Molecular Ecology Resources
January 2014, Volume 14, Issue 1, pages 139–149
http://dx.doi.org/10.1111/1755-0998.12148

Analysis of the black-chinned tilapia *Sarotherodon melanotheron heudelotii* reproducing under a wide range of salinities: from RNA-seq to candidate genes

J.-C. Avarre[1,*], R. Dugué[1], P. Alonso[1], A. Diombokho[1], C. Joffrois[1], N. Faivre[1], C. Cochet[1], J.-D. Durand[2]

[1] Institut des Sciences de l'Evolution de Montpellier, UMR 226 IRD-CNRS-UM2, Montpellier Cedex 05, France
[2] Ecologie des Systèmes Marins Côtiers, UMR 5119 IRD-UM2-CNRS-IFREMER, Montpellier cedex 05, France

*: Corresponding author : Jean-Christophe Avarre, fax. +33467143622 ; email address : jean-christophe.avarre@ird.fr

Abstract :

The black-chinned tilapia *Sarotherodon melanotheron heudelotii* is an ecologically appealing model as it shows exceptional adaptive capacities, especially with regard to salinity. In spite of this, this species is devoid of genomic resources, which impedes the understanding of such remarkable features. *De novo* assembly of transcript sequences produced by next-generation sequencing technologies offers a rapid approach to obtain expressed gene sequences for non-model organisms. It also facilitates the development of quantitative real-time PCR (qPCR) assays for analysing gene expression under different environmental conditions. Nevertheless, obtaining accurate and reliable qPCR results from such data requires a number of validations prior to interpretation. The transcriptome of *S. melanotheron* was sequenced to discover transcripts potentially involved in the plasticity of male reproduction in response to salinity variations. A set of 54 candidate and reference genes was selected through a digital gene expression (DGE) approach, and a *de novo* qPCR assay using these genes was validated for further detailed expression analyses. A user-friendly web interface was created for easy handling of the sequence data. This sequence collection represents a major transcriptomic resource for *S. melanotheron* and will provide a useful tool for functional genomics and genetics studies.

Keywords : gene expression ; non-model organism ; reference gene ; salinity ; spermatogenesis ; tilapia

Introduction

Inland water ecosystems are subjected to natural, seasonal and between-year variations in climate. Depending on their nature, duration and magnitude, these variations have contributed to the evolution of physiological adaptations in fish species. These adaptations consist in modifications of life history traits such as growth, age at sexual maturity (Duponchelle& Panfili 1998; Stearns& Crandall 1984; Stewart 1988), fecundity (Duponchelle *et al.* 2000; Legendre& Ecoutin 1989) or trophic demand (Ogari& Dadzie 1988). In this context, a better delimitation of the adaptive capacities of species, and a deeper understanding of their inner mechanisms, are tremendously needed to determine the threats upon these species. This issue is particularly overwhelming in the field of reproductive biology, which has a more straigntforward impact on fitness than any other biological function.

Species that already perform well or are tolerant to a broad range of environmental conditions are thus excellent templates for investigating the responses to such fluctuations. In this regard, the black-chinned tilapia *Sarotherodon melanotheron heudelotii* Rüppell 1852 (Teleostei, Cichlidae) is supposedly one of the record holders, since it has been reported to reproduce at salinities ranging from 0 to 120 psu (Panfili *et al.* 2004; Panfili *et al.* 2006). This tilapia is a mouthbrooding fish in which the males pick up the fertilized eggs and incubate them until they are released as free-swimming fry. In addition, the black-chinned tilapia *S. melanotheron* is an excellent model for studying the plasticity of reproductive traits since: i) it shows a remarkable adaptation to salinity and is, to our knowledge, the most plastic fish in this respect; ii) natural populations occur in many different habitats from fresh- to hypersaline waters ; iii) in culture conditions, it is capable of spawning spontaneously and has brief and frequent reproduction cycles all year round; iv) it has a relatively small size and it is thus easy to maintain adult fishes in different controlled conditions.

Using suppressive subtractive hybridization from gills of *S. melanotheron*, Tine *et al.* demonstrated how salinities impacted the expression of a small number of genes involved in osmotic homeostasis and energy metabolism (Tine *et al.* 2008; Tine *et al.* 2012), thereby highlighting a plastic regulation of gene expression in the gills. If the plasticity of the reproductive traits of *S. melanotheron* induced by salinity are now acknowledged (Legendre *et al.* 2008; Panfili *et al.* 2006), the underlying biological processes are still poorly understood, mainly because of the lack of genomic resources available for this species. The present study aimed at filling this gap, by generating a large transcript sequence collection. In non-model organisms for which there is no or limited genomic resources, next-generation sequencing (NGS) represents a valuable tool for characterizing genes involved in particular biological functions or traits (Fraser *et al.* 2011; Wang *et al.* 2009). Once a reference transcriptome is available, tag-based sequencing, or digital gene expression (DGE), represents a sensitive and cost-effective alternative for gene expression profiling of specific phenotypes or adaptive traits (Hong *et al.* 2011; t Hoen *et al.* 2008). Then, real-time PCR (qPCR) remains the simplest and probably most accurate method to substantiate quantitative data derived from NGS. Yet, obtaining, analysing and interpreting qPCR data is not a trivial issue, and requires a thorough validation of every step of any *de novo* assay design (Bustin *et al.* 2010). Therefore, the present article describes not only the development of an important transcriptomic resource for *S. melanotheron*, but also the detailed validation of a set of candidate and reference genes that will enable in-depth expression studies on both wild and experimental fish populations, complying with the MIQE (Minimum Information for publication of Quantitative real-time PCR Experiments) guidelines (Bustin *et al.* 2009).

Material and methods

Fish sampling

Natural populations of *S. melanotheron heudelotii* were sampled in Senegal during the dry season (May 2010) at 3 locations along the salinity gradient of the Sine Saloum estuary, namely Missirah (40 psu), Foundiougne (53 psu) and Kaolack (95 psu). Adult fish were caught using a cast net and then anesthetized in icy water. Size (fork length) and weight were measured and the sex determined for each fish. Animals were dissected and a portion of both liver and gonad was immediately immersed in a tube containing 10-20 volumes of RNAlater (Ambion) and placed on ice. Tubes were maintained at 4°C all along the field campaign (3 days), and stored at -20°C upon arrival to the laboratory. The stage of sexual maturity was determined macroscopically according to Legendre and Ecoutin (Legendre& Ecoutin 1989). A total of 10 males and 10 females were collected from each station.

RNA extraction

RNA was extracted with the Nucleospin-8 total RNA isolation kit (Macherey-Nagel). Fifteen to twenty mg of tissue preserved in RNA later (Ambion) were weighed and transferred into 2-ml tubes containing a 5-mm steel bead (Qiagen) as well as 360 µl lysis buffer supplemented with 1% β-mercaptoethanol (Sigma-Aldrich). Tissues were homogenized with a tissue lyzer (Qiagen) for 2 min at 50 Hz. Tubes were then centrifuged for 5 min at 20,000 g and the supernatants were transferred to new tubes and kept at -20°C overnight. RNA was extracted the following day according to the manufacturer's instructions, using a Janus automated Workstation (Perkin Elmer), and eluted in 70 µl RNase-free H_2O. In order to remove any trace of contaminating genomic DNA, RNA eluates were subjected to a second DNase treatment. Briefly, a mix of 0.2 µl RNase-free DNase and 2 µl of reaction buffer (Macherey-

Nagel) was added to 20 µl of each RNA eluate, and digestion was carried out for 15 min at 37°C. RNA quantity was measured by UV spectrophotometry (Nanodrop 1000, Thermoscientific), and its integrity was verified by capillary electrophoresis (Agilent Bioanalyzer 2100). Only samples displaying an RNA integrity number (RIN) ≥ 8 were used for subsequent analyses. Each RNA sample was diluted to a concentration of 50 ng/µl in H_2O and stored at -80°C.

RNA-seq library

Construction and sequencing

A large transcript library was generated from both liver and gonads of wild fish sampled in the Sine Saloum estuary. RNA from liver and gonads of 18 males and 16 females (~5-6 fish per salinity location) were mixed in an equimolar way. Five µg of this RNA mixture were used as template for the construction of a cDNA library, using the Illumina mRNA-Seq Paired-End kit with several modifications. In brief, polyA-containing mRNA molecules were fragmented for 5 min to yield fragments of ~250 bp. Second strand cDNA was synthesized and further subjected to end repair, A-tailing, and adapter ligation in accordance with the manufacturer supplied protocols. Purified cDNA templates were enriched by 15 cycles of PCR for 10 s at 98°C, 30 s at 65°C, and 30 s at 72°C using PE1.0 and PE2.0 primers and with Fastart taq DNA polymerase (Roche). The samples were cleaned using QIAquick PCR purification columns and eluted in 30 µl elution buffer. The purified cDNA library was quantified using Bioanalyzer DNA 100 Chips (Agilent Technology 2100 Bioanalyzer). Cluster generation was performed by applying 4 pM of cDNA to an Illumina 1G flowcell. Hybridization of the sequencing primer, base incorporation, image analysis and base calling were carried out using the Illumina Pipeline.

142

143 *Contig assembly and functional annotation*

144 Analysis and assembly of the RNA-seq library, which consisted in 50-bp paired-end
145 sequences, were performed by Skuldtech Company (www.skuldtech.com). A first assembly
146 was done using Velvet 1.0.09 (*k*-mer = 41). Sequences were then assembled into clusters
147 using MIRA v3.1. Overlapping identity percentage and minimum overlapping length
148 parameters were set to 90 % and 60 bp, respectively, in order to obtain highly reliable
149 consensus sequences. Sequences that could not be assembled at this stage were referred to as
150 singletons and were not taken into consideration in the following steps. In contrast, the
151 resulting contigs were translated into six reading frames and used as a query to search the
152 non-redundant protein databases available at the National Center for Biotechnology
153 Information (NCBI) using the BlastX algorithm with an E-value $\leq 10^{-3}$ (version # 2.2.15,
154 GenBank release number #166) (www.ncbi.nlm.nih.gov). Sequences with BlastX hits were
155 assigned to the following five sequence categories: known, uncharacterized, predicted,
156 unknown or unnamed, and hypothetical proteins. These terms correspond to the "definition"
157 category of available protein sequences deposited on GenBank
158 (http://www.ncbi.nlm.nih.gov/pubmed/). All unique sequences with BlastX hits (E-value $\leq$
159 10^{-3}) were functionally annotated using Blast2GO (http://www.blast2go.org/) by mapping
160 against gene ontology (GO) resources.

161

162 *Construction and sequencing of digital gene expression (DGE) libraries*

163 Two DGE libraries were constructed with testis RNA obtained from five males collected at
164 Missirah (salinity 40 psu) and five males collected at Kaolak (salinity 95 psu). Sequence tag
165 preparation was achieved with Illumina's Digital Gene Expression Tag Profiling Kit

166 according to the manufacturer's protocol (version 2.1B). For each library, 5µg of an
167 equimolar mix of the 5 total RNA samples were incubated with oligodT beads. Synthesis of
168 first- and second-strand cDNA was performed using superscript II reverse transcription kit
169 according to the manufacturer's instructions (Invitrogen). The cDNAs were cleaved using the
170 NlaIII anchoring enzyme. Subsequently, digested cDNAs were ligated with the GEX adapter
171 1 containing a restriction site of MmEI. A second digestion with MmeI was then performed,
172 which cuts 17 bp downstream of the CATG site. At this point, the fragments detached from
173 the beads. Then the GEX adapter 2 was ligated to the 3' end of the tags. In view of enriching
174 the samples with the desired fragments, a PCR amplification with 12 cycles using Phusion
175 polymerase (Finnzymes) was performed with primers complementary to the adapter
176 sequences. The resulting fragments of 85 bp were purified by excision from a 6%
177 polyacrylamide TBE gel. DNA was eluted from the gel debris with NEBuffer 2 by gentle
178 rotation for 2 hrs at room temperature. Gel debris were removed using Spin-X Cellulose
179 Acetate Filter (2 ml, 0.45 mm) and DNA was precipitated by adding 10 ml of 3 M sodium
180 acetate (pH 5.2) and 325 ml of cold ethanol, followed by centrifugation at 13,000g for 20 min.
181 After washing the pellet with 70% ethanol, the DNA was resuspended in 10 ml of 10 mM
182 Tris-HCl (pH 8.5) and quantified using Nanodrop 1000 spectrophotometer. Cluster generation
183 was performed by applying 4 pM of each sample to individual lanes of an Illumina 1G
184 flowcell. After hybridization of the sequencing primer to the single-stranded products, 35
185 cycles of base incorporation were carried out on the 1G analyzer according to the
186 manufacturer's instructions. Image analysis and base calling were performed using the
187 Illumina Pipeline, where sequence tags were obtained after purity filtering. This was followed
188 by sorting and counting the unique tags.
189

Tag comparison between DGE libraries, gene selection and primer design

The sequence files of each DGE library were analyzed by Skuldtech company (Montpellier, France). Comparisons of DGE libraries were performed using the exact number of tags in each library, , and assumed that each tag has an equal chance of being detected (Piquemal *et al.* 2002). The associated statistical values were obtained from Pearson correlations between tag counts and expressed as p-values (Supplementary mathematical appendix). In order to identify potentially differentially expressed genes, the two DGE libraries were scrutinized for tags that showed the most differential counts, using a p-value < 0.001. Only tags showing a minimum of 10 occurrences in at least one of the two libraries were considered. This resulted in 2214 distinct tags that showed different counts (figure 1). Among them, 711 could be assigned to the EST library. Over these 711 tags, 60 were randomly chosen such that they were over-represented in one of the two salinity conditions, with a 2-fold count difference threshold. Conversely, a p-value > 0.1 was applied to identify tags that showed conserved counts between the 2 salinities. Under such conditions, a total of 2959 distinct tags were identified (figure 1), among which 785 could be assigned to the EST library. Twelve of them were selected according to their apparent highest stability. The selected tags were locally blasted against the RNA-seq library, and all the sequences corresponding to the selected tags (100% identity) were aligned with ClustalX v2.1 software using standard settings (Larkin *et al.* 2007). Primers were designed from each resulting consensus sequence with the online RealTime PCR software tool from Integrated DNA Technologies (http://eu.idtdna.com/scitools/Applications/RealTimePCR/), using the following settings: optimal Tm of 62°C, optimal length of 22 nt and optimal GC content of 50%.

cDNA synthesis and real-time PCR

214 Reverse transcription of male RNA extracts was performed with oligodT primers on 250 µg
215 RNA, using the transcriptor first strand cDNA synthesis kit (Roche). A template-primer
216 mixture consisting of 250 µg RNA and 2.5 µM oligodT was denatured at 65°C for 10-min
217 and immediately cooled on ice. The reaction (in 20 µl final) was supplemented with reaction
218 buffer (1X), dNTPs (1 mM each), RNase inhibitor (20 U) and reverse transcriptase (10 U),
219 incubated for 1 hr at 50°C, then heated for 5 min at 85°C and immediately cooled on ice. The
220 resulting cDNAs were diluted 10 times with 180 µl H_2O and stored at -20°C until use.
221 PCR amplifications were carried out in 384-well plates with a LightCycler 480 (Roche) in a
222 final volume of 6 µl containing 3µl of SYBR Green I Master mix (Roche), 2 µl of cDNA and
223 0.5 µM of each primer. Amplifications were performed in duplicate or in triplicate with an
224 initial denaturation step of 10 min at 95 °C followed by 40 cycles of denaturation at 95°C for
225 10 s, annealing at 60°C for 10 s and elongation at 72°C for 10 s. Amplifications were
226 followed by a melting procedure, consisting of a brief denaturation at 95°C for 5 sec, a
227 cooling step at 65°C for 1 min and a slow denaturation to 97°C. Amplification efficiency of
228 each primer pair was calculated from dilution curves generated using serial dilutions (1:1, 1:2,
229 1:5, 1:10, 1:20, 1:50, 1:100) of a unique cDNA pool, consisting of a mix of 12 cDNAs (4
230 cDNAs per - 0, 35 and 70 psu). A linear regression was applied on the resulting dilution
231 curves and the regression coefficient (R^2) as well as the slope were calculated. Primer pairs
232 were validated only when their corresponding R^2 was higher than 0.99. Amplification
233 products were also verified by analyzing the shape of their corresponding melting curve and
234 by measuring their size on agarose gel electrophoresis. Only the primers yielding a single
235 product, without any primer-dimers, were validated. Each qPCR run contained a no-template
236 control for every primer pair. Cycle of quantification (Cq) values were calculated with the
237 LightCycler software, using the second derivative method. Results were expressed as changes

in relative expression according to the $2^{-\Delta\Delta Cq}$ method (Pfaffl 2001). Cq values were first corrected with the amplification efficiency of each primer pair according to the following equation: $Cq_{E=100\%} = Cq_E (\log(1+ E) / \log(2))$, where E is the efficiency and Cq_E the uncorrected Cq values. Then the corrected Cqs of each gene of interest were normalized with the mean Cq of reference genes (ΔCq), and ΔCq values were related to the average ΔCq value of all samples. All qPCR results were analyzed with the GenEx Pro package (MultiD Analyses, Sweden).

Results

Main features of the sequence data

Considering the scarcity of tilapia sequences in public databases, the first step of this study consisted in establishing a large collection of expressed sequences. It was generated from fish collected in the Sine Saloum estuary. To make this transcript collection as comprehensive as possible, individuals from the three locations (with salinities of 40, 53 and 95 psu) and at all stages of sexual maturity were represented. RNA-seq generated a total of 28,981,363 bp. Sequence assembly resulted in 30,022 contigs and 86,291 singletons, and contig length ranged from 150 to > 3000 bp. Nearly 60% of them could be annotated from public databases. The main features of sequence data are displayed in Table 1.

As a starting point to investigate differential gene expression in the testis of fish reproducing under different salinities, two DGE libraries were also constructed from 5 males collected at the locations displaying the most extreme salinities: Missirah (salinity 40 psu) and Kaolak (salinity 95 psu). Their sequencing resulted in a total of 367,813 and 537,303 tags, respectively, and represented 39,687 and 69,499 unique tags. Among these unique tags, 7,119 and 11,850 could be assigned to the transcript library, respectively.

All this sequence data was organized into an interactive navigation system. This platform includes a sequence viewer that enables exploration of consensus sequences, gene families, putative associated proteins, SNPs or allelic mutations, as well as a local BLAST alignment tool to search for peptidic and nucleotidic motifs in the database. It also allows comparisons of DGE libraries under various stringency conditions. In addition, it gives access to raw sequence data and allows exportation of sequences in fasta format. This platform has been made publicly available and can be accessed through the following address: http://www.skuldtech.com/tilapia/tilapia_menu.php. Details about the functions of this platform may be provided upon request.

Primer validation

The 72 novel primer pairs designed in the present study were first verified for their ability to amplify one single product with an acceptable efficiency. Under the conditions tested, 15 primer sets gave rise to either a lack of amplification or secondary products, as revealed by melting curve analysis and agarose gel electrophoresis. Furthermore, 3 additional pairs yielded poor amplification efficiencies, with linear regression coefficients < 0.99. For these reasons, 18 primer sets were excluded from the analyses. Amplification efficiencies of the 54 remaining primer pairs (43 potential genes of interest and 11 potential reference genes) ranged between 0.8 and 1.1. The sequence of these primers, together with the amplicon length and the amplification efficiency are displayed in Table 2.

Because genomic information regarding intron-exon boundaries was not available for *Sarotherodon melanotheron*, it was not possible to design primers spanning different exonic regions. For this reason, two DNase treatments were applied on each RNA sample: one directly on the columns during the extraction procedure, and a second one in solution, on the

RNA eluates. Relatively high levels of background genomic DNA were detected in single DNAse treated RNA extracts (Cq ranged from 23.9 to 31.5). This signal was not detected in twice DNAse treated samples (Cq>35), indicating the necessity of 2 DNAse treatments for elimination of genomic DNA in cDNA samples.

The optimal primer concentration was also assessed. For each primer pair, 4 concentrations (0.25, 0.5, 0.75 and 1 µM) were tested. Comparison of amplification plots showed that Cq values were steady for the 3 highest concentrations, whereas they were in most cases higher for 250 nM. Besides, melting profiles indicated the absence of primer-dimer or secondary peak for all tested concentrations. For these reasons, primers were used at a final concentration of 500 nM in all subsequent experiments.

Estimation of experimental reproducibility

As for any new assay, evaluation of the experimental biases that may impair quantitative results is also essential. To address this critical issue, experimental reproducibility was first assessed through the following nested protocol: RNA was extracted in duplicate, and reverse-transcription and qPCR were both performed in triplicate, which resulted in 18 Cq measurements per sample and per gene. This protocol was applied with 2 different genes (transcript_AVA2_10563 and transcript_AVA3_453) that produced nearly similar mean Cq values (~20), on 3 individual fish samples originating from 3 distinct salinities, and repeated two times independently. Results revealed that the highest source of variation (expressed as SD of Cqs), after that originating from samples, could be attributed to the reverse transcription reaction (SD ranged from 0.095 to 0.409); conversely, RNA extraction produced the lowest variation (SD ranged from 0 to 0.166), while SD of qPCR repeats varied from 0.076 to 0.130. When the same experiment was repeated using 1 µl of template cDNA instead

of 2 µl, the SD of qPCR replicates dramatically increased as it varied from 0.283 to 0.369. For this reason, the amount of cDNA used was always 2 µl, as stated in the MM section.

Since reverse transcription was the main source of variability, we also evaluated its reproducibility across a range of RNA concentrations. For this purpose, serial dilutions (1:1, 1:2, 1:5, 1:10, 1:20, 1:50, 1:100) were prepared from a pool of RNAs (50 ng/µl) and each dilution (50 ng/µl to 0.5 ng/µl) was reverse transcribed. The corresponding cDNAs were amplified with 2 primer pairs (transcript_AVA2_10563 and transcript_AVA3_453) and Cq values were plotted against the logarithm of the initial RNA concentration. This experiment was repeated twice independently. In each case, it revealed a good linearity with a $R^2 \geq 0.99$ and amplification efficiencies were comprised between 0.63 and 0.86. Results obtained with the primer pair transcript_AVA3_453 are displayed in Figure 2.

Taken altogether, these results suggest that our experimental workflow is trustworthy and should not supply substantial experimental variability to the biological results.

Validation of reference genes

Another crucial point consisted in validating suitable genes that could be used as appropriate reference genes for subsequent relative quantifications. The twelve genes previously identified as potential references were assayed with geNorm (Vandesompele *et al.* 2002) and NormFinder (Andersen *et al.* 2004) algorithms. For this end, their Cq values were measured in a set of 12 fish samples collected from 3 salinities (0, 35 and 70 psu, 4 fish / salinity), and their stability examined. Only the genes displaying an M-value < 0.55 (with geNorm software) and an SD < 0.5 (with NormFinder software) were conserved, which resulted in the selection of 5 of them. Then, these 5 genes were amplified in a set of 22 fish originating from the 3 salinity locations. Analysis of their expression stability across samples with geNorm and

NormFinder (both ignoring and taking salinity groups into account) revealed very congruent results. It led to the validation of 4 of these genes showing an M-value < 0.42 and an SD < 0.36 (Table 3). According to NormFinder, using these 4 genes as reference instead of only one would decrease the accumulated SD by nearly 2 (Table 3).

Discussion

The present study is the first large-scale analysis of the transcriptome of the black-chinned tilapia, an ecologically fascinating fish species with exceptional adaptive capacities, especially in regards to its reproductive behavior. Since *Sarotherodon melanotheron* is a non-model species, the procedure was divided in two separate stages, in order to provide a resource that will be valuable in any study dealing with the reproduction of this fish. The first one consisted of building a large transcript collection from two major organs involved in reproduction, that is liver and gonads (Mommsen& walsh 1988; Wiegand 1996), collected from fish at all stages of sexual maturity and under different salinities. The purpose of this collection was neither to evaluate the total number of transcripts expressed in these organs nor to estimate a transcriptome coverage, as is now the case for most RNA-seq projects in model animals (Wang *et al.* 2009), but rather to provide a first genomic resource in this species for which only a very limited number of sequences were available so far (Tine *et al.* 2008). It is worth mentioning that when these newly obtained sequences were annotated, the genome sequence of the Nile tilapia, *Oreochromis niloticus*, had not yet been released, which is the reason why almost none of these annotations that can be found on the tilapia database website (http://www.skuldtech.com/tilapia/tilapia_menu.php) refer to *O. niloticus*. However, a new annotation round performed on a selected set of sequences revealed no major changes in the protein prediction (data not shown), probably because the genome annotation of *O. niloticus*

was performed automatically. Moreover, the 'export' function of the sequence viewer enables easy updates of alignments and annotations.

In contrast, the second stage of this project aimed at addressing a more specific question on the reproductive biology of *S. melanotheron*, i.e. the identification of genes in male gonads subjected to changes in their expression according to salinity. As demonstrated by several groups, DGE is particularly suited for quantification of transcript abundance (Asmann *et al.* 2009), especially in non-model organisms for which no reference genome is available (Hong *et al.* 2011). For this reason, DGE libraries were compared between fish living in the most extreme salinity environments of the Sine saloum estuary. The library comparison enabled identification of hundreds of genes potentially differentially expressed between the two environments. This list of genes is likely to serve as a wealthy basis for the deeper understanding of the molecular mechanisms that allow *S. melanotheron heudelotii* to reproduce in such a wide range of salinities. Furthermore, a set of 43 genes of interest has been validated in the present work. Even though analysis of their putative role in the adaptation of male spermatogenesis to salinity is beyond the scope of this article and will be the focus of a complementary study, a first look at their predicted function indicates that several of them have already been described as playing a key role in spermatogenesis or in homeostasis. For instance, contig_Tilapia_90_27008 matches a MORC family CW-type zinc finger 2 protein, which absence was shown to trigger the stop of spermatogenesis in mice (Perry& Zhao 2003); contig_Tilapia_90_21432 corresponds to a seminal plasma glycoprotein, which harbors the faculty to immobilize sperm cells in mice as well (Mochida *et al.* 2002); contig_Tilapia_90_947 corresponds to a Na+/K+-transporting ATPase subunit alpha-1, which is involved in the active ion excretion and uptake for maintaining the

intracellular ionic balance (Lorin-Nebel *et al.* 2012). Finally, of these 43 genes, six did not match any known protein.

Although often overlooked, validation of candidate genes identified from NGS data by lab-bench-scale routine methods, such as real-time PCR, requires a number of prior evaluations. This is especially true for the accurate selection of reference genes in relative expression, as it was extensively demonstrated that stability of housekeeping genes greatly depend on the species, tissue, developmental stage, and experimental conditions (McCurley& Callard 2008; Tang *et al.* 2012). Here, the use of geNorm and NormFinder algorithms led to very congruent results, and identified 4 genes as the most stably expressed accross fish individuals and environmental salinities. It also indicated that using 4 genes simultaneously would result in lower standard deviations. Those 4 genes, which could be attributed a putative function with good confidence, all belong to the list of potential housekeeping genes described in humans (Eisenberg& Levanon 2003).

It is acknowledged that the reverse transcription step accounts for a large part of variability in a qPCR assay (Bishop *et al.* 1997). In order to limit this bias, all the primers were selected in the most 3' region of the transcripts; this was made easy by the 3' tag approach that was used for DGE. Then, combined with the use of oligo-dTs, this dramatically reduced the probability to obtain cDNAs that could not be amplified by the designed primers because of incomplete reverse transcription. Although reverse transcription was the main source of variability in the present case, it was yet very limited, as illustrated by the RNA dilution curve that showed a good linearity. Likewise, cDNA dilution curves showed excellent linearity over two logs for the 54 primer pairs. The Cq values measured with all the primer pairs on 22 individual cDNAs from fish collected at different salinities were all comprised within this range (not shown). This indicates that the range of dilutions used to measure the amplification

efficiencies was sufficient to cover most, if not all, RNA concentrations of the targeted genes that can be found in individual samples. This was expected since the sample used for dilutions consisted of a mix of 12 different cDNAs, and as such was supposed to comprise most expressed RNAs at highly variable concentrations.

In conclusion, the present study has generated a large transcriptomic resource that will be valuable for a great number of studies focusing on the functional genomics of this interesting fish, and more broadly of any species presenting salinity-related plasticity. It also identified and validated a large set of genes that will provide a significant tool for the deeper understanding of the molecular mechanisms that allow *S. melanotheron heudelotii* to reproduce in a wide range of salinities. Finally, this resource will also provide useful tools for population genetics studies on *S. melanotheron* (Consortium *et al.* 2013), but also on many other phylogenetically-related species.

Acknowledgements

This study was supported by an INSU-EC2CO grant (IPREP, 2010-2012). We are thankful to Dr. Bruno Guinand for critical reading of this manuscript. We are also grateful to Laurent Manchon and Fabien Pierrat for their input regarding mathematical treatment of the sequence data.

References

Andersen CL, Jensen JL, Ørntoft TF (2004) Normalization of Real-Time Quantitative Reverse Transcription-PCR Data: A Model-Based Variance Estimation Approach to

428 Identify Genes Suited for Normalization, Applied to Bladder and Colon Cancer Data
429 Sets. *Cancer Research* **64**, 5245-5250.
430 Asmann YW, Klee EW, Thompson EA, *et al.* (2009) 3 ' tag digital gene expression profiling
431 of human brain and universal reference RNA using Illumina Genome Analyzer. *BMC*
432 *Genomics* **10**, 11.
433 Bishop GA, Rokahr KL, Lowes M, *et al.* (1997) Quantitative reverse transcriptase-PCR
434 amplification of cytokine mRNA in liver biopsy specimens using a non-competitive
435 method. *Immunology and Cell Biology* **75**, 142-147.
436 Bustin SA, Beaulieu JF, Huggett J, *et al.* (2010) MIQE precis: Practical implementation of
437 minimum standard guidelines for fluorescence-based quantitative real-time PCR
438 experiments. *Bmc Molecular Biology* **11**, 5.
439 Bustin SA, Benes V, Garson JA, *et al.* (2009) The MIQE Guidelines: Minimum Information
440 for Publication of Quantitative Real-Time PCR Experiments. *Clinical Chemistry* **55**,
441 611-622.
442 Consortium MERPD, Arranz SE, Avarre JC, *et al.* (2013) Permanent Genetic Resources
443 added to Molecular Ecology Resources Database 1 December 201231 January 2013.
444 *Molecular Ecology Resources* **13**, 546-549.
445 Duponchelle F, Cecchi P, Corbin D, Nunez J, Legendre M (2000) Variations in fecundity and
446 egg size of female Nile tilapia, Oreochromis niloticus, from man-made lakes of Cote
447 d'Ivoire. *Environmental Biology of Fishes* **57**, 155-170.
448 Duponchelle F, Panfili J (1998) Variations in age and size at maturity of female Nile tilapia,
449 Oreochromis niloticus, populations from man-made lakes of Cote d'Ivoire.
450 *Environmental Biology of Fishes* **52**, 453-465.

451 Eisenberg E, Levanon EY (2003) Human housekeeping genes are compact. *Trends in*
452 *Genetics* **19**, 362-365.
453 Fraser BA, Weadick CJ, Janowitz I, Rodd FH, Hughes KA (2011) Sequencing and
454 characterization of the guppy (Poecilia reticulata) transcriptome. *BMC Genomics* **12**,
455 14.
456 Hong LZ, Li J, Schmidt-Kuntzel A, Warren WC, Barsh GS (2011) Digital gene expression for
457 non-model organisms. *Genome Research* **21**, 1905-1915.
458 Larkin MA, Blackshields G, Brown NP*, et al.* (2007) Clustal W and clustal X version 2.0.
459 *Bioinformatics* **23**, 2947-2948.
460 Legendre M, Cosson J, Hadi Alavi SM, Linhart O (2008) Activation of sperm motility in the
461 euryhaline tilapia *Sarotherodon melanotheron heudelotii* (Dumeril, 1859)
462 acclimatized to fresh, sea and hypersaline waters. *Cybium*, 181-182.
463 Legendre M, Ecoutin JM (1989) Suitability of brackish water tilapia species from the Ivory
464 Coast for lagoon aquaculture. I - Reproduction. . *Aquatic Living Resources* **2**, 71-79.
465 Lorin-Nebel C, Avarre JC, Faivre N*, et al.* (2012) Osmoregulatory strategies in natural
466 populations of the black-chinned tilapia Sarotherodon melanotheron exposed to
467 extreme salinities in West African estuaries. *Journal of Comparative Physiology B-*
468 *Biochemical Systemic and Environmental Physiology* **182**, 771-780.
469 McCurley A, Callard G (2008) Characterization of housekeeping genes in zebrafish: male-
470 female differences and effects of tissue type, developmental stage and chemical
471 treatment. *Bmc Molecular Biology* **9**, 102.
472 Mochida K, Matsubara T, Andoh T*, et al.* (2002) A novel seminal plasma glycoprotein of a
473 teleost, the Nile tilapia (Oreochromis niloticus), contains a partial von Willebrand

474 factor type D domain and a zona pellucida-like domain. *Molecular Reproduction and*
475 *Development* **62**, 57-68.
476 Mommsen TP, walsh PJ (1988) Vitellogenesis and oocyte assembly. In: *Fish Physiology, Vol*
477 *XI, The Physiology of Developing Fish, Part A. Eggs and Larvae* (eds. Hoar WS,
478 Randall DJ), pp. 347-406. Academic press, San Diego.
479 Ogari J, Dadzie S (1988) The food of the Nile perch, Lates niloticus (L.), after the
480 disappearance of the haplochromine cichlids in the Nyanza Gulf of lake Victoria
481 (Kenya). *Journal of Fish Biology* **32**, 571-577.
482 Panfili J, Mbow A, Durand JD*, et al.* (2004) Influence of salinity on the life-history traits of
483 the West African black-chinned tilapia (*Sarotherodon melanotheron*): Comparison
484 between the Gambia and Saloum estuaries. *Aquatic Living Resources* **17**, 65-74.
485 Panfili J, Thior D, Ecoutin JM, Ndiaye P, Albaret JJ (2006) Influence of salinity on the size at
486 maturity for fish species reproducing in contrasting West African estuaries. *Journal of*
487 *Fish Biology* **69**, 95-113.
488 Perry J, Zhao Y (2003) The CW domain, a structural module shared amongst vertebrates,
489 vertebrate-infecting parasites and higher plants. *Trends in Biochemical Sciences* **28**,
490 576-580.
491 Pfaffl MW (2001) A new mathematical model for relative quantification in real-time RT-
492 PCR. *Nucleic Acids Research* **29**, 6.
493 Piquemal D, Commes T, Manchon L*, et al.* (2002) Transcriptome analysis of monocytic
494 leukemia cell differentiation. *Genomics* **80**, 361-371.
495 Stearns SC, Crandall RE (1984) Plasticity for age and size at sexual maturity: a life-history
496 response to unavoidable stress. In: *Fish reproduction: strategies and tactics.*, pp. 13-
497 33.

498 Stewart KM (1988) Changes in condition and maturation of the Oreochromis niloticus L.
499 population at Freguson's Gulf, Lake Turkana, Kenya. *Journal of Fish Biology* **33**,
500 181-188.
501 t Hoen PAC, Ariyurek Y, Thygesen HH, *et al.* (2008) Deep sequencing-based expression
502 analysis shows major advances in robustness, resolution and inter-lab portability over
503 five microarray platforms. *Nucleic Acids Research* **36**, 11.
504 Tang Y-k, Yu J-h, Xu P, *et al.* (2012) Identification of housekeeping genes suitable for gene
505 expression analysis in Jian carp (Cyprinus carpio var. jian). *Fish & Shellfish*
506 *Immunology* **33**, 775-779.
507 Tine M, de Lorgeril J, D'Cotta H, *et al.* (2008) Transcriptional responses of the black-chinned
508 tilapia Sarotherodon melanotheron to salinity extremes. *Marine Genomics* **1**, 37-46.
509 Tine M, Guinand B, Durand JD (2012) Variation in gene expression along a salinity gradient
510 in wild populations of the euryhaline black-chinned tilapia Sarotherodon
511 melanotheron. *Journal of Fish Biology* **80**, 785-801.
512 Vandesompele J, De Preter K, Pattyn F, *et al.* (2002) Accurate normalization of real-time
513 quantitative RT-PCR data by geometric averaging of multiple internal control genes.
514 *Genome Biology* **3**, research0034.0031 - research0034.0011.
515 Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics.
516 *Nature Reviews Genetics* **10**, 57-63.
517 Wiegand MD (1996) Composition, accumulation and utilization of yolk lipids in teleost fish.
518 *Reviews in Fish Biology and Fisheries* **6**, 259-286.
519

520 **Figure legends**

521

522 Figure 1: Comparison of the tag counts in the 2 DGE libraries constructed from fish at two
523 salinity extremes (40 and 95 psu). Comparison was obtained from Pearson correlations
524 between actual tag counts and results expressed as p-values. A low p-value indicates a high
525 level of biological significance. For easier visualization, values were normalized to the
526 number of total tags of each library (count/total number of reads * 10000), and each tag was
527 color-coded according to its representation in each library (reflected by its p-value).

528

529 Figure 2: Reproducibility of the reverse transcription step. Serial dilutions were prepared from
530 a pool of RNAs and each dilution (50 ng/µl to 0.5 ng/µl) was reverse transcribed. The
531 corresponding cDNAs were amplified with the primer pair transcript_AVA3_453 (A) and Cq
532 values were plotted against the logarithm of the initial RNA concentration (B).

533

Data Accessibility

RNA-seq and DGE libraries have been organized into an interactive database that is freely accessible (http://www.skuldtech.com/tilapia/tilapia_menu.php). Moreover, the whole project, including raw DNA sequences, can be found under the SRA study accession SRP022935.

Author contributions

JCA and JDD designed the project. RD, PA, AD, CJ and NF contributed to the experiments. RD and CC were in charge of fish care. JCA, PA, AD and CJ analyzed the data. JCA and JDD wrote the paper.

544 **Table 1** Summary statistics of the RNA-seq data

Statistics for contigs	
Number of bases in all contigs	11 368 093
Number of contigs	30 022
Number of contigs in N50	8 880
Minimum contig length	150
Maximum contig length	3 099
Median contig length	423
GC content of contigs	53.13%
Statistics for singletons	
Number of bases in all singletons	17 613 270
Number of singletons	86 291
Number of singletons in N50	31 086
Median singleton length	202
GC content of singletons	52.63%

545

546

547

548 **Table 2** List of the genes validated by qPCR

Sequence name[a]	Primer sequence	Amplicon length	Amplification efficiency
Potential candidate genes			
Transcript_AVA2_6155	TGTCAGAGGGTAAAGCAAAGGG	127	0,97
	GTAACTTTCCCACACGCCCA		
Transcript_AVA1_52992	GCCACACAGAACAATGGGAATG	102	0,98
	ATCATGTTCGACGTCACTTCTCG		
Transcript_AVA1_55478	CTAACAGAGGATGAGACGGGTG	142	1,00
	TGACTTGTGGCTGCAGACTAC		
Transcript_AVA2_10563	ATTGCTCCTTGACGTACCCAC	113	1,01
	GAAAGACGTCCACCTGGCC		
Transcript_AVA2_4300	GCTGGAATTGCACTCAACGAC	149	1,01
	TGTGACATCCAGGTGAAGGAATG		
Transcript_AVA3_47460	ACGTTGGAGTTGAGTGTCATGG	115	0,97
	CTGAGAGGATGTGTTATCTGGCG		
Contig_Tilapia_90_6346	AATTCTCCCTCATTGTCGCCG	128	0,98
	CAGGTCTTGAGGCATTTTGTTCC		
Transcript_AVA1_4937	AGAGCCCTGGAACAAACTTGG	128	1,04
	CTGCCGATCTTTGTGCTTGTG		
Contig_Tilapia_90_14414	GAACCAGCGTGAACTTTGCAG	114	0,92
	ACCGGACCTTATCATTCTTGGC		
Transcript_AVA1_64597	GCTGTTCAAAATCCCACAAGG	105	0,96
	TCTCCAAGATGTTATCCATAGTGTG		
Contig_Tilapia_90_23367	TCCTCCTCATCCTCCCCTTC	111	0,97
	ACATTCATAGGCACTCCGGTG		

Contig_Tilapia_90_42	TGGACAGGAAGCAATGAGGAAG TCCAGCCTAAAGACTTTCCTGC	135	1,04
Transcript_AVA1_66083	CAAAGGAGCTTGATGCTATTGTA ACCCTGCAAATGTTCTCTTTC	116	1,10
Contig_Tilapia_90_10837	TGACCAGGCTCAGTTCAAAGATG CTGTCTGCAACTCTGGGTAAGG	109	1,01
Transcript_AVA3_28576	CATCCCACTTTGGCAGAAAACAG GCTGCTGTCATTTATTCAACACC	152	0,99
Contig_Tilapia_90_27008	GAGAACACAGGCACGGAAGAG TGGATGACAGGCTCAGTTCAATG	110	1,00
Contig_Tilapia_90_26617	TCCTCGGCTACATGCAATTACG GGCCGAACAGGCTCTTTTATG	109	0,86
Transcript_AVA3_18623	CACATGCAAGAGAAACAAGGAGC TGCCACCTTTTCCCATCCTTG	107	0,92
Contig_Tilapia_90_2777	ACCATCACCAACGATAAGGGC CGGCGATTTTGTCCCTCTGAAG	109	0,88
Contig_Tilapia_90_2321	CTGGAGCTGTAAGTGGGTGAC GCTTGTTAAAACCTGGGCGTC	129	0,99
Contig_Tilapia_90_10643	GAGTGGGCTAACAATGTCAAACG TTATTCCCAGTTCCTGCAGAGTG	106	0,85
Transcript_AVA1_24409	TCTTTGGAGGGAACATGGTGTAC GTGCTGTGACTCTGTCGGAAG	102	0,90
Contig_Tilapia_90_8343	AGAATCAGTGCCGTCCTGTTC CGATGAGGCACACCAGTATATCC	125	0,82
Contig_Tilapia_90_2464	GACCTTCCTCTGAGTGTGATGC CCAAAATCTGAAGCTGTGCGTG	128	0,86
Transcript_AVA1_35277	CGTGGCTATGGACAATTTTGGG	107	0,87

	CCTCGGCAAAAGTCAGCAAAAG		
Contig_Tilapia_90_2942	CTCTGCCCTTCTATCTGTGTTCG	111	1,02
	TCAGTCCGTTCAGTCCTCTCC		
Transcript_AVA3_14200	TGGAGCAAACAGGAAGAGAAGG	114	0,91
	CCCTGTCTTCGGAAACCAATTG		
Transcript_AVA3_33497	CTACATGCTCGGAGGGAAGATTG	115	1,01
	AGATATGGTAGAGTAGTAGGACGCC		
Transcript_AVA2_28399	ACGGTGTACTTGGACATTCAGG	136	1,00
	AAAGCAAAGGGAAGACCGGAG		
Contig_Tilapia_90_21432	ACAGAACTCGTGATCGCTGC	102	0,84
	TGCAGTCTACACAACCACACTC		
Contig_Tilapia_90_6938	TGCTCTGAACAGTTTGGGCTC	116	0,88
	ATGAGAAGCTGGTAACCGTGTG		
Contig_Tilapia_90_2253	ACCCACACCAAACTGACCAATC	117	0,89
	ACCAAAGCCGACCTCATTAACA		
Contig_Tilapia_90_2414	AGTCGGGATGGCTGGATTTG	142	0,91
	ACCAACAGTCATTGCTCCCAC		
Contig_Tilapia_90_1393	CTTTCACACCCTCTTCCCTCG	149	0,84
	CACCAACATTGAGCTGGCAAC		
Transcript_AVA1_28773	CTCCTTCTCACCCGGCAG	104	0,86
	TGGCCTCACATTCAGCCTTG		
Transcript_AVA1_9958	CGAGAACGTCAGAGAAGGTGC	150	0,86
	GAAATTTGGCAGCTCGTGGC		
Contig_Tilapia_90_2469	ATTCCGACGCCTTCTCAACC	122	0,86
	CATGCCGACCCAAACATAAATCG		
Transcript_AVA1_58357	AGACAAGAGTGCCAACATCCAG	116	0,86
	TGAGTTTGGTCTGGTTCTTGAGC		

Transcript_AVA3_142	ATTGAGAACCCCAACAGAGTGG TCCTTTGCCTTCTGCTTCTCG	119	0,87
Contig_Tilapia_90_8891	CTGGTATTGTTAAAAGGCTGGCC TCATTCTGGACTCCTGCAACAC	150	0,90
Contig_Tilapia_90_947	GTATGTCTCTTTCTCCACCCAACC ACGTTGCCCTCAGAATGTACC	143	0,85
Contig_Tilapia_90_9321	ACAGTTCATCGGACAGGTTCAG AGCGACAGAGTGAAATCATGGAC	150	0,95
Contig_Tilapia_90_26561	TCCGCACCCTTAAACTCACAAC GGAATGGCACCATGTTTACAGC	147	0,88
Potential reference genes			
Transcript_AVA2_1624	AAGGACTGGCTTATGCTGATAA GTGATTCAGGTAAGTGACAATGC	86	1,01
Contig_Tilapia_90_13722	GCAATTCGGCTTTCCATGACAG AGCAGAGATAGACATGATTTGGGAG	135	0,96
Transcript_AVA3_453	TGCACACTCTCAAAGATCCCG AGTGACAGAGCCCAAGAAACG	105	1,01
Contig_Tilapia_90_1351	TCGGATAACATCGCCACACTG GAGCCATCTGTCATCTAAACCCTC	148	1,00
Transcript_AVA2_10375	AAATTCATTGGTTTGGAGCGGC GCCTGCTAGTGGGATACCATTC	110	1,04
Contig_Tilapia_90_29868	GTGCCCGAAACAATCCAGAAAC TCTGAACTCGAACTGCCGAAC	149	0,93
Contig_Tilapia_90_1736	CTGCGGTCTAAGATGAGCCAAG AACGGTGGATGCTGGATACTTG	126	1,07
Contig_Tilapia_90_21150	AGGCCACTCAAACATCACCC GTTTGCCCTGGCTTTGATGTG	110	0,97

Transcript_AVA3_16746	GATGTTCTTATTGCAGACGGTGG ACTGCGGACTCTGTGTAATTGC	118	0,97
Contig_Tilapia_90_7452	TTTTGATCGTAGTCGTGGGCTC CATGAGGATGCTTGTGGAAGATAAAC	102	0,87
Contig_Tilapia_90_3058	CATGTTGCTTTCTGCCTCAGTG CGCATCTCCGAGCAGTTCAC	108	0,91

[a] Names of the sequences as they appear on the web sequence viewer (http://www.skuldtech.com/tilapia/tilapia_menu.php)

Table 3 Expression stability of the 4 selected reference genes

	geNorm	NormFinder	
Sequence name	M-value	SD	Acc. SD
Contig_Tilapia_90_7452	0,337	0,297	0,297
Contig_Tilapia_90_13722	0,395	0,319	0,218
Transcript_AVA3_453	0,412	0,352	0,187
Contig_Tilapia_90_3058	0,337	0,355	0,166

A low M-value or standard deviation indicates high expression stability. The combined use of the 4 genes as reference decreases the accumulated standard deviation.

Figure 1

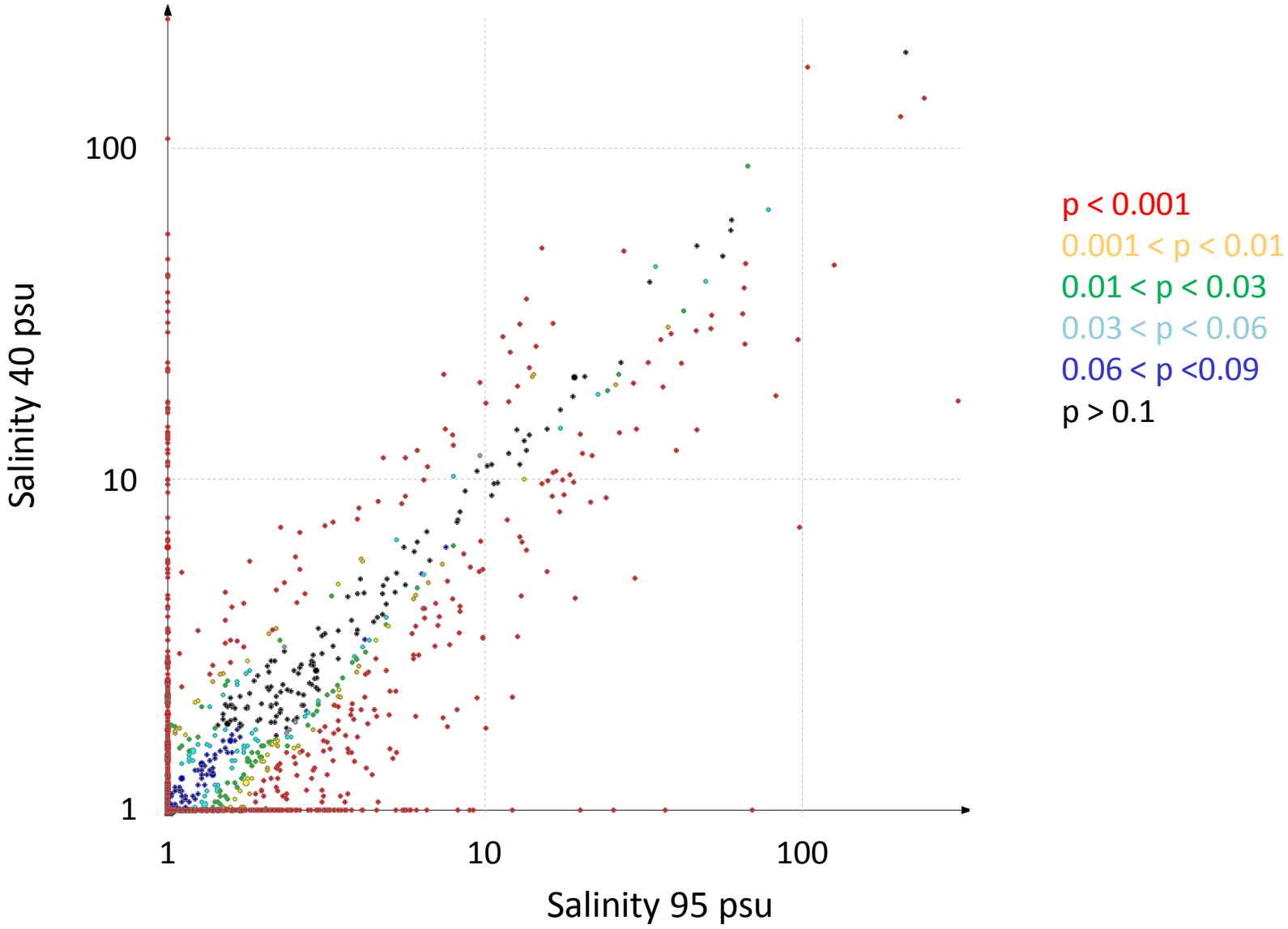
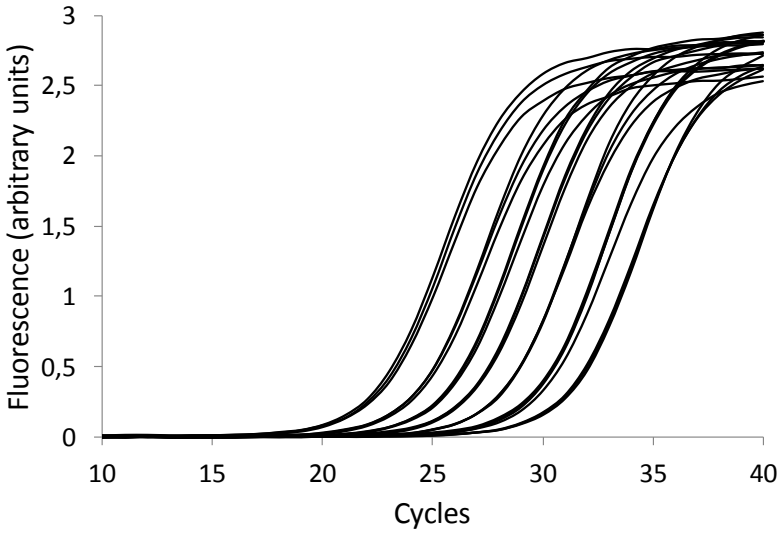
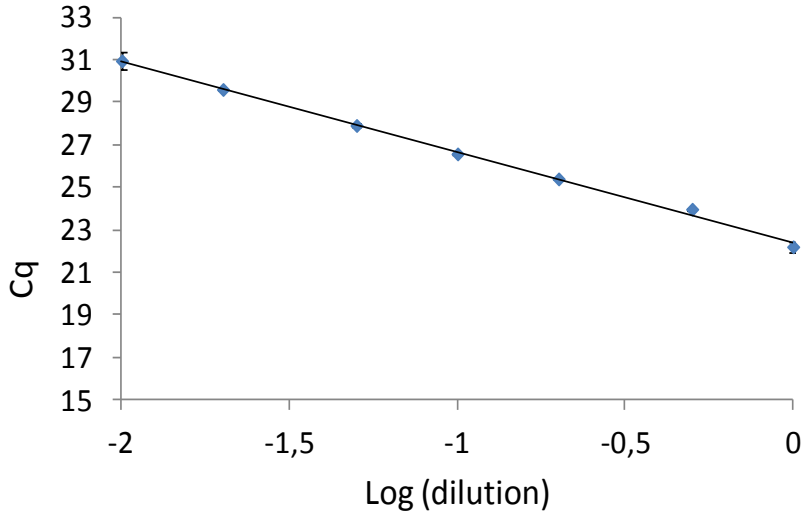

100
10
1
Salinity 40 psu
1
10
100
Salinity 95 psu
p < 0.001
0.001 < p < 0.01
0.01 < p < 0.03
0.03 < p < 0.06
0.06 < p <0.09
p > 0.1

Figure 2

A)


Fluorescence (arbitrary units)
3
2,5
2
1,5
1
0,5
0
10
15
20
25
30
35
40
Cycles

B)


Cq
33
31
29
27
25
23
21
19
17
15
-2
-1,5
-1
-0,5
0
Log (dilution)

Mathematical approach: analysis of differential expression

When comparing two populations of tags, the problem is to distinguish random fluctuation from a biologically significant change. One must calculate the probability for a given tag to be observed x times in a sample of N_1 elements and y times in a sample of N_2 .Using actual experimental series of tag counts, we checked (data not illustrated) that registered variations were in fair agreement with the assumption that x values are binomially distributed:

$$P_1(x\,;\,N_1) = C_{N_1}^{x}\; p^{x}\,(1-p)^{N_1-x} \quad (1)$$

with: $C_N^x = \dfrac{N!}{x!(N-x)!}$, $\boldsymbol{p}$ being the mathematical individual probability.

When performing an other round of analysis and picking up y tags among N_2, the probability is:

$$P_2(y;\,N_2) = C_{N_2}^{y}\; p^{y}\,(1-p)^{N_2-y} \quad (2)$$

and, when picking up $x + y$ times the same tag in $N_1 + N_2$, the probability is :

$$P_{1,2}(x\,+\,y\,;\,N_1\,+\,N_2) = C_{N_1+N_2}^{x+y}\; p^{x+y}\,(1-p)^{N_1+N_2-(x+y)} \quad (3)$$

Let us consider now the conditional probability $P(x, y\,;\,N_1, N_2 \mid x + y\,;\,N_1 + N_2)$ that, knowing the overall result of two random samplings of $x + y$ tags among $N_1 + N_2$, then x copies of a tag had actually been picked among N_1 and y among N_2 . According to Bayes' theorem, for two events A and B, the probability for A occuring knowing B is :

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)} \quad (4)$$

Using (4) in the present case, we obtain:

$$P(x, y\,;\,N_1,\,N_2 \mid x\,+\,y\,;\,N_1\,+\,N_2) = \frac{P(x\,+\,y\,;\,N_1\,+\,N_2 \mid x, y;\,N_1, N_2)\,P(x, y;\,N_1, N_2)}{P(x\,+\,y\,;\,N_1\,+\,N_2)} \quad (5)$$

Since the fact that x copies of a tag have been found in N_1, and y in N_2 implies that $x + y$ tags occured in $N_1 + N_2$, it follows that : $P(x + y\,;\,N_1 + N_2 \mid x, y\,;\,N_1, N_2) = 1$. Now, considering the two independent binomial distributions (1) et (2), we can write $P(x, y\,;\,N_1, N_2) = P_1(x\,;\,N_1)\,P_2(y\,;\,N_2)$.

Equation (5) then becomes :

$$P(x, y;\,N_1,\,N_2 \mid x\,+\,y\,;\,N_1\,+\,N_2) = \frac{P_1(x\,;\,N_1)\,P_2(y\,;\,N_2)}{P_{1,2}(x\,+\,y\,;\,N_1\,+\,N_2)} \quad (6)$$

Combining (1), (2) et (3) in (6), we obtain:

$$P(x, y \mid x+y) = \frac{C_{N_1}^{x}\,C_{N_2}^{y}}{C_{N_1+N_2}^{x+y}} \quad (7)$$

The symetry of equation (7) allows to compare independent experiments, with tag libraries of different sizes. A low P value will allow to consider the variation as being biologically relevant.