Blbp, Brain
36 Lipid Binding protein; bp, base pair(s); Btg1, B Cell Translocation gene 1; cDNA,
37 DNA complementary to RNA; CE, cornified envelope; dazl, deleted in azoospermia
38 gene; DDD, Digital Differential Display; dmrt1, double-sex and mab-3 related
39 transcription factor gene 1; EST, Expressed Sequenced Tag; fabp7a, fatty acid
40 binding protein gene 7a; gap43, growth associated protein 43 gene; Gb, giga base;
41 gdf9, growth and differentiation factor gene 9; glast, glutamate aspartate transporter
42 gene; gsdf, gonadal soma-derived factor gene; Na/K-ATPase, sodium-potassium
43 adenosine triphosphatase; nrn1, neuritin gene 1; neurod1, Neurogenic differentiation
44 1 gene; NGS, Next-Generation Sequencing; rec8, Meiotic recombination protein
45 REC8 homolog; RNA-seq, RNA sequencing; ocpi, oocyte cysteine protease inhibitor
46 gene; POMC, pro-opiomelanocortin; pou5f1, POU domain class V transcription factor
47 1 gene; RyR1, Ryanodine receptor 1; SAGE, Serial analysis of gene expression;
48 Sox9a, SRY-box containing gene 9a; Smyd1, SRA, NCBI Sequence Read Archive;
49 sycp1, 2 and 3, synaptonemal complex protein; rhcg2, Rh type C glycoprotein2a
Abstract

Rainbow trout, *Oncorhynchus mykiss*, is an important aquaculture species worldwide and, in addition to being of commercial interest, it is also a research model organism of considerable scientific importance. Because of the lack of a whole genome sequence in that species, transcriptomic analyses of this species have often been hindered. Using next-generation sequencing (NGS) technologies, we sought to fill these informational gaps. Here, using Roche 454-Titanium technology, we provide new tissue-specific cDNA repertoires from several rainbow trout tissues. Non-normalized cDNA libraries were constructed from testis, ovary, brain and gill rainbow trout tissue samples, and these different libraries were sequenced in 10 separate half-runs of 454-Titanium. Overall, we produced a total of 3 million quality sequences with an average size of 328 bp, representing more than 1 Gb of expressed sequence information. These sequences have been combined with all publicly available rainbow trout sequences, resulting in a total of 242,187 clusters of putative transcript groups and 22,373 singletons. To identify the predominantly expressed genes in different tissues of interest, we developed a Digital Differential Display (DDD) approach. This approach allowed us to characterize the genes that are predominantly expressed within each tissue of interest. Of these genes, some were already known to be tissue-specific, thereby validating our approach. Many others, however, were novel.
candidates, demonstrating the usefulness of our strategy and of such tissue-specific resources. This new sequence information, acquired using NGS 454-Titanium technology, deeply enriched our current knowledge of the expressed genes in rainbow trout through the identification of an increased number of tissue-specific sequences. This identification allowed a precise cDNA tissue repertoire to be characterized in several important rainbow trout tissues. The rainbow trout contig browser can be accessed at the following publicly available web site (http://www.sigenae.org/).

1. Introduction

Rainbow trout, *Oncorhynchus mykiss*, is an important aquaculture species worldwide and, in addition to being of commercial interest, it is also a model research organism of considerable scientific importance (reviewed by Thorgaard et al. (Thorgaard et al., 2002)). Many genomic resources are available for this species, including extensive classical Expressed Sequenced Tag (EST) repertoires (Rexroad et al., 2003; Govoroun et al., 2006; Koop et al., 2008); however, because of the lack of a currently available whole genome sequence in that species, genomic analysis has been inhibited by the limited knowledge of the rainbow trout transcriptome. The recent development of next-generation sequencing (NGS) has been a major technological breakthrough (Metzker, 2010). This technique provides new opportunities to better characterize gene repertoires, even in non-model species for which a whole genome sequence is unavailable (Crawford et al., 2010; Mizrachi et al., 2010; Fraser et al., 2011). These NGS technologies are now widely used for transcript expression profiling, and this approach, also referred to as RNA-seq, is progressively replacing classical techniques such as SAGE analysis or DNA microarrays (Costa et al., 2010).
Using NGS and 454-Titanium technology, this study sought to complement the current knowledge of the rainbow trout transcriptome by providing new tissue-specific cDNA repertoires from different rainbow trout tissues. 454-Titanium technology has already been used to characterize rainbow trout transcripts (Salem et al., 2010); however, tissue-specific information is currently lacking. We thus developed sequencing resources and bioinformatics tools to initiate expression profiling analyses on the vast new collection of tissue-specific sequence information generated by this study. These first analyses demonstrate that our approach is reliable. These new resources may now be utilized to search for more subtle genetic regulation events, allowing the identification of genes that are predominantly expressed under different physiological conditions.

2. Materials and Methods

2.1 Tissue sampling and RNA extraction
Total RNA was extracted from the collected tissues (gill, testis, brain and ovary) with TRIzol™ reagent (Invitrogen, Cergy-Pontoise, France) according to the manufacturer's instructions. The total RNA yield was estimated using a Nanodrop 1000 spectrophotometer (Labtech, Palaiseau, France), and RNA quality was examined with an Agilent Bioanalyzer (Agilent Technologies, Massy, France). Only good quality total RNA samples (RNA Integrity Number, RIN > 8) were selected for further use. The total RNA from several extractions was pooled to minimize the impact of biological variations and to obtain the amount of RNA required for library preparation. Because of the presence of mucus, the gill sample pools were purified with the NucleoSpin® RNA II kit (Macherey-Nagel, Hoerdt, France) according to the "Clean-up of RNA from reaction mixtures" protocol. This purification was made from
200 µg of the total RNA mix, and the elution was performed twice with 70 µl of H_2O (2 successive elutions). mRNA was then isolated from 250 µg of total RNA using the Nucleotrap® mRNA kit (Macherey-Nagel, Hoerdt, France) as described in the manufacturer's manual, with the exception of the elution procedure. Here, 40 µl of pre-warmed (68°C) H_2O were used per 40 µl of Oligo(dT) latex beads. Three rounds of elution were performed. The quality of the mRNA and the depletion of ribosomal RNA were determined with the Agilent Bioanalyzer (Agilent Technologies, Massy, France).

### 2.2 Double stranded-cDNA preparation

Double-stranded cDNA was prepared as previously described (Leroux et al., 2010) using modifications of the SuperScript Double-Stranded cDNA Synthesis Kit (Invitrogen, Cergy-Pontoise, France) protocol. Briefly, we used a modified primer (5’ GAGAGAGAGACTGGAG(T)16VN 3’) containing the GsuI restriction site and methylated dCTP (10 mM dm5CTP, Fermentas, Saint-Remy-les-Chevreuse, France) during the first-strand cDNA synthesis. The first strand was synthesized from 3 µg of mRNA. The next steps (second strand synthesis, RNase I digestion, phenol/chloroform extraction and ethanol precipitation) were performed according to the manufacturer's instructions. The pellets were resuspended in 9 µl of water. Three to five independent syntheses were made for each sample and pooled to obtain a total of 3.5 to 5 µg of total cDNA. Next, the double-stranded cDNA was digested by GsuI (5 U) for 90 min at 30°C to remove the polyA tail. The cDNA samples were quantified using the Quant-iT PicoGreen kit (Invitrogen, Cergy-Pontoise, France).
2.3 454 The Roche sequencing method

After cDNA nebulization, the libraries were prepared using the GS DNA Library Preparation Kit according to the manufacturer's recommendations (Roche Applied Science, Meylan, France); 3.5 to 5 µg of each cDNA sample were used to generate the libraries. Sheared DNA was ligated to the linker for emPCR and sequencing. The final libraries were quantified with the SlingShotTM kit using the Fluidigm® Digital Array according to the manufacturer's instructions. Pyrosequencing using 454/Roche GS FLX Titanium chemistry was performed with the GS Titanium LV emPCR and sequencing kits according to the manufacturer's instructions. Each sample was sequenced on one region of a two-region Picotiterplate.

2.4 Real-time PCR analysis

Real-time PCR was performed using the Step One Plus system (Applied Biosystems, Foster City, USA) as previously described (Desvignes et al., 2009). For each tissue, three separate reverse transcription reactions were conducted using 2 µg of DNase-treated RNA that originated from three different fishes. Control reactions were performed in the absence of reverse transcriptase. RT products and control reaction samples were diluted 1/25, and 4 µl was used for each real-time PCR reaction. Real-time PCR was performed with the Fast SYBR Green Master Mix (Applied Biosystem) and either 300 or 600 nm of each primer. All real-time PCR reactions were performed in duplicate. After amplification, a melting curve was performed according to the manufacturer recommendations to check the specificity of amplification. The relative expression of each gene was normalized to the level of 18S gene amplification and was calculated as the percentage of the highest expression.
level recorded for each gene. The sequences of the primers used are listed in Supplementary file 2.

2.5 In situ hybridization

Digoxigenin-labeled antisense RNA probes were synthesized from a PCR-amplified template (EST clone: BX085208 containing the rainbow trout β-enolase sequence) using T3 RNA polymerase. In situ hybridizations were performed on transverse sections of rainbow trout embryos as previously described (Gabiillard et al., 2003), with minor modifications.

2.6 Bioinformatic analysis of the sequence data

The initial sequence set used for this transcriptome assembly was composed of publicly available mRNA sequences from Genbank, Sanger ESTs from dbEST, GS-Flx 454 RNA-Seq from SRA (including a complete rainbow trout muscle 454 dataset publicly available under the SRA accession number SRA029716) and the locally produced 454 Titanium RNA-Seq reads that are available under the SRA accession number SRA026493 (Table 1). The reads were cleaned to remove sequences composed of less than 100 base pairs, polyA tails, and *E. coli* and yeast contaminating sequences using SeqClean (http://compbio.dfci.harvard.edu/tgi/software/). To ease the assembly process, our next step was to remove redundancy using the cd-hit-est software (Li and Godzik, 2006). This software discarded those reads that were included in other reads. A threshold of 0.98 was chosen, producing a non-redundant representative read set (parameters -M 30000 -d 0 -n 10 -l 11 -r 1 -p 1 -g 1 -G 1 -c 0.98 -aS 0.5 -T 8). The discarded reads were repositioned in the resultant contigs after assembly using the inclusion coordinates. The assembly was performed using TGICL (Pertea et al., 2003), which first clusters the reads using megablast, given a similarity...
threshold (-p 96) and a similarity length (-l 60), and then, using the mRNAs as backbones, splits the clusters larger than a given size (-s 100000). The program then produces alignment files and consensus sequences with CAP3. All of the 454 singlets were removed from the final results, which contained 219,814 contigs and 22,373 singlets. The contigs and singlets were annotated by similarity searches using a large set of nucleotide and protein reference databases, including UniProtKB/Swiss-Prot, RefSeq protein, RefSeq RNA, PFAM, DFCI gene indices, UniGene clusters and Ensembl transcripts for numerous species. The full list of databases is available on the home page of the assembly (http://public-contigbrowser.sigenae.org:9090/Oncorhynchus_mykiss/index.html). The e-value thresholds that were used to determine similarities were dependent upon the database that was used; $e^{-10}$ was used for transcript databases, $e^{-2}$ was used for species-specific nucleic databases, and $e^{-5}$ was used for all others. The UniProtKB/Swiss-Prot similarity search was performed using a bi-directional best hit approach to identify putative orthologs, and the resulting annotation character string had a [BBH] tag.

Variation detection was also performed using the contigs. The in-house detection script screens the CAP3 alignment files to find substitutions that occur at least twice at a position with a depth of at least 6 sequences and that have no other substitution in a twenty base pair window around this position. All of the resulting files from the assembly, the annotation and the variation detection steps were then reformatted to populate the contig browser database, which was used to show the contig layout, and the BioMart (Haider et al., 2009) database, which enables a large variety of queries. The contig browser was built using the Ensembl 40 web code (Birney et al., 2006) and provides such new features as the Digital Differential Display view and the Venn
diagram view. The contig and singlet FASTA files and the BioMart tabular files can be downloaded from the web server.

3. Results and Discussion

3.1 The generation of tissue-specific transcriptomic resources using a RNA-Seq approach

Eight non-normalized cDNA libraries were constructed from rainbow trout testis, ovary, brain, and gill tissue samples and were subsequently sequenced in 10 separate half-runs using 454-Titanium (Table 1). Overall, we produced a total of 3 million good-quality sequences (testis, 1,234,260; ovary, 738,510; brain, 481,363; and gill, 613,223) with an average sequence size of 328 bp. Collectively, these sequences constitute more than 1 Gb of expressed sequence information. This raised the total number of expressed sequences that are publicly available for rainbow trout to 4.8 millions. In contrast to previous large NGS datasets produced in the rainbow trout (Salem et al., 2010), our strategy also provided tissue-specific information that was not previously available for this species. All of the acquired sequences were deposited in the NCBI Sequence Read Archive (SRA) and were assembled along with the publicly available rainbow trout sequences. This resulted in a total of 242,187 clusters of putative transcripts (contigs) and 22,373 singletons. These sequences and their assembly can be accessed at the following publicly available web site (http://www.sigenae.org/). This browser is based on an Ensembl architecture (Hubbard et al., 2002; Stabenau et al., 2004). One can utilize many classical features to browse all of the computed information on each contig and can extract large user-specific datasets using the Biomart generic interface (Smedley et al., 2009). In silico
strategies have often been used to mine sequence databases and discover new genes
with a predominant expression pattern in one tissue or a group of tissues (Murray et
al., 2007); however, the use of such strategies has been restricted to species with large
available sequence datasets. In non-model species, this approach has, in contrast,
often been limited by the rather low number of sequences available. With the recent
advent of NGS, this limitation is no longer critical. Information on many non-model
species is now more accessible due to RNA-Seq approaches (Wilhelm and Landry,
2009), allowing for the characterization of genes with differential expression patterns
(Crawford et al., 2010; Mizrachi et al., 2010; Fraser et al., 2011). Therefore, we
applied this approach to our own dataset to validate its utility.

3.2 The use of the 454 tissue-specific transcriptomic resource for the
identification of tissue-predominant transcripts
A Digital Differential Display (DDD) analysis tool was implemented within the
rainbow trout sequence browser to characterize contigs with a predominant expression
pattern. DDD analysis was subsequently performed using the 454 sequencing data
that was obtained from this study. For each tissue of interest (ovary, testis, brain, gill
and muscle), we compared the number of reads contained in each of the contigs in the
tissue of interest to all of the remaining tissue libraries. Some of these examples are
discussed below, and the DDD software is available within the rainbow trout contig
browser to allow for further user-specific queries. To confirm that this strategy is
effective for the identification of genes that are predominantly expressed in a given
tissue, we carried out qPCR validation for 9 genes (see Fig 1 and Tables 2 to 6).

3.2.1 The identification of genes that are preferentially expressed in the rainbow trout
ovary
When performing a DDD analysis on the trout ovarian samples using a 10-fold cut-off, 614 contigs were found to be over abundant in the ovarian/oocyte libraries in comparison to the non-ovarian libraries. Among those 614 contigs, 438 contigs were found only 1 or 2 times in the non-ovarian libraries (see additional file 1). In addition, 432 contigs were found at least 10 times in the ovarian libraries but were never detected in the non-ovarian libraries. Among the most overrepresented genes found in the ovarian libraries, we identified several ZP domain-containing genes. Our results show that, in agreement with existing data from medaka (*Oryzias latipes*) (Kanamori et al., 2003), the *zpc1*, *zpc5*, and *zpax* genes are predominantly expressed in the ovaries in the rainbow trout (Table 2). Similarly, several genes that are known to be specifically or predominantly expressed in rainbow trout oocytes or ovaries were identified. In agreement with existing data on rainbow trout (Bobe et al., 2008), zygote arrest 1 (*zar1*) and growth and differentiation factor 9 (*gdf9*) exhibited a very strict ovarian-predominant profile. Similarly, oocyte cysteine protease inhibitor (*ocpi*) was dramatically overexpressed in the ovarian libraries, again in agreement with existing data (Bobe and Goetz, 2001). In addition to these previously documented patterns, which validate our approach, the DDD analysis has led to the identification of many genes that also putatively exhibit predominant or specific ovarian and/or oocyte expression. Interestingly, our data also revealed the predominant expression of several ovarian genes in the rainbow trout that are known to be strongly expressed in the zebrafish (*Danio rerio*) oocyte, such as bucky ball (Marlow and Mullins, 2008), or in the medaka, such as the transcription factor *pou5f1* (Sanchez-Sanchez et al., 2010) (Fig. 1 and Table 2). Furthermore, we have shown the ovarian-predominant expression of *wnt11* (Fig. 1 and Table 2). In sexually mature rainbow trout, *wnt11* mRNA was detected in all assayed tissues. In comparison to skin, which had the
lowest level of \(wnt11\) expression, we observed a 15,000-fold higher expression of
\(wnt11\) in post-ovulatory ovaries and a 1,700-fold higher expression in oocytes (Fig.
1). Our results are consistent with previous studies reporting the expression of \(wnt11\)
in mouse granulosa cells and oocytes (Harwood et al., 2008). Interestingly, \(wnt11\) is a
maternal transcript. Maternal \(wnt11\) mRNA has been shown to accumulate in
Xenopus oocytes (Ku and Melton, 1993), where it plays a role in axis formation in the
embryo (Tao et al., 2005). Together, these results suggest the participation of
maternally inherited \(wnt11\) mRNA in teleost early development. The DDD analysis
also allowed the identification of novel oocyte-predominant genes in rainbow trout
(see Additional File 1). Such genes deserve further attention.

3.2.2 The identification of genes that are preferentially expressed in trout testes
When performing a DDD analysis of the trout testicular samples, we characterized
503 contigs that were only expressed in the testes, with a minimum of 10 reads used
as a threshold. In addition, 590 contigs were significantly overexpressed in the testes,
with a 10-fold minimum change and a threshold of 10 reads (see additional file 1). As
expected, numerous contigs corresponded to transcripts that are known to be
preferentially expressed in the testes of several vertebrate species, including the
rainbow trout (Rolland et al., 2009). These transcripts are involved in the sexual
differentiation of the male gonad (\(Sox9a, dmrt1, amh\)) and in specific testicular
functions such as the regulation of germ cell proliferation (\(amh, gsdf\)), meiosis (\(sycp1,\)
\(sycp2\) and \(sycp3, rec8, dazl, vasa, tudor1\)), and spermiogenesis (\(Dynein intermediate\)
\(chain 1, microtubule associated protein 6\)). These observations illustrate the reliability
of the DDD method in the analysis of the 454 sequencing data. Interestingly, our
analysis revealed several transcripts that were unexpectedly overexpressed in trout
testes. For instance, one contig sequence encodes the trout counterpart of the
evolutionarily conserved TBX19 transcription factor (Table 3). TBX19 is a member of a multigene superfamily that harbors a conserved DNA binding domain called the T-box (see (Plageman and Yutzey, 2005) for review). The TBX transcription factors are involved in the regulation of developmental processes. In humans and mice, TBX19 expression is restricted to two pituitary cell types that express pro-opiomelanocortin (POMC). In trout, enhanced TBX19 expression was also detected in the pituitary along with the testes (Fig. 1). A mutation of the tbx19 gene in humans results in the impaired differentiation of the POMC cell lineages as well as ACTH deficiency (Vallette-Kasic et al., 2007). An understanding of the biological relevance of the enhanced expression of the tbx19 gene in trout testes will require further investigation. We also noted that several transcripts that originated from transposable elements were among the most frequently detected transcripts in trout testes. These transcripts encode reverse transcriptases, transposases and partial envelope proteins (Table 3). Transposable elements have long been considered “junk” DNA, but there is accumulating evidence that they may have been recruited by the progenote to fulfill important biological functions (Volff, 2006, Kleene et al., 1998). For instance, retroviral envelope proteins are required for the cell contact and fusion of human trophoblast cells (Vargas et al., 2009). Moreover, the transposition of active mobile elements in the vicinity of any gene can modulate gene expression, and it has been proposed that these events may participate in the epigenetic regulation of genes in somatic cells (see (Muotri et al., 2007, Collier and Largaespada, 2007) for reviews). In addition, transposition is likely in germ cells, and this could be considered a major source of genetic variation that can be passed on to further generations. The identification of mobile transposable elements in the testes contributes to our understanding of the mechanisms that lead to dynamic changes in the genome
structure and may provide an explanation for the observed differences in the reproductive capacity of individuals.

3.2.3 The identification of genes that are preferentially expressed in muscle

When performing a DDD analysis for the trout muscle samples using a 10-fold cut-off, 393 contigs were found to be overabundant in the white muscle libraries in comparison to the non-muscle libraries (see Additional File 1). Among these, 266 contigs were found only 1 or 2 times in the non-muscle libraries. Of the top ten overabundant contigs (exhibiting overexpression ranging from 2517- to 793-fold), six were found to encode sarcomeric proteins such as myosins and troponins (Table 4). Other sarcomeric protein-encoding genes were found to be overexpressed in the muscle libraries including Lim domain-binding protein 3, nebulin and myozenin-1. Additionally, sarcosin (kbtd10, for Kelch repeat and BTB domain-containing protein 10, see Fig. 1), which is involved in the assembly of myofibrils (Greenberg et al., 2008), and Stars, an actin-binding Rho-activating protein that promotes actin polymerization (Arai et al., 2002), were predominantly expressed in the muscle library. The DDD analysis is also strongly validated by the observation that most of the enzymes identified in the muscle libraries were known to be muscle-specific isoforms or enzymes. Among these enzymes were creatine kinase, Phosphoglycerate mutase (pgam2) (see Fig. 1), Adenylate kinase isoenzyme 1, L-lactate dehydrogenase A, Fructose-bisphosphate aldolase A, 6-phosphofructokinase, Glycogen phosphorylase, AMP deaminase 1 and β-enolase (Table 4). Consistently, the in situ hybridization of β-enolase revealed a strong hybridization signal specifically in glycolytic white fast muscle fibers (Fig. 2). In agreement with existing data, the histone methyl transferase Smyd1 (Tan et al., 2006) and the antiproliferative protein Btg1 (Busson et al., 2005), which both regulate muscle differentiation, were clearly
overexpressed in the muscle libraries. In addition, DDD analysis has led to the
identification of the genes encoding Ryanodine receptor 1 (RyR1) and calsequestrin,
which are known to participate in the formation of a complex that is required for Ca^{2+}
ion release from the sarcoplasmic reticulum. This is an essential step in muscle
contraction (Zhang et al., 1997).

3.2.4 The identification of genes that are preferentially expressed in the brain

When performing a DDD analysis with a 10-fold cut-off in the trout brain samples
that were taken 40 days post-fertilization, 313 contigs were found to be overabundant
in the brain libraries in comparison to the non-brain libraries. In addition, 365 contigs
were found at least 10 times in the brain libraries but were never detected in the non-
brain libraries. At the stage examined, the brain is organized and regionalized, but it is
still actively growing. Accordingly, many of the genes that were identified in the
brain samples are known to be involved in neurogenesis and/or neuronal
differentiation (Table 5). Some of these genes (glast and fabp7a) are likely to be
expressed in the radial glial cells that are the main progenitors at this stage, whereas
others are involved in neuronal growth or neuron differentiation (synapsin,
neuromodulin, neurtin, and neuroD). Among the genes expected to be strongly
expressed in the developing brain is fabp7a, and Brain Lipid Binding Protein (Blbp,
see Fig. 1). fabp7a, is a member of the fatty acid-binding protein family that was
retrieved several times. This finding is of particular interest because, in mice, Blbp is
a recognized marker of radial glial cells, which are known to be progenitor cells in the
developing mammalian brain (Anthony and Heintz, 2008). Recently, it was shown
that blbp is also expressed in the brains of larval and adult fish (Diotel et al., 2010;
Marz et al., 2010). Ependymin (epd1) is a meningeal-derived extracellular matrix
glycoprotein that was also overrepresented in the brain in comparison to other tissues.
Little is known about the function of such proteins; however, they are thought to increase in abundance during periods of neuronal plasticity in mammals (Rother et al., 1995). The hypothesis that ependymin is also involved in neurogenesis in fish is of particular interest and will need to be further examined. The gene slc1a2b encodes a glial high-affinity glutamate transporter that was previously identified in a screen of the zebrafish hypothalamus (Toro et al., 2009). In mammals, these glutamate transporters are strongly expressed in astrocytes. In these cells, the excess of glutamate is quickly removed from the extracellular space, where the glutamate level is kept low. Such transporters are also expressed in the radial glial cells of rodents and, likely, of fishes, although the latter has not been documented. Neuromodulin (gap43) was also found to be expressed only in the brain (see Fig. 1), in agreement with its recognized functions as a 'plasticity' protein that is highly expressed in neuronal growth cones during development (Udvadia et al., 2001). Neurogenic differentiation 1 (neurod1) is a member of the NeuroD family of basic-helix-loop-helix (bHLH) transcription factors that are involved in brain differentiation and patterning. Tubulin alpha chains (tubal) were also highly represented, which was expected based on their known roles in neurotubule formation (Hieber et al., 1998). Neuritin and synapsin, respectively, are involved in neurite growth and synapse formation. Neuritin (nrn1) is expressed in the postmitotic neurons of the developing nervous system and in neuronal structures that are undergoing plasticity in adults (Naeve et al., 1997). Interestingly, neuritin was also represented in the libraries constructed from non-nervous tissues. Synapsins are a family of proteins that are implicated in the regulation of neurotransmitter release at synapses. Recent data suggest that synapsins are not essential for synaptic function, but instead serve important modulatory functions at the synapse (Kile et al., 2010).
3.2.5 The identification of genes that are preferentially expressed in the gills

When comparing gill with other tissues using a DDD analysis with a 10 fold cut-off, we found 313 contigs that were overrepresented in the gill libraries. Moreover, 365 contigs were only expressed in the gill libraries (see Additional File 1). As expected, several of these genes are known to be specifically or predominantly expressed in the gills (Table 6). Included in the gill-specific genes is a new calpain, which was reported to be expressed in the trout gill by Salem and coworkers (Salem et al., 2006). Additionally, we identified several ion transporters such as Na/K-ATPaseδ1a, which was previously shown to be overexpressed in gill tissue by Richards et al. (Richards et al., 2003). Proteins that are known to be regulators of cellular ion transport, such as the FXYD proteins that are known to regulate the substrate affinity of the Na/K-ATPase enzyme (Garty and Karlish, 2006), were also overexpressed in gill tissue. In the Atlantic salmon (Salmo salar), the expression of FXYD11 isoform was reported to be 1,000-fold higher in gill than in other tissues (Tipsmark, 2008). In agreement with this finding, our data revealed that this FXYD11 isoform is overexpressed in gill tissue. Other interesting candidate genes that were found are known to be linked to ammonia excretion. Ammonia is predominantly excreted in the gills in teleost fish. Consequently, it is not surprising to find a higher expression of genes encoding ammonia transporters (Rhesus (Rh) proteins) in gill tissue than in other organs. For instance, we observed the predominant expression of Rh type C glycoprotein (rhcg2) in the trout gill, in agreement with data reported by (Nawata et al., 2007). DDD analysis also allowed us to identify new putative genes that are predominantly expressed in gill tissue in comparison to other tissues. These include the zymogen granule protein 16 (zg16) (see Fig. 1), which is considered to be a secretory lectin that is involved in the regulation of granule secretion (Zhou et al., 2007), and the
envoplakin (*evpl*) and corniferin genes, which are known in mammals to be involved in maintaining the structure of the epithelium. Envoplakin (see Fig. 1) is a component of desmosomes and cornified envelope (CE) and is expressed in keratinizing and non-keratinizing squamous epithelia (Ruhrberg et al., 1996). Cornifelin is also a component of the CE (Michibata et al., 2004). The cornified envelope provides a protective barrier between the environment and the living layers of the skin and is thought to play an important role in the maintenance of the integrity of the epidermis in mammals. The presence of several CE proteins in gill tissue could suggest their importance in maintaining the barrier that protects the gill epithelium from the water environment.

4. Conclusions

The present work utilized 454-Titanium technology to uncover tissue-specific cDNA repertoires from various rainbow trout tissues. This important information has already allowed us to develop an *in silico* data mining program that searches for genes that are differentially expressed between tissues; however, these analyses are still difficult because of the inherent complexity of the *de novo* assembly of these large transcript sequencing datasets. This complexity has been already described (Costa et al., 2010) and involves, for instance, the presence of different splice forms of genes, as well as of allele-specific variants, which can split different contigs after the assembly process. The current lack of a whole genome reference sequence in rainbow trout is a major problem; however, ongoing projects should soon provide this genome resource, allowing researchers to directly map these short read sequences onto the genome, partially solving the difficulties inherent in *de novo* transcriptome assembly.
**References**


receptors and myogenic factors induces the major BTG1 influence on muscle

Collier, L.S. and Largaespada, D.A. Transposable elements and the dynamic somatic

Costa, V., Angelini, C., De, F.I. and Ciccodicola, A. Uncovering the complexity of

Crawford, J.E., Guelbeogo, W.M., Sanou, A., Traore, A., Vernick, K.D., Sagnon, N.
and Lazzaro, B.P. De novo transcriptome sequencing in Anopheles funestus

Desvignes, T., Pontarotti, P., Fauvel, C. and Bobe, J. Nme protein family evolutionary

Diotel, N., Le Page, Y., Mouriec, K., Tong, S.K., Pellegrini, E., Vaillant, C., Anglade,
I., Brion, F., Pakdel, F., Chung, B.C. and Kah, O. Aromatase in the brain of
teleost fish: expression, regulation and putative functions. *Frontiers in

Fraser, B.A., Weadick, C.J., Janowitz, I., Rodd, H. and Hughes, K.A. Sequencing and
characterization of the guppy (Poecilia reticulata) transcriptome. *BMC

Gabillard, J.C., Duval, H., Cauty, C., Rescan, P.Y., Weil, C. and Le Bail, P.Y.
Differential expression of the two GH genes during embryonic development of
rainbow trout Oncorhynchus mykiss in relation with the IGFs system.

Garty, H. and Karlish, S.J. Role of FXYD proteins in ion transport. *Annual review of


Vargas, A., Moreau, J., Landry, S., LeBellego, F., Toufaily, C., Rassart, E., Lafond, J.

and Barbeau, B. Syncytin-2 plays an important role in the fusion of human

Volff, J.N. Turning junk into gold: domestication of transposable elements and the
creation of new genes in eukaryotes. *BioEssays : news and reviews in

Wilhelm, B.T. and Landry, J.R. RNA-Seq-quantitative measurement of expression

Zhang, L., Kelley, J., Schmeisser, G., Kobayashi, Y.M. and Jones, L.R. Complex
formation between junctin, triadin, calsequestrin, and the ryanodine receptor.
Proteins of the cardiac junctional sarcoplasmic reticulum membrane. *The

Z.Q. and Han, Z.G. hZG16, a novel human secreted protein expressed in liver,
was down-regulated in hepatocellular carcinoma. *Biochemical and

**Acknowledgments**

This work was supported by funds from the "Institut National de la Recherche
Agronomique" (AIP bioressources INRA 2009) and by an EC's Seventh Framework
Program (222719-LIFECYCLE). B.N. was supported by a joined Région Bretagne
and INRA PhD fellowship.
**Tables**

**Table 1 - Summary of the sequencing run results.**

<table>
<thead>
<tr>
<th>Tissues</th>
<th>Number of reads</th>
<th>Number of base pairs</th>
<th>Average sequence length</th>
<th>Number of 454 runs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gill</td>
<td>613.223</td>
<td>232.066.637</td>
<td>378 bp</td>
<td>1</td>
</tr>
<tr>
<td>Brain</td>
<td>481.363</td>
<td>159.232.104</td>
<td>331 bp</td>
<td>1.5</td>
</tr>
<tr>
<td>Ovary</td>
<td>738.510</td>
<td>234.665.107</td>
<td>318 bp</td>
<td>1</td>
</tr>
<tr>
<td>Testis</td>
<td>1.234.260</td>
<td>381.485.303</td>
<td>309 bp</td>
<td>1.5</td>
</tr>
<tr>
<td>Total</td>
<td>3.067.356</td>
<td>1.007.449.151</td>
<td>328 bp</td>
<td>5 runs of 454</td>
</tr>
</tbody>
</table>
Table 2 – Selected genes that were over-expressed in the ovarian libraries.
The expression scores (representing the number of reads in the corresponding contig) for the genes that are discussed within the text are shown for the ovarian and non-ovarian libraries. Genes with expression profiles that were confirmed by qPCR are identified in the table with an asterisk (*) following the contig name.

<table>
<thead>
<tr>
<th>Contig name</th>
<th>Best Blast Hit</th>
<th>Ovarian</th>
<th>Non-ovarian</th>
</tr>
</thead>
<tbody>
<tr>
<td>BX082756.p.om.8</td>
<td>ZPC domain containing protein 5</td>
<td>1042</td>
<td>1</td>
</tr>
<tr>
<td>CU072510.p.om.8</td>
<td>ZPC domain containing protein 1</td>
<td>132</td>
<td>0</td>
</tr>
<tr>
<td>GAY7CU01AN5KC.p.om.8</td>
<td>ZPA domain containing protein</td>
<td>68</td>
<td>0</td>
</tr>
<tr>
<td>CB486811.p.om.8*</td>
<td>Protein Wnt-11</td>
<td>80</td>
<td>4</td>
</tr>
<tr>
<td>BX079798.p.om.8*</td>
<td>bucky ball</td>
<td>87</td>
<td>0</td>
</tr>
<tr>
<td>BX074171.p.om.8</td>
<td>oocyte protease inhibitor-2</td>
<td>248</td>
<td>2</td>
</tr>
<tr>
<td>CR373079.p.om.8</td>
<td>oocyte protease inhibitor-2 (opi-2)</td>
<td>170</td>
<td>2</td>
</tr>
<tr>
<td>BX303838.p.om.8</td>
<td>serpin peptidase inhibitor</td>
<td>169</td>
<td>0</td>
</tr>
<tr>
<td>CB490333.p.om.8*</td>
<td>POU domain, class 5, transcription factor 1</td>
<td>40</td>
<td>0</td>
</tr>
<tr>
<td>BX088048.p.om.8</td>
<td>zygote arrest 1 (zar1)</td>
<td>87</td>
<td>3</td>
</tr>
<tr>
<td>CX041779.p.om.8</td>
<td>growth/differentiation factor 9</td>
<td>32</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 3 – Selected genes that were over-expressed in the testicular libraries.
The expression scores (representing the number of reads in the corresponding contig) for the genes that are discussed within the text are shown for the testicular and non-testicular libraries. Genes with expression profiles that were confirmed by qPCR are identified in the table with an asterisk (*) following the contig name.

<table>
<thead>
<tr>
<th>Contig name</th>
<th>Best Blast Hit</th>
<th>Testicular</th>
<th>Non-testicular</th>
</tr>
</thead>
</table>

- 30 -
<table>
<thead>
<tr>
<th>PDB Accession</th>
<th>Gene Description</th>
<th>Expression Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>DQ489287</td>
<td>GSDF</td>
<td>2445</td>
</tr>
<tr>
<td>BX857318</td>
<td>SYCP1</td>
<td>198</td>
</tr>
<tr>
<td>GBPNQUD01A6W4I</td>
<td>Myelin associated glycoprotein</td>
<td>372</td>
</tr>
<tr>
<td>CU063452</td>
<td>AMH</td>
<td>176</td>
</tr>
<tr>
<td>BX080625</td>
<td>Cdk6</td>
<td>124</td>
</tr>
<tr>
<td>BX080735</td>
<td>Cohesin subunit</td>
<td>119</td>
</tr>
<tr>
<td>F0DLFKL01BENKG</td>
<td>Unknown</td>
<td>105</td>
</tr>
<tr>
<td>FYV3OTN02HHPOI</td>
<td>Rec8</td>
<td>106</td>
</tr>
<tr>
<td>BX314487</td>
<td>L-asparaginase</td>
<td>85</td>
</tr>
<tr>
<td>FYV3OTN01B8PA8</td>
<td>ERV-FRD provirus ancestral Env polyprotein</td>
<td>68</td>
</tr>
<tr>
<td>F6VG15V02FTMKM</td>
<td>Histone-lysine N-methyltransferase</td>
<td>61</td>
</tr>
<tr>
<td>FYV3OTN01A1QIV</td>
<td>Zinc finger CW-type PWWP domain protein 1</td>
<td>53</td>
</tr>
<tr>
<td>CB487402</td>
<td>LINE-1 reverse transcriptase</td>
<td>62</td>
</tr>
<tr>
<td>FYV3OTN01E3SYW</td>
<td>Transposable element Tc1 transposase</td>
<td>47</td>
</tr>
<tr>
<td>F6VG15V02HO88H</td>
<td>Pro-Pol polyprotein</td>
<td>88</td>
</tr>
<tr>
<td>CX036404</td>
<td>DAZL</td>
<td>24</td>
</tr>
<tr>
<td>GBPNQUD01BLTP4</td>
<td>TBX19</td>
<td>23</td>
</tr>
<tr>
<td>GBPNQUD01AIF47</td>
<td>Retrotransposon protein 1</td>
<td>138</td>
</tr>
<tr>
<td>BX076816</td>
<td>Envelope protein</td>
<td>51</td>
</tr>
</tbody>
</table>

**Table 4 – Selected genes that were over-expressed in the muscle libraries.**

The expression scores (representing the number of reads in the corresponding contig) for the genes discussed within the text are shown for the muscle and non-muscle libraries. Genes with expression profiles that were confirmed by qPCR are identified in the table with an asterisk (*) following the contig name.
<table>
<thead>
<tr>
<th>Contig name</th>
<th>Best Blast Hit</th>
<th>Muscle</th>
<th>Non-muscle</th>
</tr>
</thead>
<tbody>
<tr>
<td>CA343693.p.om.8*</td>
<td>Phosphoglycerate mutase 2-1 (muscle)</td>
<td>5035</td>
<td>2</td>
</tr>
<tr>
<td>CR372691.p.om.8</td>
<td>myosin, light polypeptide 3-3</td>
<td>2141</td>
<td>1</td>
</tr>
<tr>
<td>CA365356.p.om.8</td>
<td>fast skeletal troponin C</td>
<td>21213</td>
<td>10</td>
</tr>
<tr>
<td>CA345074.p.om.8*</td>
<td>Kelch repeat and BTB domain-containing protein 10</td>
<td>224</td>
<td>2</td>
</tr>
<tr>
<td>CA361268.p.om.8</td>
<td>SET and MYND domain-containing protein 1</td>
<td>103</td>
<td>2</td>
</tr>
<tr>
<td>CX252008.p.om.8</td>
<td>Calsequestrin-1</td>
<td>83</td>
<td>2</td>
</tr>
</tbody>
</table>

Table 5 – Selected genes that were over-expressed in the brain libraries.

The expression scores (representing the number of reads in the corresponding contig) for the genes that are discussed within the text are shown for the brain and non-brain libraries. Genes with expression profiles that were confirmed by qPCR are identified in the table with an asterisk (*) following the contig name.

<table>
<thead>
<tr>
<th>Contig name</th>
<th>Best Blast Hit</th>
<th>Brain</th>
<th>Non-brain</th>
</tr>
</thead>
<tbody>
<tr>
<td>BX082627.p.om.8*</td>
<td>BLBP</td>
<td>1125</td>
<td>95</td>
</tr>
<tr>
<td>BX316810.p.om.8</td>
<td>glial high affinity glutamate transporter</td>
<td>261</td>
<td>1</td>
</tr>
<tr>
<td>CA368669.p.om.8</td>
<td>Neuritin</td>
<td>12</td>
<td>6</td>
</tr>
<tr>
<td>CA381682.p.om.8*</td>
<td>Neuromodulin</td>
<td>176</td>
<td>2</td>
</tr>
<tr>
<td>CX153668.p.om.8</td>
<td>Brain-specific polypeptide PEP-19</td>
<td>51</td>
<td>13</td>
</tr>
<tr>
<td>CA343423.p.om.8</td>
<td>Synapsin</td>
<td>47</td>
<td>7</td>
</tr>
</tbody>
</table>

Table 6 – Selected genes that were over-expressed in the gill libraries.

The expression scores (representing the number of reads in the corresponding contig) for the genes that are discussed within the text are shown for the gill and non-gill libraries. Genes with expression profiles that were confirmed by qPCR are identified in the table with an asterisk (*) following the contig name.
### Contig name | Best Blast Hit | Gill | Non-gill
---|---|---|---
CA345680.p.om.8 | gill-specific calpain | 43 | 1
CA362824.p.om.8 | Rh type C glycoprotein (rhcg2) | 132 | 0
CA345079.p.om.8 | Na/K ATPase alpha subunit isoform 1a | 319 | 21
CA349792.p.om.8 | FXYD domain containing ion transport regulator 11 | 294 | 0
CA365840.p.om.8* | Envoplakin | 58 | 1
BX321817.p.om.8 | cornifelin-like protein b | 57 | 1
CA345392.p.om.8* | Zymogen granule membrane protein 16 | 2075 | 10

### Additional files

**Additional File 1 – Genes that were detected by Digital Differential Display analysis to be predominantly expressed in rainbow trout ovary, testis, brain, muscle and gill tissues.**

This file contains the number of reads per condition for the contigs that were identified to be predominantly expressed in rainbow trout ovary, testis, brain, muscle and gill tissues (one folder page per tissue of interest) using the digital differential display analysis. For each contig containing 10 or more reads in the tissue of interest (FP in contig) and no hits in the other tissues (SP in contig), or for each contig having a fold change of more than 10 (FP/SP), this file provides information on the significance of the enrichment using Fisher’s Exact Test (no value is given for contigs with no hit in the SP condition). The file also provides information regarding the top blast hit results for all of the contig sequences (best hit description, best hit database,
the percentage of sequence coverage with the best hit, and the percentage of sequence conservation with the best hit). For each contig, a dynamic internet link with the corresponding contig view webpage is available (click directly on the contig name; Excel file, .xls).

Additional File 2 – Primer sequences used for qPCR validation of some differentially expressed genes.