FACULDADE DE CIÊNCIAS E TECNOLOGIA
Departamento de Ciências Biológicas e Biotecnologia

Identification of reliable reference genes for quantitative gene expression in the Mozambique tilapia, *Oreochromis mossambicus*.

Master's Degree in Marine Biology

(especialização em aquacultura)

André Alçada Baptista Rebelo de Andrade

Supervisor: Professor Doctor Adelino Canário

Co-supervisor: Doctor Laurence Deloffre

Faro, September 2010

Acknowledgments

Gostaria de agradecer ao Professor Adelino Canário pela oportunidade que me concedeu de realizar este projecto sob a sua supervisão e pelas correcções.

Queria também agradecer especialmente à Laurence por todos os preciosos ensinamentos, pela paciência nos meus piores momentos, por adaptar os horários do laboratório ao trabalho no cinema e pelas conversas divertidas.

À Eduarda pelas correcções preciosas e a força transmitida.

À Alexandra por todos os qRT-PCRs e extracções que fizemos juntos, acho que formamos uma boa equipa.

À Rita e Ângela pela ajuda na histologia, e simpatia.

Ao Pedro e Nicolai pelos almoços, cafés e óptimas conversas

À Rute, Ricardo, Cristina e Marco, enfim a todo o pessoal que frequentava o 2.28 e que adorei conhecer.

Ao Peter pela disponibilidade do laboratório quando foi preciso dissecar os ovos naqueles infindáveis dias.

Especialmente à Débora, pela força, companheirismo e paciência quando eu não arranjava tempo para nada.

Aos meus pais e irmãos que adoro, e aos meus amigos.

I. Abstract

Gene-expression analysis is increasingly important in biological research, with real-time reverse transcription polymerase chain reaction (qRT-PCR) becoming the method of choice for accurate expression profiling of selected genes. One or more appropriate reference genes must be selected to accurately compare mRNA across different samples and tissues. Although reference genes constitute the best normalizers, a major problem is that there is substantial evidence in the literature that most of the commonly used reference genes are regulated under some circumstances, such as at different stages of development or in response to different treatments, indicating that there may be no universal reference gene with a constant expression.

In this study we evaluated the expression of ten candidate reference genes (*18S* ribosomal RNA, *β-actin*, HPRT-1, *GADPH*, Tubulin A, TATA binding protein, Elongation factor 1 alpha, Beta-2 microglobulin, Cathepsin D, Cathepsin Z) by qRT-PCR in 13 distinct adult tissues and during five stages of oogenesis in *Oreochromis mossambicus*. Two different software tools (Genorm and Normfinder) were used to analyze the data given by qRT-PCR in order to validate the stability of the reference genes and to determinate of the most suitable one(s). The analysis revealed that the more stable genes in somatic tissues were *18S* ribosomal RNA, Elongation factor 1 alpha, Beta-2 microglobulin and TATA binding protein while during oogenesis it were *β-actin*, Elongation factor 1 alpha and Cathepsin D. It was also observed that *GADPH* clearly had large variability in its expression in both cases tested in our study and is not recommended for normalization.

In order to test and to compare the sets of reference genes chosen by the two programs they were used in the normalization of the expression of two bone morphogenetic protein antagonists, *BAMBI* and *Gremlin* in the five groups of oocytes at different stages of development. The test of the reference genes revealed that the software geNorm had more satisfactory results then Normfinder software and that it should be used for best normalization the set of four reference genes *CTSD, B-actin, EFa1* and *CTSZ*.

II. Resumo

A importância na análise de expressão genética tem vindo a crescer significativamente nas últimas década, sendo o real-time reverse transcription PCR (qRT-PCR) o método por excelência devido à sua grande eficiência e precisão. Um dos requisitos para medir a expressão genética através do qRT-PCR é utilização de um ou mais genes de referência, que funcionam como controlos internos para eliminar factores de variação inerentes a este método. Embora os genes de referência sejam a melhor ferramenta para a normalização, um dos grandes constrangimentos que se tem vindo a verificar é que cada vez mais estudos são reportados em que demonstram que muitos dos genes de referência até agora utilizados, são regulados em algumas circunstâncias tais como em diferentes estados de desenvolvimento ou em resposta a diferentes tratamentos. Indicando assim que não existe um gene de referência universal que tenha uma expressão constante em todas as células e diferentes situações.

Neste estudo avaliámos a expressão genética de 10 possiveis genes de referência (*18S* RNA ribossomal, *β-actin*a, HPRT-1, *GADPH*, Tubulina A, proteína de ligação TATA, factor de alongamento alpha 1, Beta-2 microglobulina, Catepsina D e Catepsina Z) através de qRT-PCR, em 13 tecidos de indivíduos adultos da espécie *Oreochromis mossambicus* e em cinco grupos de ovocitos em desenvolvimento de fêmeas desta espécie. A selecção de genes de referência é considerado um problema circular, já que mesmo estes genes precisam de ser normalizados. Diversas ferramentas informáticas baseadas em algoritmos têm sido desenvolvidas para contornar este problema. Duas destas ferramentas são os programas geNorm e Normfinder, os quais foram utilizados neste trabalho para analisar os dados fornecidos pelo qRT-PCR com a finalidade de validar a estabilidade destes genes e para permitir a determinação do(s) mais adequado(s). A análise efectuada por estes programas revelou que os genes mais estáveis ao longo dos tecidos foram os genes *18S* RNA ribossomal, factor de alongamento alpha 1, Beta-2 microglobulina e o proteína de ligação TATA, e que nos cinco grupos de ovocitos em desenvolvimento foram os genes *β-actin*a, factor de alongamento alpha 1 e Catepsina D. Outra conclusão obtida neste estudo foi o facto de o gene *GADPH*, muito usualmente usado como gene de

referência revelar-se pouco recomendado para a normalização no nosso caso, já que foi o gene que revelou uma maior variabilidade genética quer nos tecidos como nos ovocitos.

Para testar e comparar o(s) gene(s) de referência apontados por estes dois programas, estes foram utilizados para a normalização da expressão genética de duas proteínas morfogenéticas ósseas antagonistas, *BAMBI* e *Gremlin*, nos cinco grupos de ovocitos em desenvolvimento. O teste revelou que o programa geNorm apresentou resultados mais satisfatórios que o programa Normfinder e que se deverá utilizar para a melhor normalização o conjunto de quatro genes *CTSD, B-actin, EFa1* e *CTSZ* .

III. Abbreviations

18S* -18S* Ribosomal RNA
***B2m* -** Beta-2 microglobulin
B-act - Beta-actin
***BAMBI* -** Bone morphogenetic protein and activin membrane-bound inhibitor
BMP - Bone morphogenetic protein
Bp - base pairs
CAS - Cortical alveoli stage
cDNA - Complementary DNA
Cq – Quantification cycle
CNS - Chromatin nucleolar stage
CSTD - Cathepsin D
CSTZ - Cathepsin Z
Ct – Threshold cycle
DEPC - Diethylpyrocarbonate
DNA - Deoxyribonucleic Acid
dsDNA – Double stranded DNA
dNTPs - Deoxynucleotides triphosphates
***EF1a* -** Elongation factor- 1 alpha
FSH - Follicle stimulating hormone
***GADPH* -** Glyceraldehyde 3-phosphate dehydrogenase
GSI – Gonodosomatic Index
GVBD - Germinal vesicle breakdown
Gre – *Gremlin*
***HPRT1* -** Hypoxanthine phosphoribosyltransferase 1
MIS - Maturation-inducing steroids
mRNA - Messenger RNA
MVB - Multivesicular bodies
MS - Maturation stage
PCR - Polymerase chain reaction
PGC - Primordial germ cells
PS - Perinuclear stage
qRT-PCR – Real Time Quantitative PCR
RACE-PCR - Rapid amplification of cDNA ends PCR
rDNAse - RNAse free DNAse
RNA - Ribonucleic acid
RT-PCR - Reverse transcriptase PCR
SD - Standard deviation
ssDNA – Single stranded DNA
PGCs - Primordial germ-cells
***TBP* -** TATA Binding Protein
***Tub A* -** Tubulin, alpha 1
Ta – Annealing temperature
Tm – Melting temperature
UTR - Untranslated region
VS - Vitellogenic stage
Vtgs - Vitellogenins

IV. Table of contents

I. ABSTRACT III

II. RESUMO IV

III. ABBREVIATIONS VI

IV. TABLE OF CONTENTS VII

1. INTRODUCTION 1

1.1 REAL-TIME QUANTITATIVE PCR (QRT-PCR) 4
1.1.1 FLUORESCENT MARKERS 6
1.1.2 ONE-STEP VS TWO-STEP REACTION 7
1.1.3 QUANTITATION OF RESULTS 8
1.2 REPRODUCTION AND FOLLICLE DEVELOPMENT IN TELEOSTS 9
1.2.1 SEX DETERMINATION AND REPRODUCTION STRATEGIES 10
1.2.2 OVARY 11
1.2.3 FOLLICLE DEVELOPMENT 11
1.3 TILAPIA AS A MODEL 15
1.3.1 THE TILAPIAS 15
1.4 OBJECTIVES 17

2. MATERIAL AND METHODS 18

2.1 CLONING OF PUTATIVE REFERENCE GENES 18
2.1.1 CDNA SYNTHESIS 18
2.1.2 POLYMERASE CHAIN REACTION 19
2.1.3 CLONING 20
2.2 CLONING OF TWO ANTAGONISTS OF BMPS, *GREMLIN* AND *BAMBI* 22
2.2.1 PARTIAL CDNA CLONING OF GREMLIN 22
2.2.2 FULL LENGTH CDNA CLONING OF BAMBI 23
2.2.3 MULTISEQUENCE ALIGNMENT AND IDENTIFICATION OF CONSERVED DOMAINS 25
2.3 TISSUE COLLECTION 25
2.4 DISSECTION OF OOCYTES AT DIFFERENT STAGES 26
2.5 HISTOLOGY 27
2.6 QUANTITATIVE-POLYMERASE CHAIN REACTION 27
2.6.1 PRIMER DESIGN 27
2.6.2 QUANTITATIVE RT-PCR ON ADULT TISSUES 28
2.6.3 QUANTITATIVE RT-PCR OF OOCYTES 31

3. RESULTS 39

3.1 HISTOLOGY 39
3.2 QUANTITATIVE RT-PCR 41
TABLE 8 – ACCESSION NUMBER OF ALL GENES CLONED IN THIS STUDY. 41
3.2.1 QUANTITATIVE PCR ON ADULT TISSUES 42
3.2.2 QUANTITATIVE RT-PCR ON OOCYTES DURING DEVELOPMENT 48
3.2.3 BAMBI AND GREMLIN EXPRESSION IN OOCYTES 55

4. DISCUSSION 60

5. REFERENCES 68

ANNEX 1: DNA LADDER A

ANNEX 2: P-GEM-T EASY VECTOR RESTRICTION MAP B

ANNEX 3: CALCIUM CHLORIDE METHOD C

ANNEX 4: SOLUTIONS D

ANNEX 5 – MULTISEQUENCE ALIGNMENT FOR ALL PUTATIVE REFERENCE GENES G

ANNEX 6 – MULTISEQUENCE ALIGNMENT, MATRIX OF IDENTITIES AND PHILOGENETIC TREES FOR *BAMBI* AND *GREMLIN* GENES P

ANNEX 7: STANDARD AND MELTING CURVES FOR ALL GENES V

1. INTRODUCTION

Understanding the patterns of gene expression is expected to provide insight into complex regulatory networks and will most probably lead to the identification of the genes relevant to a biological process. In fact, the importance of gene-expression analysis is increasing in many fields of biological research. The real-time reverse transcription quantitative polymerase chain reaction (qRT-PCR) technique is one of the tools of choice to analyze gene expression by measuring transcript abundance and has gained much popularity in recent years [3].

This method is well suited for validation of microarray expression screening results and for studies of a selected number of candidate genes or pathway constituents in an experimental setup (biopsies, treated cell cultures or any other sample collection) [30]. qRT-PCR can now be performed on 384-well block thermal cyclers that allow gene expression analysis of even higher number of genes and samples (1 to 48 samples for 48 to 3072 different genes) [30].

Any measured variation in gene expression is caused by two sources. One is the true biological variation, and the other is non-specific variation which includes mainly factors relative to template input quantity and quality, pipetting errors, inhibitory compounds, yields of the extraction process and the enzymatic reactions (reverse transcription and polymerase chain reaction amplification). One of the major difficulties in obtaining reliable expression patterns is the removal of this experimentally induced non-biological variation from the true biological variation. This can be done through normalization by controlling as many of the variables as possible [30].

There are several strategies to remove experimentally induced variation but most of these methods cannot completely reduce all sources of variation. That is why it is crucial to standardize each step of the work, if not, variation will be introduced in the results that cannot be eliminated by applying the final normalization. Indeed it has been shown to be very important to try to control all sources of variation along the entire workflow of PCR based gene expression analysis [31].

It is generally agreed that the use of reference genes, also called housekeeping genes, is the preferred way of normalizing qRT-PCR data. This method is particularly attractive because the reference genes are internal controls that are affected by all sources of variation during the experimental workflow in the same way as the genes of interest [30]. The reference genes are expressed in the cells, and their mRNAs are present during nucleic acid extraction, storage, and any enzymatic processes such as DNase treatment and reverse transcription [17].

The ideal reference gene is a gene which expression occurs at a constant level in all tissues, independently of experimental conditions and developmental stages. These genes are generally well conserved throughout evolution and play key roles in cell survival, and their synthesis occurs in all nucleated cells types [7] [8].

Although reference genes constitute the best normalizers, a major problem is that there is substantial evidence in the literature that most of the commonly used reference genes are regulated under some circumstances, such as at different stages of development or in response to different treatments, indicating that there may be no universal reference gene with a constant expression [1] [24].

Genes frequently used, and considered as classical reference genes, are highly expressed genes such as glyceraldehyde-3-phosphate dehydrogenase, Beta-actin and ribosomal protein *18S* that due to their key role in metabolism, cytoskeleton and ribosome structures, respectively, it was generally assumed that their expression levels were constant [1]. However it has been reported that these so called classical reference genes are also susceptible to regulation in certain circumstances [40]. Because of this, many other genes have been investigated as potential reference genes, including TATA binding protein, beta-2-microglobulin, tubulin alpha, L3 ribosomal protein and hypoxanthine phosphoribosyl transferase 1, but they all present the same problems of expression variation as the classical reference genes and are therefore of limited value as universal quantitative references [6]. These variations can partially be explained by the fact that some of the proteins encoded by the reference genes can participate in functions other than basal cell metabolism [7].

Therefore it is crucial to validate for each experimental situation if a candidate reference gene is suitable as reference gene for normalization. If unrecognized, unexpected changes in reference gene expression can result in erroneous conclusions about real biological effects [30].

In order to identify the most suitable reference gene a set of potential reference genes should be analyzed and validated for each different condition in a specific experimental study [8]. Typical experiments are performed by measuring the relative gene expression of 10 candidate genes in 10 representative samples from each tissue or sample group. The relative gene expression values are then imported to appropriate software and ranked according to their expression stability, choosing among the set the ones who have minimal variability. It has been reported that using a single non-validated reference gene may result in a significant bias and typically between 2 and 5 reference genes are required for accurate normalization [8].

Selection and validation of a reference gene could be a circular problem because the expression of this target gene itself also needs to be standardized. In order to try to circumvent this paradox and to analyze expression stability, several algorithms and software tools have been developed. Genorm [3] or Normfinder [47] are examples of such software that can be used to determine the most adequate reference genes to select.

This thesis is included as part of a more extensive science project. The objective of this work is to select suitable reference genes for qRT-PCR to be used in normalization of gene expression in *Oreochromis mossambicus*. Because there is evidence that classical reference genes may have expression variations at different stages of development, it is also important to evaluate the reference genes suitability in ovary follicules at different stages. During ovogenesis great structural and functional changes happened, regulated by gene expression. It is expected that the most suitable reference genes for follicles will be different from those of somatic tissues. If successful in recognizing the most suitable reference genes, this work will be a useful tool to apply in normalization of gene expression in studies with this species.

1.1 Real-Time quantitative PCR (qRT-PCR)

The qRT-PCR has become one of the most widely used methods for gene expression analysis due to its large dynamic range, sensitivity, as well as accurate quantification of specific gene expression [2].

The process of qRT-PCR is similar to traditional PCR except that the quantification of PCR product is measured after each cycle of amplification (along the reaction) instead of an end-point quantification which is not necessarily proportional to the amount of target present in the samples [26] [24]. PCR is an *in vitro* process that increases the amount of a specific DNA template starting with a small amount of the template in a three-step cycling process.

For the PCR amplification to occur the essential components are 2 synthetic oligonucleotide primers, complementary to the regions that flank the target DNA sequence on the opposite strands and that presents the 3' hydroxyl end oriented toward each other, a target sequence in the DNA sample, a thermostable DNA polymerase that can withstand multiple cycles of heating and dNTPs (19).

The PCR cycling process begins with the thermal denaturation of the DNA sample by raising the temperature to approximately 95ºC. In the second step, renaturation or annealing, the primers present in the reaction, in a vast molar excess, ligate to their complementary DNA sequences as the temperature is slowly reduced to near 55ºC (depending on the primer). The step of synthesis or extension is the third and final step and the temperature is raised up to 72ºC, the optimal temperature for the catalytic function of the *Taq* DNA polymerase for the extension of the sequences initiated in the primers, thus making copies of the target sequence [19].

For quantification to occur in qRT-PCR a fluorescent marker is commonly used. This marker or reporter produces a fluorescent signal measured by a camera or detector and during the reaction the reporter generates a signal that reflects the amount of product formed. In the initial cycles the signal is weak (small amount of product) and during the remaining of the reaction, as the amount of product accumulates, the signal increases exponentially, until it reaches saturation [24].

The fluorescent signal readings during a qRT-PCR can be analyzed as follows (**Figure 1**): linear ground phase, the beginning of the PCR reaction, in which the fluorescence emission at each cycle has not yet risen above the background; the early exponential phase, where the fluorescence emission will be statistically significantly higher than the background or baseline levels and is in this phase that the fluorescence of each reaction crosses the threshold level. The cycle in which it occurs is the threshold cycle (CT) or quantitative cycle (Cq), and is a parameter that characterizes each individual reaction. The Cq value is inversely correlated with the logarithm of initial copy number of template and is used for the calculation of the experimental results, log-linear phase follows the early exponential phase and corresponds to the phase when the PCR reaction reaches its optimal amplification and is expected that the amplification curves for different reactions to be parallel in this phase. The final PCR phase is the plateau stage, and is when the amplification rate drops to near zero and little more product is synthesized as the limiting reagent of the reaction is completely used [2].

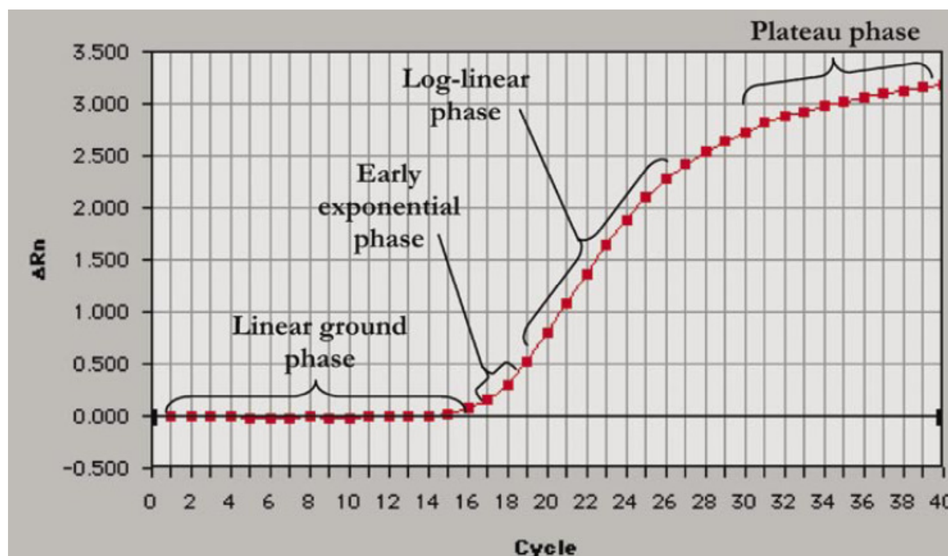


Figure 1 - Phases of the PCR amplification curve. PCR amplification curve charts the accumulation of fluorescent emission at each reaction cycle. The curve can be divided in four different phases: the linear ground, early exponential, log-linear, and plateau phases. Data gathered from these phases are important for calculating background signal, cycle threshold (Ct), and amplification efficiency. Rn is the intensity of fluorescent emission of the reporter dye divided by the intensity of fluorescent emission of the passive dye (a reference dye incorporated into the PCR master mix to control for differences in master mix volume). ΔRn is calculated as the difference in Rn values of a sample and either no template control or background, and thus represents the magnitude of signal generated during PCR. This graph was generated with ABI Prism SDS version 1.9 software (Applied Biosystems) (in Wong et al, 2005 [2]).

1.1.1 *Fluorescent markers*

Currently, several detection chemistries are used in real-time PCR but the most common are TaqMan® probes and the SYBR Green dye. When TaqMan® probes are used specific primers and probes for the amplicon are added to the reaction. The probes - synthetic oligonucleotides with a fluorescent reporter dye attached to the 5' end - are designed to hybridize to the cDNA and to be incorporated to the internal region of the PCR product. When the polymerase replicates the template, the probe is incorporated and the fluoropher and quencher are cleaved, thus resulting in the increase of the emitted fluorescence (**Figure 2**). The use of TaqMan® probes is very specific to the target amplicon, requires less optimizations in qRT-PCR method compared to other markers used, although it can be considerably expensive [4]. On the other hand, SYBR Green dye is not as specific as TaqMan probes as it binds to any double stranded DNA present in the reaction and upon excitation emits light (**Figure 2**). As this dye does not bind in a sequence-specific manner, the assay is prone to false positives. A strategy that can be used to discard the presence of false positives is the dissociation curve analysis, where the presence of different PCR products is reflected on the melting peaks [2] Melting peaks or melt curves allow a comparison of the melting temperatures of amplification products and, to be produced, the final PCR products are exposed to a temperature gradient from about 50 °C up to 95 °C while the fluorescence signal is continuously measured. Products of different lengths and sequences will melt at different temperatures and are observed as distinct peaks [23].

The use of SYBR Green dye requires more optimization, nevertheless is the simplest, flexible and most economical format used for detecting and quantifying PCR products in real-time PCR reactions [4].

Nonetheless, a good reporter should have low background fluorescence, emit a high signal upon amplicon formation and high target specificity [24].

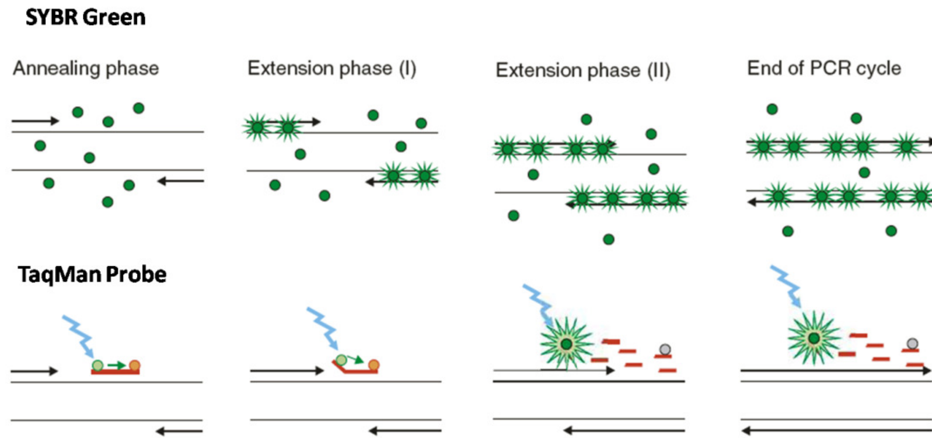


Figure 2 – Schematic representation of the functioning of the SYBR Green Dye and TaqMan probe. SYBR Green binds to the dsDNA as it is being synthesized, and starts to emit fluorescence. The TaqMan probe binds to the DNA with the primers and as the fluorophore (green) is excited, it transfers the energy to a quencher molecule (orange) present in the probe; as the extension occurs, the probe is broken down, the fluorophe is released emitting fluorescence. (Taken from Velden et al 2003 [18]).

1.1.2 One-step VS two-step reaction

For the qRT-PCR to be applied in gene expression analysis, the messenger RNA (mRNA) must be copied into complementary DNA (cDNA) by reverse transcription. This is a crucial step as it has to reflect in the cDNA the amounts of the target present in the mRNA [24]. qRT-PCR can be performed in one-step or two-step reactions.

One-step reaction is characterized by cDNA synthesis and PCR amplification in a single tube, while in the two-step reaction both reactions are done separately. The one-step reaction is thought to minimize experimental variation but it may not be suitable in situations where the sample is assayed for several times and is less sensitive than the two-step reaction. On the other hand the two-step reaction, by initially doing the reverse transcription, allows the use of the same amount of cDNA in several different assays by using dilutions of the original cDNA template [2].

1.1.3 Quantitation of results

Depending on the level of quantitation needed to the experiment, several approaches can be adopted in order to quantify the amount of template. Among the different methods used for quantification are included the absolute standard method, the relative standard method and the comparative threshold (C_t) method.

1.1.3.1 Absolute standard method

In this method the input copy numbers of the transcript of interest are determined usually by relating the PCR signal to a standard curve [27]. This curve is constructed from a serial dilution of cRNA, ssDNA or dsDNA of known concentrations [2] and is used as a reference standard for extrapolating quantitative information for mRNA targets of unknown concentrations [27]. The external standard used for the curve construction should have the same primer binding sites of the target sequence in order to eliminate potential differences in quantification due to annealing. Ideally, it should contain sequences that are the same as those in the target sequence or which vary slightly [2].

A DNA standard must be a pure species and can be synthesized by cloning the target sequence into a plasmid, purifying a conventional PCR product, or directly synthesizing the target nucleic acid [2]. Spectrophotometric measurements at 260nm are normally used to determine the concentrations of these DNAs.

For correct application of this method the amplification efficiencies of the standard and the target gene should be the same. Normalization to a housekeeping gene or reference gene should be performed in order to control variations in the efficiency of the reverse transcription step and also to correct the variation introduced to variable RNA inputs [4].

1.1.3.2 Relative standard method

In the relative standard method, the quantity of each experimental sample is first determined using a standard curve and then expressed relative to a single

calibrator sample [2]. The target quantity is determined from the standard curve and divided by the target quantity of the calibrator to generate relative expression levels. The calibrator is designated as unity, and all the other quantities are expressed as an n-fold difference relative to the calibrator [27].

This method is often used also when the amplification efficiencies of the reference and target genes are not equal and results should also be normalized by endogenous controls, usually housekeeping genes.

1.1.3.3 Comparative threshold (C_t) method

The comparative threshold method is a mathematical model that calculates changes in gene expression as a relative fold difference by comparing C_t values between experimental samples and a control or calibrator such as a non-treated sample or RNA from normal tissue [2] [4]. The C_t values of both the calibrator and the samples of interest are normalized to an appropriate endogenous housekeeping or reference gene [4].

The comparative Ct method is also known as the $2^{-[delta][delta]Ct}$ method, where $[delta][delta]C_t = [delta]C_{t,sample} - [delta]C_{t,reference}$ and $[delta]C_{t,sample}$ is the Ct value for any sample normalized to the endogenous housekeeping gene and $[delta]C_{t,reference}$ is the C_t value for the calibrator also normalized to the endogenous housekeeping gene [4].

For the $[delta][delta]C_t$ calculation to be valid, the amplification efficiencies of the target and the endogenous reference must be approximately the same. This can be established by looking at how $[delta]C_t$ varies with template dilution. If the plot of cDNA dilution versus delta C_t is close to zero, it implies that the efficiencies of the target and housekeeping genes are very similar [4].

1.2 Reproduction and follicle development in Teleosts

The infraclass Teleostei, commonly referred to as teleosts or bony fishes, embraces a large number of species distributed in a great variety of aquatic environments, and is the most representative group among vertebrates [9]. This diversity leads to a wide range of reproductive adaptations, or "reproductive

strategies", of interest in comparative biology. Also the numerous biological models that comprise this group can be particularly convenient for the acquisition of physiological or genetic knowledge [12], as the knowledge in the molecular pathways involved in reproduction processes in vertebrates, even in mammalian models is still limited [9].

Among several different aspects involving teleost reproduction, ovarian follicle development is a crucial step for their success on larvae recruitment. Studies on this development process can provide useful information not only for basic research, but also for aquaculture, fisheries management, as well as environmental and biomedical sciences. Increasing the knowledge in molecular pathways involved in this process may turn very useful to improve the industry of aquaculture since one of their biggest concerns is to produce "high-quality eggs".

1.2.1 Sex determination and reproduction strategies

Teleosts have a great plasticity in respect of sex determination with a range of gonochoristic species - where individuals are either male (M) or female (F) - or species in which gender change may occur during their lifetime (protogynous F→M; protandrous M→F) [11].

Reproductive behavior is also much diversified among teleosts depending on the species, from spawning without egg care, until nest building and complex parental care [12].

There are three main types of ovarian development among teleosts, the synchronous, where all oocytes develop and ovulate at the same time, the "group-synchronous", where at least two different stages of oocyte development are present in the ovary, and asynchronous, where all stages of oocytes are present without a dominant population. Group synchronous fishes spawn in a single episode or in a short period of time, while the asynchronous have several spawning batches during the spawning season. The reproductive effort can be indicated by the gonadosomic index (GSI) - the ratio between the gonad weight and total weight of the fish [13].

1.2.2 Ovary

The ovary of teleosts shows a similar general structure. In almost all the cases they usually appear as paired elongated organs oriented longitudinally within the abdominal cavity, under the kidney (and, if present, the swim bladder). An oviduct is situated at the posterior part of each ovary and is connected to the genital papilla [12].

Ovaries are compartmentalized by numerous septa formed by folds of the germinal epithelium projected into the ovarian lumen. These septa are called ovigerous lamella and contain nests of oogonia and oocytes at early stages, and also follicules at various stages of oocyte maturation. In adult teleost females the oogonia keep on proliferating contrarily to mammals, renewing the stocks of young oocytes and follicles. At ovulation, mature oocytes are released from their follicle into the ovarian cavity, before being laid through the oviduct and the genital papilla [12].

1.2.3 Follicle development

The formation of a mature egg in all teleosts undergo the same basic pattern (**Figure 3**): formation of primordial germ-cells (PGCs); the transformation of PGCs into oogonia; transition of oogonia into oocytes (onset of meiosis); followed by vitellogenesis (meiotic arrest at the end of prophase I); maturation with resumption of meiosis and germinal vesicle breakdown (GVBD); finishing with ovulation where mature eggs are expelled from its follicule [11].

1.2.3.1 Primordial germ cells and primary growth

The first step of oogenesis is the formation of PGCs that are produced from germ line cells developed by association with the accumulation of maternal RNA [11]. PGCs have a distinct nucleus membrane and one or two prominent nucleoli, they are of relatively large size and have large nuclei.

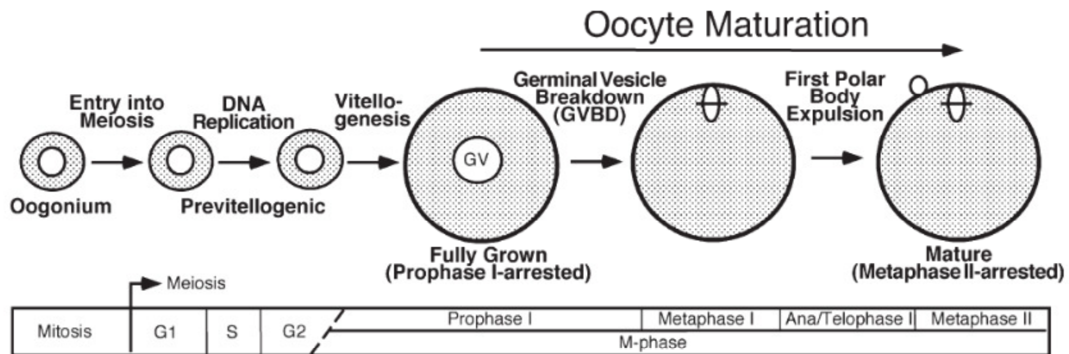


Figure 3 - Oocyte growth, oocyte maturation, and meiosis. Oocytes produced by the entry of mitotically proliferating oogonia into meiosis stop their meiotic cell cycle at prophase I, during which they grow by the accumulation of yolk (vitellogenesis). The prophase I-arrested oocytes are immature. Upon hormonal stimulation, the immature oocytes resume meiosis and proceed to metaphase II, at which stage they mature. (taken from [25]).

The PGCs are transformed into oogonia through structural changes and then oogonia multiply by mitotic divisions forming oogonial nests.

At this stage each oogonium becomes surrounded by a monolayer of somatic granulosa cells that secrete a basement lamina. Outside the basement lamina it is formed the theca which is a monolayer of somatic cells associated with blood vessels [11]. The oocyte with its surrounding granulosa cells, basement lamina and theca somatic layer constitutes the ovarian follicle and forms the primary oocyte. The transition from oogonium to a primary oocyte is also characterized by the initiation of the first meiotic division, before leaving the oogonial nest [11]. The primary oocyte growth encompasses the meiotic chromatin-nucleolus stage until early cortical stage and is involved with the development of the follicle layers surrounding the oocyte (**Figure 4**). It is also known that at this stage the organelles and molecules required for later stages are synthesized [11].

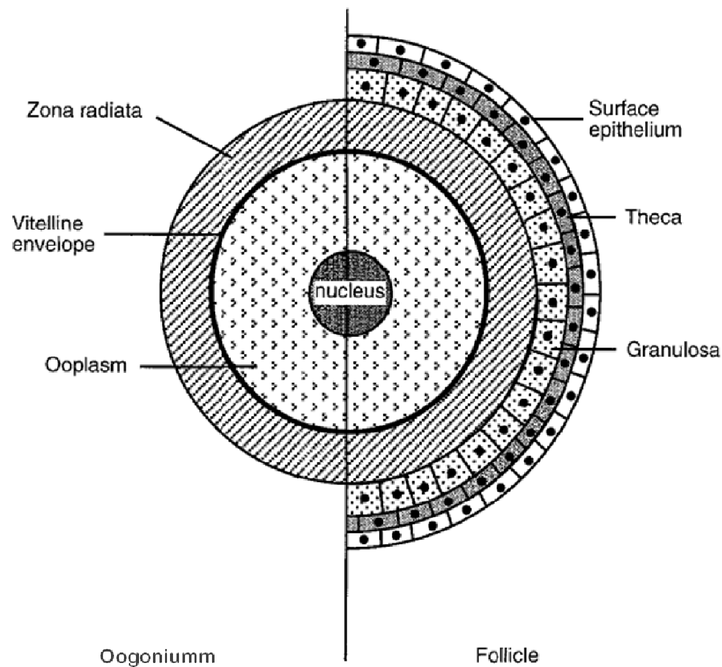


Figure 4 - Diagrammatic representation of the distinction between an oogonium and an intact ovarian follicle oocyte (adapted from [12]).

Oocyte enlargement occurs associated to the appearance of cortical alveoli at the periphery of oocytes that characterizes the so called "primary vittelogenesis" stage. At this stage the cortical alveoli multiplies, increasing the oocyte size by filling the cytoplasm with vesicles which differ in composition among species. At the end of this stage the content of cortical alveoli that are displaced to the periphery of the oocyte, is released to the egg surface leading to the formation of the chorion by restructuring of the egg envelop proteins [11].

1.2.3.2 Vitellogenesis

Vittelogenesis, in almost all teleosts, is the stage in which the oocyte shows a massive growth and may account as much as 95% of the final egg size and GSI increasing between 50 and 100 fold [13]. This growth is due to the incorporation by the oocyte of mainly of vitellogenins (Vtgs), but also vitamins and carbohydrates, all carried from the blood stream. These components will be crucial for the future development of the embryo. Vtg is a glycophospholipoprotein composed of 79% protein and 19% lipid. It is synthesized by the liver in response to estrogenic stimulation and its cleavage into yolk proteins is probably caused by the proteolytic activity of the enzyme

cathepsin D. Vtg enters in the oocyte passing through thecal capillaries until the granulosa layer and then through pore canals of the zona radiata into the oocyte membrane. At the membrane it binds to specific receptors in the endocytotic clathrin-coated pits of the vesicles. These vesicles then move into the peripheral ooplasm and fuse with lysosomes forming multivesicular bodies (MVB) [20].

1.2.3.3 Final Maturation and Ovulation

Oocyte final maturation processes are characterized by the reduction or the end of endocytosis, resumption of meiosis, germinal vesicle breakdown (GVBD), the formation of a monolayer of cortical alveoli under the oolemma and yolk platelet dissolution. The first meiotic division gives rise to two cells differing in size, the small cell with first polar body degenerates and the large mature oocyte is formed.

Numerous studies have shown that the transition of vitellogenesis into maturation is associated with an increase of plasma LH levels and increased expression of the LH receptor, leading to the production of maturation-inducing steroids (MIS) regulated by a LH driven switch [22]. The MIS is a derivative of progesterone and binds to oocyte-specific receptors to activate the maturation promoting factor in the ooplasm, which leads to the dissolution of the germinal vesicle and reinitiates meiosis. The stage of final maturation is also characterized by yolk clarification, oil droplet fusion and an increase of the oocyte volume. This increase is due to the intake of water and it happens mainly on marine species as the embryos need to have a water reservoir, in order to survive in hyper-osmotic seawater and for flotation [11].

Ovulation is the process of release of the mature oocytes from its follicle into the ovarian or abdominal cavity. The mature oocytes are released by separation from the granulosa layer and follicle layer. For some species, the stimulation by MIS leads not only to maturation but also to ovulation, in other species the ovulation is stimulated by arachidonic acid and its metabolites including prostaglandins [11].

1.3 Tilapia as a model

The complexity of tilapias, along with their economic interest worldwide makes them an important object of study. The relative ease of handling and rearing make them also a good study model for researchers. Recent research on tilapia has been mainly to define the mechanisms of the female gonad and development organization [28]. Among tilapia species, the Nile tilapia (*Oreocheromis niloticus*) has been used in the majority of studies, although, Mozambique tilapia (*Oreocheromis mossambicus*) (**Figure 5**) shares several common features that also make this species a suitable model for research in various fields.

Figure 5 - *Oreochromis mossambicus* (Peters, 1852). http://www.fao.org/fishery/species/2408: accessed 8 March 2011.

1.3.1 The Tilapias

Tilapias is the common name given to a group of subtropical to tropical freshwater fish of the family *Cichlidae* (Order *Perciformes*) that are native to Africa and the south-western Middle East [14]. Tilapias are grouped into three genera, according to parental care patterns: *Oreochromis* (arena-spawning maternal mouthbrooders); *Sarotherodon* (paternal or biparental mouthbrooders); *Tilapia* (substrate spawners) [14] [15].

Many species of Tilapia have been dispersed almost worldwide primarily for use in biological control, as bait fish, ornamental fish and for food supply by

farming in aquaculture systems [14]. Tilapias are now the second group of fish most farmed worldwide (behind carps) with an annual production of 2.5 million tons [29]. The species more representative in this farming industry are Mozambique tilapia (*Oreochromis mossambicus*), Nile tilapia (*O. niloticus*) and Blue tilapia (*O. aureus*). These species are well-suited for aquaculture production mainly because they are fast growers and tolerant of a wide range of environmental conditions, such as salinity and oxygen. Also the fact that they are omnivorous and, in some cases, disease resistant and overcrowding tolerant, makes them good targets to use in aquaculture [14]. Farming is based on all male monosex populations due to higher growth rates of males compared to females. This is explained because females of these species become sexually mature very early , which, associated with time spending in elaborated parental care, requires a large energy effort that reduces their growth rates. The monosex populations are normally achieved by androgen treatments and despite the success of its use, there is a preference for genetic control which is more suited to avoid environmental contaminations by hormones [15].

All the previously described characteristics made this group a case of success not only in aquaculture farming but also a successful invasive species since they are probably the most widely distributed group of exotic fish in the world [14], which is another strong argument to better understand this species in the attempt to minimize their destructive impact on the invasion of new habitats.

1.3.1.1 Reproduction

Tilapia species are sexually mature when they are 8-10cm and 2-3 months old. Adult fish can live until 6 to 8 years and in some cases even 12 years. The spawning season of tilapia usually starts in the spring months in temperate regions when the temperature rises, and continues throughout the year as long as water temperatures are above 22 ºC [16].

Tilapias make nests in the form of shallow pits in the substrate where courting and spawning happens. In most cases, males are territorial and develop aggressive behavior [16]. Female release her eggs in the nest and, in most cases, after fertilization, parents collect them into their mouth to undergo incubation. After hatching parents can maintain their young in the mouth for a period of time

for protection [16]. Females have multiple spawns due to their asynchronous ovaries and can produce a large variation in the number of eggs, depending on the species, body size and age [16].

1.3.1.2 Determination of sex

Tilapias are heterogametic (XX/XY). “All male populations” can be achieved genetically through the development off YY “supermales” and it has been shown that sex determination on tilapia species is more complex than a simple XX/XY monofactorial system [15].

Although sex determination is strongly determined by chromosomes, it is now clear that other factors such as the environment are also acting on sex. It has been demonstrated in various tilapia species that temperature can have a strong effect on sex determination. As an example, Baroiller et al., in 1995 masculinised XX progenies (100% females) with functional male phenotypes by raising water temperature above 32 ºC [15].

1.4 Objectives

The main objective of this thesis is the selection of suitable reference gene(s) for qRT -PCR in the Mozambique tilapia.

The study will start by cloning of ten potential reference genes (**Table 1**), followed by qRT-PCR and statistical analysis to determine their suitability.

The method of qRT-PCR will be used to measure the mRNA transcript level in different somatic tissues (brain, esophagus, stomach, anterior intestine, medium intestine, posterior intestine, liver, kidney, head kidney, spleen, heart, gill, and gonad) and during follicle development.

Gene expression levels will be compared among tissues according to sex and between oocytes maturation stages. Two different software tools (Genorm (version 3.5) http://medgen.ugent.be/~jvdesomp/genorm [3] and Normfinder http://www.mdl.dk/publicationsnormfinder.htm [47]) will be used to validate the stability of the reference genes and to allow the determination of the most suitable one(s).

After validation, the reference genes chosen will be tested in the analysis of the expression of two bone morphogenetic protein antagonists, *BAMBI* and *Gremlin*. The coding region of the two antagonists will be cloned and expression levels analyzed by qRT-PCR.

Table 1

Candidate reference genes which were evaluated.

Abbreviation	Gene name	Function
18S	*18S* Ribossomal RNA	Ribosome structure
β-act	Beta-actin	Cytoskeletal structural protein
GADPH	Glyceraldehyde 3-phosphate dehydrogenase	Glycolytic enzyme
HPRT1	Hypoxanthine phosphoribosyltransferase 1	Purine nucleotide synthesis
TBP	TATA Binding Protein	Transcription
B2m	Beta-2 microglobulin	Major histocompatibility complex
TubA	Tubulin, alpha 1	Cytoskeletal structural protein
EF1a	Elongation factor- 1 alpha	Protein synthesis
CSTD	Cathepsin D	Lipoprotein uptake and yolk processing
CSTZ	Cathepsin Z	Lipoprotein uptake and yolk processing

2. Material and methods

2.1 Cloning of putative reference genes

2.1.1 cDNA synthesis

Complementary DNA (cDNA) is a single stranded DNA molecule that is synthesized by the action of a reverse transcriptase enzyme from a template of single-stranded RNA. This molecule is an important tool for molecular biology, providing the information from the RNA, in a stable form.

Total RNA from *O. mossambicus* ovary, treated with DNase I, was available in the laboratory and was used for this task. Three micrograms of total RNA were denaturated at 65ºC for 10 min. Reverse transcription was carried out in a 40 µl

reaction volume, 40 U of MMLV reverse transcriptase (Promega), 8 U of Ribolock RNase inhibitor (Fermentas), 0.2 mM dNTPs and 10 pmol oligo(dT) adaptor (5'-CGAGTCGACATCGATCGT$(T)_{16}$-3'). A negative control without reverse transcriptase was used. Synthesis reactions were incubated for 1 hour at 37ºC.

2.1.2 Polymerase Chain Reaction

The polymerase chain reaction or PCR is a method for selective replication of a particular DNA fragment. It uses DNA polymerase that catalyzes the replication of a DNA template in a buffered reaction with an excess of a pair of short oligonucleotides primers and four deoxynucleoside triphosphates (dNTPs) to make millions of copies of the target sequence.

2.1.2.1 Primer design

To ensure primer efficiency different parameters were controlled with the software following the common established rules for primer design. The percentage of nucleotides G and C were between 40% and 60%, melting temperature in the range between 54 and 68ºC (2ºC for each A or T base and 4ºC for each G or C base) and differing no more than 5ºC between forward and reverse primers, size between 20 and 30 base pairs.

For each of the putative reference genes with the exception of the *18S* ribosomal RNA, for which the sequence is available in the non-redundant nucleotide database (nr) at the National Center for Biotechnology Information (NCBI), a mutisequence alignment was carried out using available vertebrate nucleotide sequences for that particular gene (http://multalin.toulouse.inra.fr/multalin/multalin.html). Primer pairs were designed in the conserved region of the sequences. Properties of the primers were controlled using the online software, Oligonucleotide Properties Calculator (http://www.basic.northwestern.edu/biotools/oligocalc.html). All primers were synthesized by Sigma-Aldrich and are presented on table 2 at the end of methods section.

2.1.2.2 PCR reactions

The optimization of PCR reactions conditions determined the optimal annealing temperature and $MgCl_2$ concentration for each pair of primers.

Frequently the Ta estimated by the software or given by the primer manufacturer may not produce the best results with the protocol used in the laboratory. To determinate the optimal Ta a gradient of temperature was tested.

Magnesium is an important DNA polymerase co-factor and excessive concentration used in the reaction will result in specificity loss. To optimize this parameter, different concentrations were tested in parallel to the temperature gradients.

PCR reactions (50 µl volume) were performed using ~ 20 ng of cDNA, appropriate concentration of $MgCl_2$, 0.1 mM of dNTPs, 50 pmol of each sense and antisense primers and 0.6 U of DreamTaq™ DNA polymerase (Fermentas). All reactions were carried out on a MyCycler thermocycler (Biorad), using the following conditions: initial denaturation at 94ºC for 5 min, 30 cycles at 94ºC for 45 sec, adapted annealing temperature (table 3) for 1min, 72ºC for 1 min and final extension at 72ºC for 10 min.

The PCR products were analyzed in a 1.5% agarose gel electrophoresis. PCR products with the expected size were purified on columns with the GeneJET™ Gel Extraction Kit (Fermentas, USA) according to the manufacturer's instructions. The purified PCR product was eluted on 30 µl of water, quantified by a nanodrop ND-1000 spectrophotometer.

2.1.3 Cloning

The pGEM®-T Easy (Promega, Spain) (Annexe 2) was used as cloning vector. It is a linearized plasmid with a single 3´-terminal thymidine at both ends which allows the direct cloning of PCR products since they have 3'-deoxyadenosine (A) residues at the extremities, as a result of the activity of DreamTaq™ DNA polymerase used in this work. This vector has multiple cloning sites within the α-peptide coding region of the enzyme β-galactosidase. This enzyme is responsible for X-Gal (5-bromo-4-chlor-3-indolyl-beta-D-

galactopyranoside) degradation; as a result it produces blue colored colonies of bacteria. When a DNA fragment is ligated into the vector, the α-peptide is inactivated resulting in white colonies. This test is called the blue/white screen.

2.1.3.1 Ligation

To ligate the purified and quantified PCR products into the pGEM®-T Easy vector, the mass ratio insert:vector of 3:1 was used. The reaction was done as following: ~ 50 ng of PCR product were added to 15 ng of pGEM®-T Easy (Promega, Spain), 1.5 weiss U of T4 ligase (Promega, Spain), 1 µl of 10x ligation buffer and sterile water up to 10 µl. The mixture of ligation reaction was incubated overnight at 4ºC and then used for the transformation of competent bacteria.

2.1.3.2 Transformation

For transformation, 5 µl of ligation reaction were added to 100 µl of competent XL-1 Blue MRF` cells previously prepared for competence by the calcium chloride method (Annex 3). After incubation on ice for 20 min, the competent cells were heat shocked at 42ºC for 2 min. Transformed bacteria were plated on LB/ampicillin/IPTG/X-Gal plates and incubated overnight at 37ºC to allow bacteria to grow.

2.1.3.3 Colony PCR

Twelve of the white colonies were picked from each plate and put to grow in 96 well plates for 4 hours with 75 µl of LB Broth plus 50 µg/ml of ampicillin. After growing, colony PCR was performed, in 12.5 µl final volume, with 1.5 µl of growth colony, 2 mM $MgCl_2$, 0.1 mM of dNTPs, 50 pmol of each T7 and SP6 primers and 0.25 U of DreamTaq™ DNA polymerase (Fermentas). PCR conditions were: 94ºC for 10 min to allow cell breakage and DNA denaturation, 30 cycles of 30 sec at 94ºC to denature DNA into single strands on every cycle, 30 sec at 48ºC for primer annealing and 30 sec to 1 min at 72ºC, variable depending on expected size of the amplicon. PCR products were run on 1.5% agarose gel electrophoresis stained by ethidium bromide.

2.1.3.4 Plasmid extraction

Extractions of the bacteria plasmids were performed by the alkaline lysis method, which uses an alkaline solution (NaOH) and a detergent (sodium dodecyl sulphate) to disrupt the bacterial cell wall and dissolve the cell membrane for releasing the plasmids to the solution.

Positive colonies were grown overnight in 3 ml of LB/ampicilin at 37ºC and constant agitation. Bacterial cultures were transferred to 1.5ml fresh tubes and centrifuged for 5 min at 12000 g. The supernatant was discarded and the pellet resuspended in 100 µl of cold GTE (Annex 4), in which RNase was added (50 µg/ml), and incubated 5 min at room temperature. Two hundred microliters of lysis buffer were added and the tubes were mixed by inversion and kept on ice until the solutions clarified (maximum of 5 min to avoid DNA degradation). Neutralization of the lysis reaction was done by addition of 150 µl of cold KAc 3M, pH 4.8. Tubes were mixed by inversion until flocculation and were incubated 15 min on ice. After centrifugation at 12000 g at 4ºC for 15 min, supernatants were recovered into fresh tubes and 450 µl of cold isopropanol were added to precipitate the plasmids. After mixing, the tubes were left 20 min at -20ºC and centrifuged at 12000 g at 4ºC for 20 min. Supernatants were discarded and the pellets were washed with ethanol at 75%. After centrifugation at 12000 g for 5 min, ethanol was discarded and the pellets were air dried at room temperature. Pellets were resuspended in 30 µl of sterile water. Plasmids quality was analyzed on a 1.5% agarose gel. After quantification, they were sequenced at the CCMAR sequencing facilities and stored at -20ºC. The sequences obtained were compared by similarity searches against the non-redundant nucleotide database at NCBI using BLASTn (http://blast.ncbi.nlm.nih.gov/).

2.2 Cloning of two antagonists of BMPs, *Gremlin* and *BAMBI*

2.2.1 Partial cDNA cloning of Gremlin

To carry out cloning of *Gremlin*, the strategy was similar to the one described previously for the putative reference genes.

Briefly, primers were designed based on conserved regions determined by a multisequence alignment of published nucleotide sequences. Two pairs of primers were used for the cloning: the second pair allows the nesting of the PCR product obtained with the first pair. Primers sequence and localization on the predicted cDNA are presented in table 4. Oligo(dT) adaptor primed cDNA was synthesized using DNA free total RNA from mature ovary.

The first PCR was carried (50 µl volume) with ~ 20 ng of cDNA, 2mM of $MgCl_2$, 0.1 mM of dNTPs, 50 pmol of each sense and antisense primers and 0.6 U of DreamTaq™ DNA polymerase (Fermentas). All reactions were performed using the following conditions: initial denaturation at 94ºC for 5 min, 35 cycles at 94ºC for 45 sec, 58°C for 45 sec, 72ºC for 45 sec and final extension at 72ºC for 10 min. After control by nested PCR, products were purified, cloned into pGemT-easy plasmid (Promega) and sequenced.

2.2.2 Full length cDNA cloning of BAMBI

A mature mRNA molecule is composed a 5'-UTR, a coding region (initiate by a methionine and terminate by a stop codon), a 3'UTR and poly(A) tail (**Figure 7**). Rapid amplification of cDNA ends (RACE) is a polymerase chain reaction-based technique, which facilitates the cloning of full-length cDNA sequences when only a partial cDNA sequence is available.

The 5' RACE which allows the cloning of the initial methionine and the 5' UTR and the 3' RACE, from which the objective is to obtained the stop codon and the 3' UTR with the poly-adenylation signal.

A large cDNA fragment of the coding region of *BAMBI* had been previously cloned which was used to obtain the full length cDNA.

2.2.2.1 5'- RACE-PCR

Classical techniques designed to clone the 5' region of cDNAs are based on the addition of an adaptor to the 5' extremity of the cDNA. The added adaptor, for which the sequence is known, will be used as site to anchor a primer during the PCR. This addition is done by template switching, extension of a poly(A) tail in the 5' of the cDNA or ligation of the adapter on the 5' end of the RNA after deletion of the protective CAP.

We have experimented an unconventional protocol in which the addition of the adaptor on the 5' extremity is not necessary. Figure 7 outlines this technique. Briefly, after cDNA synthesis using a gene-specific primer, a first PCR is done with a unique reverse primer designed upstream of the available sequence. A second PCR is done with this reverse primer and a primer designed to be complementary to the first one. During the PCR, some of primers will hybridize together but the majority anneals to the cDNA and allows the amplification of the 5' of the cDNA.

A cDNA was synthesized using a *BAMBI*-specific primer designed from available sequences (**Table 5**). Three micrograms of total RNA were denaturated at 65ºC for 10 min. Reverse transcription was carried out in a 40 µl reaction volume containing 40 U of MMLV reverse transcriptase (Promega), 8 U of Ribolock RNase inhibitor (Fermentas), 0.2 mM dNTPs and 10 pmol *BAMBI* cDNA primer. A negative control without reverse transcriptase was used. Synthesis reactions were incubated for 1 hour at 37ºC.

For the first PCR, ~ 20 ng of cDNA, 2 mM of $MgCl_2$, 0.1 mM of dNTPs, 50 pmol of *BAMBI*-5'-nor primer and 0.6 U of DreamTaq™ DNA polymerase (Fermentas) were used in a 50 µl reaction volume. PCR conditions were the following: initial denaturation at 94ºC for 5 min, 35 cycles at 94ºC for 45 sec, 55°C for 45 sec, 72ºC for 30 sec and final extension at 72°C for 10 min.

Two µl PCR of the product obtained was used as template for the second PCR and the conditions were the same as described in PCR-1 with the exception of the addition of 50 pmol of *BAMBI*-5'-opp primer. The PCR product was run on a 1.5% agarose gel before purification, ligation in pGemT-easy plasmid and sequencing.

2.2.2.2 3'- RACE-PCR

For the 3'-RACE-PCR, PCR reaction was done using an oligodT primed ovary cDNA as template. Two forward primers were designed based on the available sequence and are described in table 5.

The reaction volume was 50 µl and contained ~ 20 ng of cDNA mixed to 2mM of $MgCl_2$, 0.1 mM of dNTPs, 50 pmol of *BAMBI*-3' primer, 50 pmol of oligo(dT) adaptor and 0.6 U of DreamTaq™ DNA polymerase (Fermentas).

Reactions were performed using the following conditions: initial denaturation at 94ºC for 5 min, 35 cycles at 94ºC for 45 sec, 58°C for 45 sec, 72ºC for 45 sec and final extension at 72ºC for 10 min. After control by nested PCR with *BAMBI*-3'-nest primer, the PCR product was purified, inserted in plasmid and sequenced as previously described. The sequences were submitted to the ncbi public database. (http://blast.ncbi.nlm.nih.gov/).

2.2.3 Multisequence alignment and identification of conserved domains

Alignments were performed using ClustalW [33]. Phylogenetic and molecular evolutionary analyses were conducted using MEGA version 4 [34].

To identify conserved domains in the *BAMBI* protein sequence, the sequence was compared to the NCBI's Conserved domain database (CDD; http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml). The existence of BMP receptors transmembrane domains in *BAMBI* protein was determined using the online tool, SOSUI (http://bp.nuap.nagoya-u.ac.jp/sosui/sosui_submit.html). The main objective of this program is to determinate whether the protein is a soluble or transmembrane protein by prediction of helices in the amino acid sequences.

2.3 Tissue collection

Tilapias were obtained from a broodstock reared at the University of Algarve. Fish were maintained in re-circulating freshwater tanks under a simulated natural photoperiod and fed once a day with commercial cichlid food (Nutrafin basix®; Rolf C. Hagen Inc.).

All animals were euthanized with MS-222 (3-aminobenzoic acid ethyl ester; Sigma-Aldrich), administrated in a concentration of 1 g/L. Animals were then decapitated and dissected.

Samples were collected from 13 different tissues (brain, esophagus, stomach, anterior intestine, medium intestine, posterior intestine, liver, kidney, head kidney, spleen, heart, gill and gonad) of three males and three females. Fish were measured and weighted. Gonadosomatic index was calculated as

GSI=gonad weight/body weight*100. Immediately after dissection all tissues were flash frozen in liquid nitrogen and later stored at -80ºC until RNA extraction. Portions of the ovaries were separated previously to be frozen and fixed in Bouin-Hollande (Annex 4) to confirm maturation by histology.

2.4 Dissection of oocytes at different stages

Five sexually mature ovaries were collected from five females. Fish length, body weight and ovary weight were recorded. A portion of the ovary was fixed in Bouin-Hollande for posterior confirmation of the maturation stage by histological analysis. Gonads were dissected manually using a binocular microscope (Olympus SZX7) in Cortland medium and follicles were separated into 5 different groups according to their maturation stage: chromatin nucleolar stage (CNS), perinuclear stage (PS), cortical alveoli stage (CAS), vitellogenic stage (VS) and final maturation stage (MS) (**Figure 6**). For the first 4 stages, ~ 50 eggs were pooled while for the last stage, groups of 25 eggs were pooled and the yolk emptied. After dissection all samples were immersed in liquid nitrogen and stored at -80ºC until RNA extraction.

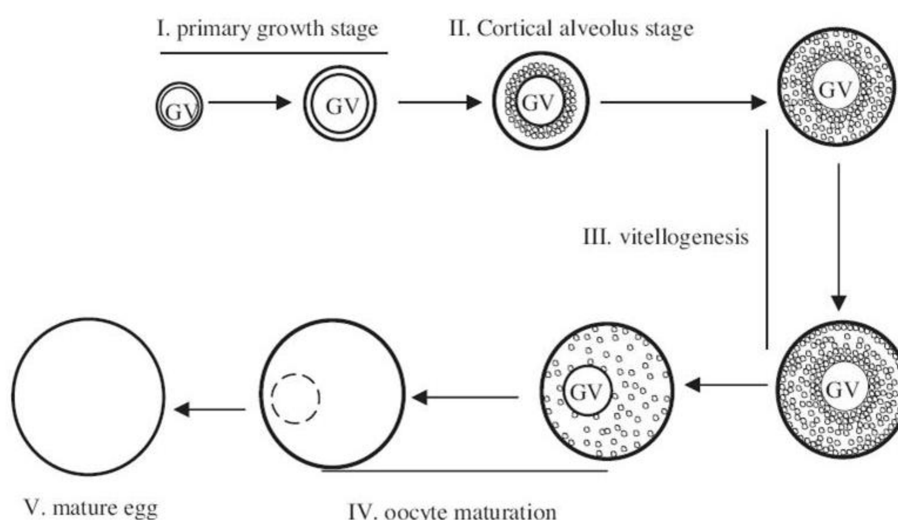


Figure 6 – Schematic drawing of zebrafish folliculogenesis. In the primary growth stage (stage I), follicles appear translucent and rangefrom 7–140 µm in size. Stage II follicles (0.15–0.34 mm) have cortical alveoli, typically surrounding the germinal vesicle. Vitellogenesis takes place in stage III follicles, whereas oocyte maturation occurs in stage IV. During oocyte maturation, germinal vesicle (GV) migrates from center to the periphery and the GV membrane breaks down. Stage V is the mature egg. (From Clelland et al., 2007 [32]).

2.5 Histology

Histological analysis was carried out on three ovaries. After 24 hours in Bouin-Hollande's, tissues were washed twice with a solution of 1x PBS/20% sucrose overnight. Ovaries were treated in a semi-enclosed tissue processor Leica TP1020 starting with dehydration in increasing ethanol gradient: 70% (10min); 95% (30 min x2); 100% (1 hour x2) - followed by a mixture of ethanol:xylene (1:1) (1 hour), xylene 100% (1 hour and 1.5 hour), a mixture of xylene:paraffin (1:1) (2 hours) and paraffin 100% (2 hours). Paraffin embedded tissues were set in paraffin blocks using the thermal, dispensing and cryo consoles (Miles Scientific).

Serial sections of 7µm were obtained using a Leica RM 2135 microtome and mounted on poly-L-lysine-coated slides. Sections were dried overnight and stained with haematoxylin-eosin (Annex 4). For the staining, paraffin was removed from the sections by immersion in xilol (2x 15min) followed by hydration through a graded ethanol series: 100% (10 min); 95% (5 min); 70% (5 min), and distilled water (5 min). Sections were immersed in haematoxylin for 5 min and washed with running water and distilled water. Sections were then immersed in eosin Y for 2 min and washed with distilled water with a few drops of acetic acid and with distilled water. Dehydration was preformed through an increasing series of ethanol (70%, 95% and 100%, 5 min each), followed by immersion in xylene (2x 15 min) and mounted with DPX.

2.6 Quantitative-Polymerase Chain Reaction

2.6.1 Primer design

Different factors were taken into account following the established common guidelines to design primers for the qRT-PCR method. To prevent the reducing of annealing efficiency it was avoid runs of 4 or more bases. Amplicon sizes were kept short, ideally between 50-200 bases in length, since this allows for faster and more efficient reactions and increased consistency of results. The choice of forward and reverse primers that span exon-exon junctions, can avoid amplification of potential contaminating genomic DNA in cDNA targets. To

reduce the occurrence of secondary sequences such as hairpins, self-dimers and cross-dimers, complementary sequences between the primers were avoided. The percentage of G/C in primers was kept between 40 to 60%.

Primers were designed for each gene based on the cloned sequences and using the software Beacon Designer 7.7 (Premier biosoft international) (**Table 3**). Primers were tested and optimized on ovary cDNA in a MyCycler gradient thermocycler, with different temperatures (50 to 60ºC). PCR products were ran on a 3% gel and sequenced to confirm primer specificity. PCR efficiencies and coefficients of determination (**Table 3**) were determined with duplicated standard curves generated from 10-fold dilution series (from 1 ng to 1 fg) of purified PCR fragments as templates and with an IQ-5 real-time PCR detection system using SsoFast™ Evagreen® supermix (Biorad).

The PCR products used for making the standard curve were prepared as the following: PCR were performed as described on section **2.1.2** for all target genes using as templates the correspondent plasmid DNA extracted (see section **2.1.3**). It was used 2 µl of plasmids samples (diluted to have the concentration of 100 ng/ul) in reactions of 100 µl. PCR products were purified on columns with GeneJET™ gel extraction kit (Fermentas) and after quantified using nanodrop ND-1000 spectrophotometer.

2.6.2 Quantitative RT-PCR on adult tissues

2.6.2.1 RNA extraction and removal of genomic contamination

Total RNA extraction was performed using 50 mg of each tissue and the Maxwell® 16 Instrument (Promega) with the respective kit and following the protocol provided by the manufacturer.

To each tissue sample, 500 µl of Lysis buffer previously prepared was added and the tissue was immediately disrupted using an ultra Turrax® homogenizer (IKA). Homogenized samples were then incubated on ice for 10 min to ensure complete lysis and transferred to fresh tubes. Blue RNA dilution buffer (835 µl) was added to the lysate followed by 125 µl of clearing agent. The tubes were mixed thoroughly, and incubated at 70ºC for 3 min. After vortexing

for 30 s; tubes were newly incubated for 5 min at room temperature, transferred to clearing columns and centrifuged at 12000 g for 2 min. The flow trough was transferred to the Maxwell® 16 RNA cartridge rack. Elution tubes were prepared by the addition of 300µl nuclease-free water. At the end of the program, eluted samples were transferred into fresh tubes.

RNAs were concentrated by precipitation. To each tube, 2.5 volumes of absolute ethanol and 0.1 volumes of NaAc (5 M, pH =8) were added. The solution was mixed thoroughly and incubated at -20ºC overnight. Following incubation, RNA was precipitated by centrifugation at the maximum speed at 4ºC during 20 min. The pellet was washed in 75% ethanol (in DEPC treated water), air dried and resuspended in 10 or 20 µl of nuclease free water depending on the total RNA pellet size. RNAs were ran in an agarose gel to check the quality and quantified by spectrophotometer.

Tissue total RNA extracted was treated with DNA-free™ kit (Ambion, USA), to remove potential genomic DNA (gDNA) contamination. The removal of the gDNA was carried out using 8 µg of RNA that were transferred into fresh tubes and to which 5 µl of rDNAse I 10x Buffer, 2 U of rDNAse I and nuclease free water were added, up to a final volume of 50 µl. Samples were incubated at 37ºC for 30 min. After addition of 5 µl of inactivation buffer to terminate the reaction, the mix was incubated 5 min at room temperature. The tubes were centrifuged at 10000 g for 1.5 min to pellet the DNase inactivation reagent. To avoid any inhibition during the cDNA synthesis related to the inactivation buffer, only 40 µl of the reaction was transferred in a new tube and stored at -20ºC.

Before cDNA synthesis, the control of DNase treatment efficiency was performed by PCR amplification with primers designed against the housekeeping gene beta-actin (*β-actin*) (Wenjing Ruan, Mao de Lai, 2007).

PCR was performed, in 25 µl final volume, with ~ 10 ng of cDNA, 2 mM MgCl2, 0.1 mM of dNTPs, 50 pmol of each sense and antisense primers (sense, 5'-GGCCGCGACCTCACAGACTAC-3' and antisense, 5'-ACCGAGGAAGGATGGCTGGAA-3') and 0.5 U of DreamTaq™ DNA polymerase (Fermentas). PCR conditions were: initial denaturation at 94ºC for 5 min, 30 cycles at 94ºC for 30 sec, 48ºC for 30

sec, 72ºC for 30 sec. PCR products (size 220 bp) were run on a 1.5%. agarose gel stained with ethidium bromide. A negative control reaction was performed in the absence of cDNA when a positive control was carried out with the cDNA prepared for the cloning of the putative reference genes (section 2.1).

RNA quantification is crucial in this protocol and, to get an accurate value, each sample was quantified three times by the nanodrop ND-1000 spectrophotometer. Ratio of the A260/A280 absorbency was recorded and all the values are between 1.8 and 2.1.

2.6.2.2 cDNA synthesis

The synthesis of cDNA was made using the iScript™cDNA Synthesis Kit (Bio-Rad). iScript is a modified MMLV-derived reverse transcriptase and is provided with RNAse inhibitor pre-blended. The iScript Reaction Mix includes a blend of oligo (dT) and random hexamer primers. The first strand synthesis was carried out with 500 ng of total RNA to which was added 4 µl of 5x iScript Reaction Mix, 1 µl of iScript reverse transcriptase and nuclease-free water up to 20 µl. The mixture was incubated as described in the protocol of the manufacturer: 5 min at 25ºC, 30 min at 42ºC and 5 min at 85ºC to conclude the synthesis.

The efficiency of the synthesis was tested by a PCR using *β-actin* as reference and PCR conditions described previously.

2.6.2.3 Measurement of gene expression

Before analysis of gene expression in tissues, a dilution of cDNA (2.5-fold dilution series dilution of 1/10 (2.5 ng equivalent RNA) to 1/3125 (8 pg equivalent of RNA) and primer concentrations were tested to determine the optimal dilution for the reaction. The determination of the sense and antisense primer concentrations was performed as described by [35].

For each experimental sample, 2 ml of the cDNA diluted 1/20 was added to 7.5 µl of SsoFast™ Evagreen® supermix and 100 nM of each reverse and forward primers (final volume: 15 µl). After activation of the enzyme by a step at 94ºC during 30 sec, 40 cycles were performed as described in the manufacturer's

instructions: denaturation at 95ºC during 5 sec and annealing/extension at 60ºC during 10 sec. The melt curve was determined by increasing the temperature of 0.5ºC by cycle from 65ºC to 95ºC.

2.6.3 Quantitative RT-PCR of oocytes

2.6.3.1 Total RNA extraction and poly(A) purification

After an unsuccessful RNA extraction from eggs using Maxwell® 16 Instrument (Promega), total RNA from the 25 samples of oocytes was extracted manually using the PureZol RNA isolation reagent. This product allows the processing of small amounts of starting material (50 cells or 5 mg of tissue). It is a monophasic solution of phenol and guanidine isothiocynate and facilitates immediate and effective inhibition of RNase activity, while lysing cells and eliminating other cellular components.

Samples were maintained in liquid N_2 after their removal from storage at – 80ºC to avoid RNA degradation. Each sample was transferred into a pyrex tissue homogenizer into which 1ml of PureZol reagent was added and cells were disrupted manually. The homogenate was transferred into a fresh tube and incubated at room temperature for 5 min for complete dissociation of nucleoprotein complexes. After incubation, 0.2ml of chloroform was added to each sample. The tube was shaken vigorously for 15 seconds and incubated for 5 min at room temperature. Samples were centrifuged at 12000 g for 15 min at 4ºC to obtain separation of the mixture into 3 phases. The lower phase or red organic phase contained the DNA and proteins as well as the white interphase situated in the middle phase. The upper, colorless aqueous phase containing the total RNA was transferred into a new RNase-free tube without disturbing the interphase thus avoiding contamination of RNA by other nucleic acids and proteins. After addition of 0.5 ml of isopropyl alcohol, the samples were mixed thoroughly and let overnight at -20ºC to precipitate the RNA. The mixture was centrifuged at 12000 g for 10 min at 4ºC. Supernatants were discarded carefully and pellet was washed with 1 ml of 75% ethanol. Pellets were air-dried for 5 min and resuspended in 10 µl or 20 µl of DEPC treated water accordingly to each pellet

size. After complete resuspension of the pellets, RNA was stored at -20ºC. RNA was quantified by nanodrop as previously described.

All samples of total RNA extracted from tilapia oocyte were treated with the MicroPoly(A)Purist™ Kit (Ambion, Applied Biosystems) for isolation of poly(A) RNA according to the manufacturer's instructions using 4 to 8 µg of RNA. This kit uses the characteristic of eukaryotic mRNAs that contain a stretch of "A" residues at their 3' ends by using a binding buffer that promotes the hybridization between the poly(A) sequences found on most mRNAs and the oligo(dT) cellulose. As an overview of the procedure the binding buffer was added to the sample of RNA as well as the oligo(dT) cellulose and incubated with continual shaking for 60 min to promote hybridization. Oligo(dT) cellulose was transferred to a spin column and washed to remove nonspecifically bound material and ribosomal RNA. Poly(A) RNA was eluted using 5-50 µl of pre-warmed THE RNA Storage Solution, accordingly, and samples were stored at -20ºC until cDNA synthesis. As previously, mRNA was carefully quantified in triplicate by spectrophotometer.

2.6.3.2 cDNA synthesis.

To confirm the absence of inhibitors in RNA samples after the poly(A) purification, we have used as an external control the Alien QRT-PCR Inhibitor Alert* (Agilent Technologies). The Alien RNA transcript used is an *in vitro*-transcribed RNA molecule that has no significant homology to any known nucleic acids.

After addition to each sample, of 10^5 copies of ALIEN RNA, first strand cDNA was performed with 50 ng of mRNA and the Iscript™ cDNA synthesis kit (Biorad) following manufacturer's instructions. To control the inhibition of the reaction, a tube containing only the ALIEN RNA was reverse transcribed in the same conditions.

2.6.3.3 Quantification of gene expression

Before analysis of biological samples, the potential inhibition of the reverse transcription was controlled by a qRT-PCR against ALIEN RNA following the

instructions of the manufacturer (Agilent Technologies). The C_t from the sample containing only the ALIEN RNA was compared to the C_t obtained for the biological samples.

Quantitative RT-PCR reactions were carried out in the same conditions as described in the section 2.6.2.3. For each maturation stage, 5 independent samples were tested in technical duplicates. The absence of interference with the ALIEN transcript was tested for each couple of primers. Negative reactions without cDNA were also performed.

2.6.4 Quantitative PCR data analysis

Amplification specifity for each qRT-PCR analysis was confirmed by melting curve analysis. Amplification efficiencies for each gene and coefficients of determination (R^2) between concentration and signal intensity were calculated from the slopes of the standard curves for all ten potential reference genes. The quantification cycle (Cq) data were then collected for all selected tissues and oocytes of each of the five groups. These data were either used directly for stability calculations or were first transformed to relative quantities using the delta-Cq method and the gene-specific amplification efficiency.

Data analysis was performed using two software packages, namely geNorm (version 3.5) http://medgen.ugent.be/~jvdesomp/genorm [3] and Normfinder http://www.mdl.dk/publicationsnormfinder.htm [47].

GeNorm makes pairwise comparisons between one endogenous control and all other reference genes, in all samples. The software ranks endogenous controls gene stability by average expression stability value (M). All the candidate reference genes across tissues and oocytes were ranked according to their *M* values. The program recommends using an *M* value below the threshold of 1.5 to identify (sets of) reference genes with stable expression. The determination of the optimal number of reference genes was also determined with geNorm, by calculating the pairwise variation (V score) V_n/V_{n+1} between two sequential normalization factors NF_n and NF_{n+1} that contain an increasing

number of reference genes. Normalization factors are calculated by stepwise inclusion of extra, less stable reference gene to determine how many reference genes should be used. A large variation means that the added gene has a significant effect on normalization factor and should be included for calculation.

Vandesompele et al [3] propose pair-wise variation of 0.15 as a cut-off under which the inclusion of additional reference gene is not required. However, accordingly to geNorm manual (http://medgen.ugent.be/~jvdesomp/genorm/geNorm manual.pdf) [53], the proposed 0.15 value must not be taken as a too strict cut-off. The pairwise variation is only intended to be guidance for determination of the optimal number of reference genes. Sometimes, the observed trend (of changing V values when using additional genes) can be equally informative. Just using the 3 best reference genes (and ignoring this second graph) is in most cases a valid normalization strategy, and results is much more accurate and reliable normalization compared to the use of only one single reference gene.

The NormFinder software is an application for Microsoft Excel, that provides information on intra- and inter-group variability, the information regarding best endogenous control as well as best combination of endogenous controls. More stable gene expression is indicated by lower average expression stability values. In summary, the program ranks the various candidate reference genes according to their expression variation inter and intra groups and finds both the single gene with most stable expression and the best pair of genes with combined most stable expression. Also, in this program the accumulated standard deviation is a good indicator of the optimal number of reference genes.

Table 2 – Primers used for the cloning of the putative reference genes and expected size of the amplicons.

Gene abrev.		Primer sequences	Expected size (bp)	Temperature (ºC)
β-actin	sense	5'– ATGGAAGATGAAATCGCCGCACTGG –3'	1120	58
	antisense	5'– TTAGAAGCACTTGCGGTGGACGAT –3'		
B2m	sense	5'– ATCTGCCAGGTGAGTTATTTCCACC –3'	220	58
	antisense	5'– GGTGTTTACATGTTTGACTCCCA –3'		
CTSD	sense	5'– GCCAACTCCAGCACATATGTGAAGAA –3'	790	58
	antisense	5'– ATGAACACATCTCCCAGAATCCACAG –3'		
CTSZ	sense	5'– CCACAGTACTGCTCCTGCTGGGC –3'	650	58
	antisense	5'– ATAGGATCTCCGTACATGCAGT –3'		
EF1a	sense	5'– CACATCAACATCGTGGTCATTGGCC–3'	1030	58
	antisense	5'– CCGGGGTGGTTCAGGATGATGACCT –3'		
GADPH	sense	5'– GACGTGGAGTACATGGTCTACATGTT –3'	510	58
	antisense	5'– GTGGAGGCAGGGATGATGTTCTG –3'		
HPRT1	sense	5'– GGCCAGAGAGATCATGAAGGAAATGGG –3'	350	58
	antisense	5'– TCACCACCAAACTTGCAACTTAACCATT –3'		
TBP	sense	5'– AATGCTGAATACAATCCAAAGCG –3'	800	58
	antisense	5'– TAGATTCTCTCACCTTGGCACCTGT –3'		
Tub A	sense	5'– GACGACTCCTTCAACACCTTCTT –3'	1120	58
	antisense	5'– CTCTGAGTACTCTCCCTCCTCCAT –3'		

Table 3 – Primers used for the q-PCR of the putative reference genes. Amplicon sizes were calculated based on the sequences known for each gene. Efficiency and r^2 were determined for each couple of primers for standard curves.

Gene abbrev.		Primer sequences	Product size (bp)	Efficiency	r^2
18S	sense	5'– AATCGCTCCACCAACTAAGAACGG –3'	-	94.9 %	0.989
	antisense	5'– TGACGGAAGGGCACCACCAG –3'			
β-actin	sense	5'– TGACCTCACAGACTACCT –3'	179	92.5 %	0.995
	antisense	5'– GCTCGTAACTCTTCTCCA –3'			
B2m	sense	5'– GAAGAATGGGGCAGAAAT –3'	79	97.7%	0.998
	antisense	5'– CTTGGTCAGATGGAAGTG –3'			
CTSD	sense	5'–CAGAAGAAGGTGGAGAAGA –3'	130	96.3 %	0.998
	antisense	5'– TGTTGGCATAGTGGAAGT –3'			
CTSZ	sense	5'– AGCCTACATCAACCACAT –3'	156	103.2 %	0.999
	antisense	5'– TACTTACTGCCACTTCCTC –3'			
EF1a	sense	5'– GTTCAAGGGATGGAAGATTG–3'	189	95.6 %	0.996
	antisense	5'– TTCAGGATACCAGTCTCAAC–3'			
GADPH	sense	5'– GCTGGTCATCGGTAACA –3'	126	101.8 %	0.998
	antisense	5'– GCCTTCTCAATGGTGGTA –3'			
HPRT1	sense	5'– TGAACCAGAACAGCGATA –3'	162	102.5 %	0.999
	antisense	5'– TCTCCACAATATCCTCAACA –3'			
TBP	sense	5'– GGCAACTGTGATGTGAAG –3'	135	96 %	0.997
	antisense	5'– GATGAGCAGGACAATTCG –3'			
Tub A	sense	5'– TTCCTCATCTTCCACTCATT –3'	98	94.1 %	0.994
	antisense	5'– AACTTAGACTTCTTGCCGTA –3'			
Gremlin	sense	5'– CCAGTATTTGAAACGGGAC–3'	96	102.6 %	0.998
	antisense	5'– CAGAAGCGGTTGATGATG–3'			
BAMBI	sense	5'– AGTCCAGTTTCATTTCCATTT –3'	87	100.2 %	1
	antisense	5'– CACAGTAGCACCTAATTTCTC–3'			

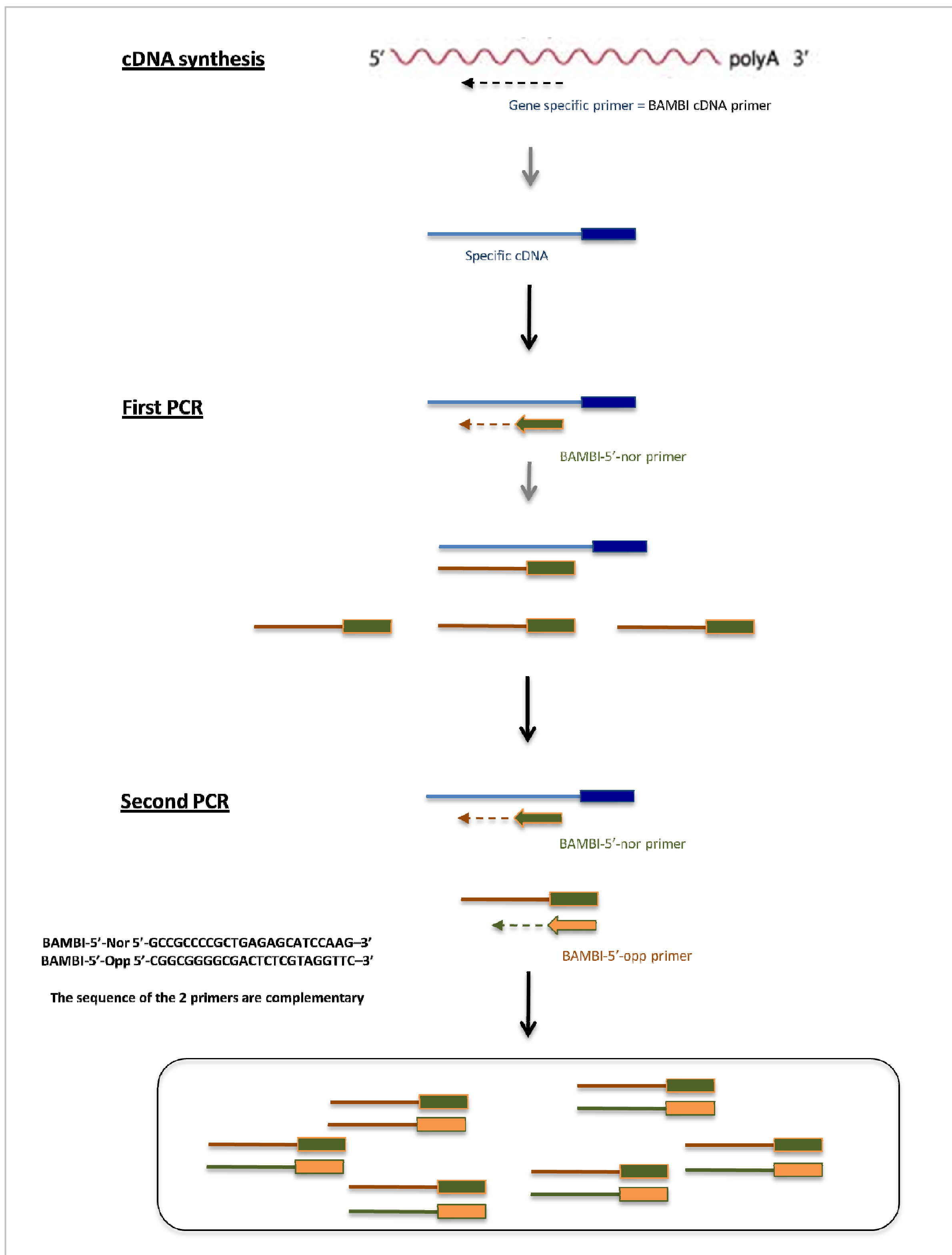


Figure 7 – Scheme of the different steps of RACE-PCR to obtain *BAMBI* full length cDNA.

Table 4 - Primers for partial cloning of *Gremlin*. Expected sizes were calculated based of the multialignment of available vertebrate protein sequences.

		Primer sequences	Product size (bp)
<u>Cloning primers</u>			
Grem-Sense	sense	5'- GACGAGGTGCTGGAGTCCAG –3'	320
Grem- Antisense	antisense	5'– CAACGGCACTGTTTGACGCGC –3'	
<u>Nested primers</u>			
Grem-Nest-Sense	sense	5'– CATCATCAACCGCTTCTGCTACGG –3'	215
Grem-Nest-Antisense	antisense	5'– GGATGTAGAAGGAGTTGCACTG –3'	155

Table 5 – Primers for RACE-PCR to obtain *BAMBI* full length cDNA.

	Primer sequences	Minimum product size (bp)
5'-RACE-PCR		
<u>CDNA synthesis</u>		
***BAMBI* cDNA**	5'– TAGCGGCCATCTGTCGAGTCTCTGGTGTG –3'	270
<u>PCR primers</u>		
***BAMBI*-5'-Nor**	5'– GCCGCCCCGCTGAGAGCATCCAAG –3'	24
***BAMBI*-5'-Opp**	5'– CGGCGGGGCGACTCTCGTAGGTTC –3'	
3'-RACE-PCR		
***BAMBI*-3'**	5'– ACGAAATCAGGTGCTACTGTGA –3'	700
***BAMBI*-3'-nest**	5'– AATTAGGTGCTACTGTGATGCGCCG –3'	700

3. RESULTS

3.1 Histology

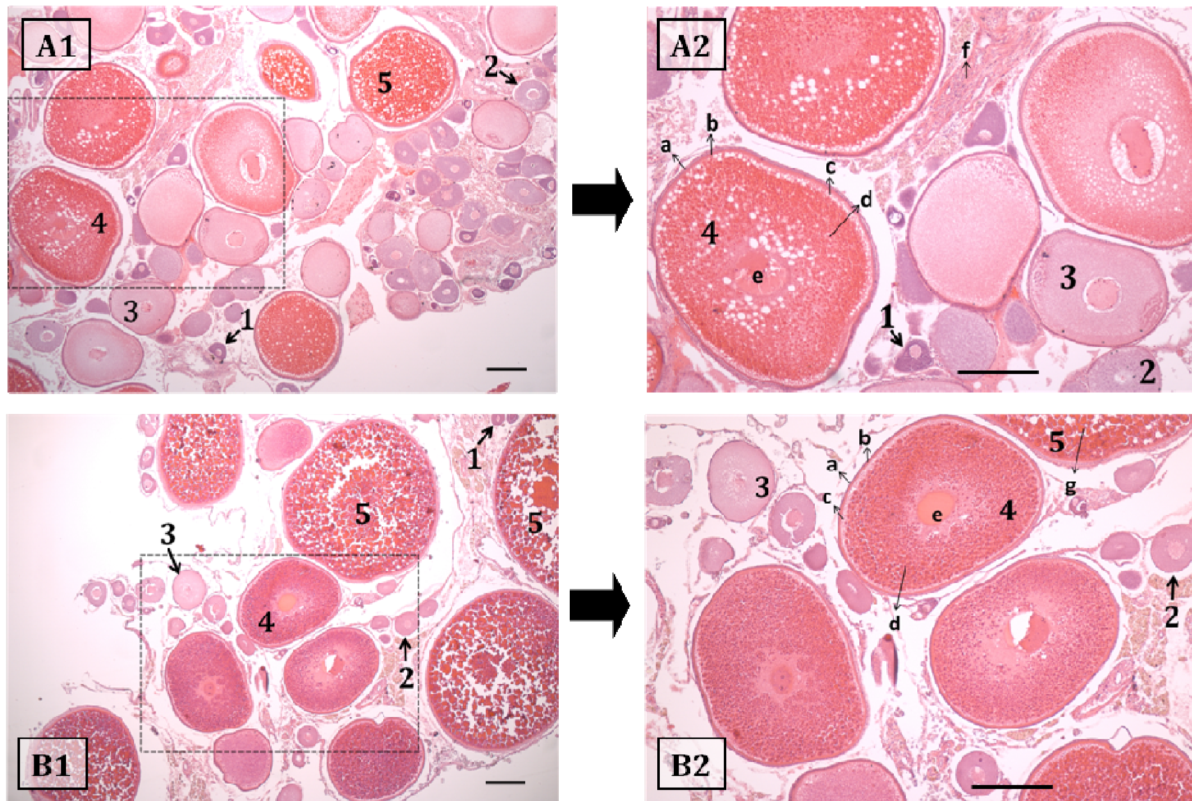


Figure 8 – Photomicrographs of *O. mossambicus* ovaries showing overall histological structures (H & E). Panel **A1** (5x) and **A2** (10x) are representative sections of ovaries from females used for quantification of gene expression in somatic tissues where **A2** corresponds to the dashed rectangle area in **A1**. Panel **B1** (5x) and **B2** (10x) are representative sections ovaries used for the determination of gene expression in groups of oocytes where **B2** corresponds to the dashed rectangle area in **B1**. Ovary of *O. mossambicus* contains different stages of oocytes visible in all pictures: Oogonia (**1**); Primary oocytes (**2**); early vitellogenic oocytes (**3**); late vitellogenic oocytes (**4**) and mature oocytes (**5**). Different structures that compounds a late vitellogenic oocyte (**4**) are visible namely the theca (**a**), granulosa (**b**), zona radiata (**c**), ooplasm with yolk granules (**d**), and germinal vesicle (**e**). Connective tissue (**f**) is visible in picture **A2**. Yolk globules are stained intensely in red by eosin (**g**) and are identified in picture **B2** incorporated in the ooplasm. Bar scale = 500μm.

The histological sections in figure 8 are representative of the females used in this work. Panel **A1** and **A2** were obtained from an ovary of a female used for analysis of gene expression of somatic tissues, while panels **B1** and **B2** are from a histological section of an ovary from a female used in the analysis of follicular gene expression.

Five different stages of follicle development have been identified in the sections. Oogonia (**1**) are small spherical cells with large nucleus surrounded by an ooplasm which stains dark purple with H&E. Primary oocytes (**2**) are identified by having a larger nucleus and an increased mass of ooplasm visible as more pale purple than oogonia. The third stage identified is the early vitellogenic oocyte (**3**) where oocytes continue to enlarge, clear vesicles (cortical alveoli) start to appear at the periphery of the ooplasm and zona radiata is also visible. In the late vitellogenic oocyte (**4**) the oocytes reach their maximum size with small yolk granules accumulated in the ooplasm. In this stage it is clearly visible the germinal vesicle (**e**) and different layers are also visible: theca (**a**); granulosa (**b**); zona radiata (**c**). The last stage identified are the mature oocytes (**5**) in which oocytes are markedly increased and large yolk globules (**g**) are dispersed on the ooplasm, visible by intense red staining by eosin. The germinal vesicle is no longer visible at this stage.

The GSI for all individuals sampled is shown in tables 6 and 7.

Table 6- Values recorded of the total weight, gonads weight and standard length of the 3 females and 3 males dissected to obtain tissue samples. The gonad somatic index (GSI) was calculated (gonad weight/total weight x 100) for all fish.

Individual	Total Weight (g)	Gonads Weight (g)	Standard Length (cm)	GSI
F1	244	1.60	24.0	0.656
F2	189	1.69	22.8	0.894
F3	228	2.40	23.0	1.053
M1	258	1.98	25.8	0.767
M2	290	2.13	26.7	0.734
M3	295	2.90	26.0	0.983

Table 7- Values recorded of the total weight, gonads weight and standard length of the 5 females dissected to obtain oocytes samples. The gonad somatic index (GSI) was calculated as gonad weight/total weight x 100 for individual fish.

Individual	Total Weight (g)	Gonads Weight (g)	Standard Length (cm)	GSI
F1	75.8	1.4	13.9	1.843
F2	72.8	1.3	12.8	1.78
F3	75	1.2	13.5	1.24
F4	68.8	0.6	13.1	0.872
F5	81	0,7	15	0.987

3.2 Quantitative RT-PCR

All ten potential reference genes were successfully cloned, multisequence alignment was carried out for each predicted protein sequence to analyze their conservation (**Annex 5**). The cloned sequences were submitted to the NCBI public database (http://blast.ncbi.nlm.nih.gov/) and identified with the accession numbers listed in table 8.

The genes that were used to test the reference genes, *BAMBI* and *Gremlin*, were also successfully cloned and the respective accession numbers are also listed in table 8. Multisequence alignments were performed for the two sequences and matrix of identities and phylogenetic threes were constructed showing the conservation and evolutionary history of both genes (**Annex 6**).

Table 8 – Accession number of all genes cloned in this study.

Gene abbrev.	Accession number
18S	*AF497908*
β-actin	FN673689
B2m	FN673687
CTSD	FN646079
CTSZ	FN673688
EF1a	FN597061
GADPH	FN673690
HPRT1	FN673691
TBP	FN673691
Tub A	FN673691
Gremlin	FR715328
BAMBI	FN543097

Based on the cloned sequences, primers for qRT-PCR reactions were designed. All primer pairs were highly specific as demonstrated by the single peaks of the melting curve analysis (**Annex 7**). The amplification efficiency range of the qRT-PCR for all of the genes tested was between 92.5 % and 103.2 % and the R^2 values of the standard curves were between 0.982 and 1 (**Table 4** in section 2; **Annex 7**).

3.2.1 Quantitative PCR on adult tissues

Analysis of individual relative expressions levels showed that all selected genes were expressed in all tissues under study.

Figure 9 presents an overall expression representation of the candidate genes in the tissues. The gene with the lower Cq value is *18S* and, therefore, is the most expressed gene. In contrast, GAPDH is the gene with the highest Cq (lower expression in tissue) and with the highest variation in the Cq values. This figure also show that *TBP* is the second gene with less variation and that the genes *Tub A* and *CTSZ* have the most identical results when combining variation and Cq values. This figure allows us to have a first estimation of the variability of the candidate genes in all male and female tissues.

In figure 10 it is possible to compare the Cq variation of each gene in each female and male tissues. This representation is coherent with the previous figure and shows *18S* as the most stable gene and *GADPH* the most variable. It also reveals that there are no large differences in gene expression between females and males with few exceptions in the ovary. In this tissue there are considerable differences between Cq values of *GADPH*, *TBP*, *B2m* and *CTSD*.

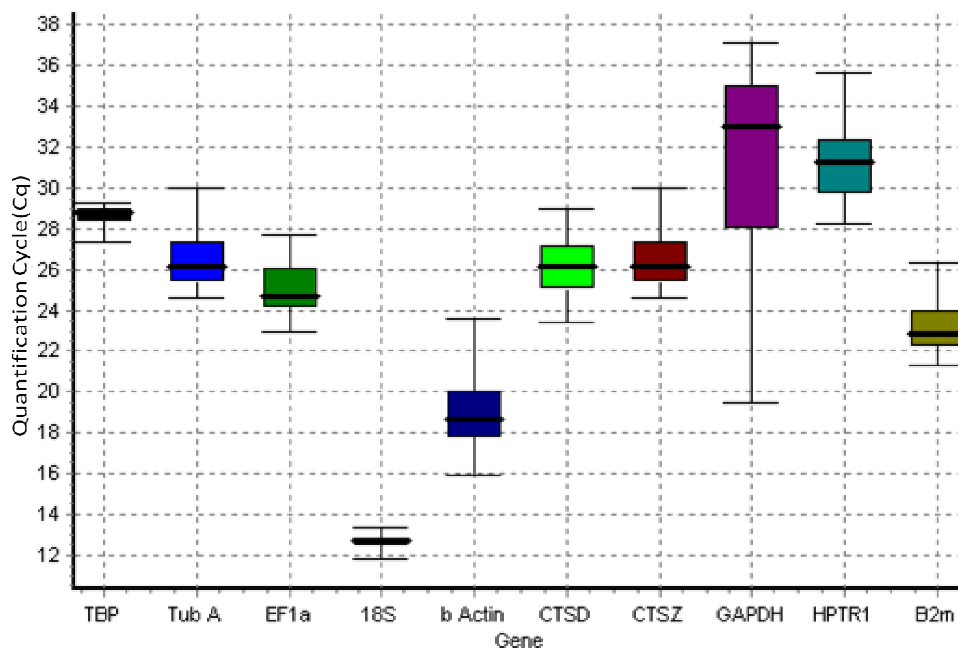


Figure 9 - Overall expression representation of the candidate reference genes in adult tissues. For each gene 13 tissue from male and female were analyzed. The median is marked via a bold line. Bars indicated the 25/75 percentiles.

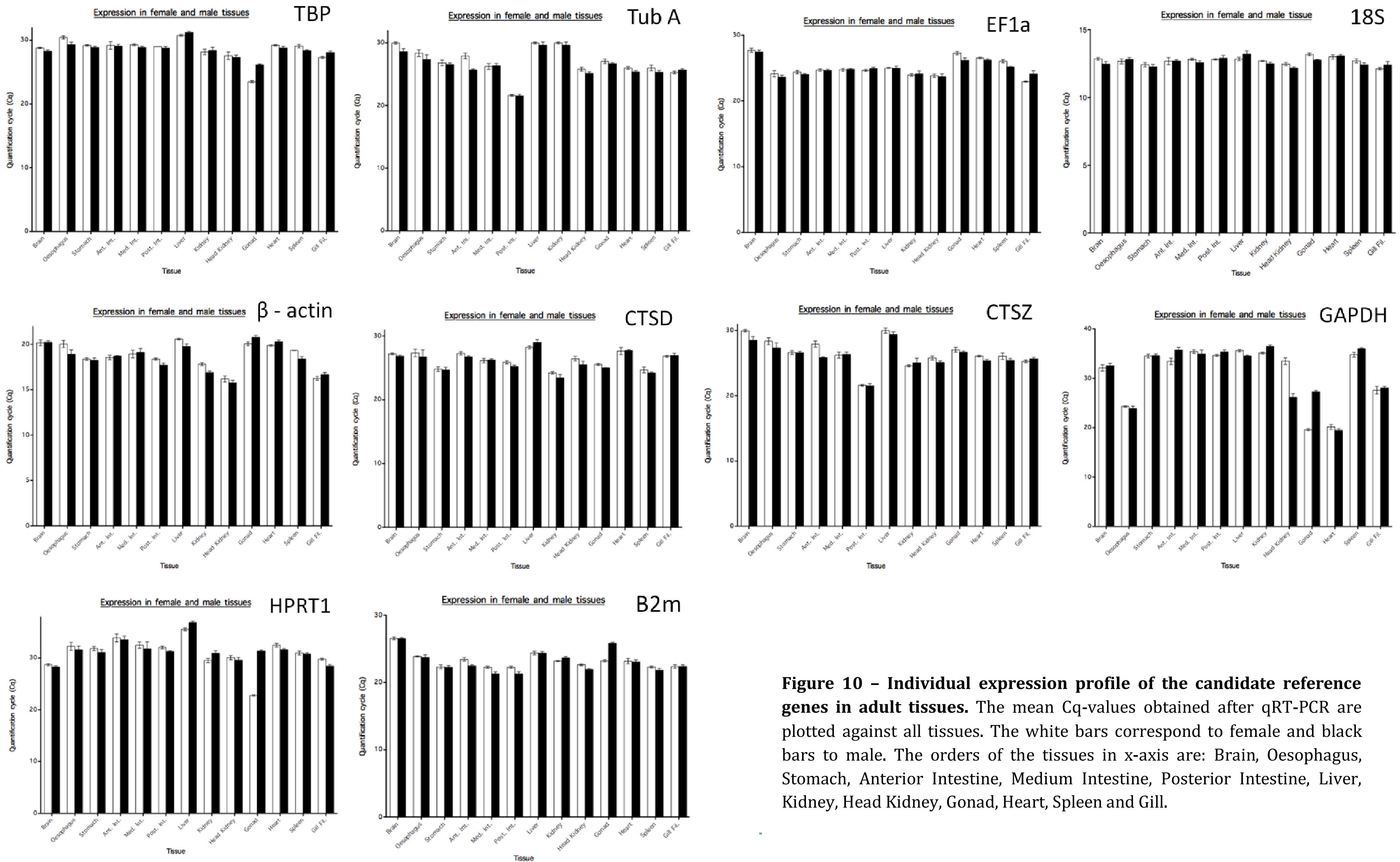


Figure 10 – Individual expression profile of the candidate reference genes in adult tissues. The mean Cq-values obtained after qRT-PCR are plotted against all tissues. The white bars correspond to female and black bars to male. The orders of the tissues in x-axis are: Brain, Oesophagus, Stomach, Anterior Intestine, Medium Intestine, Posterior Intestine, Liver, Kidney, Head Kidney, Gonad, Heart, Spleen and Gill.

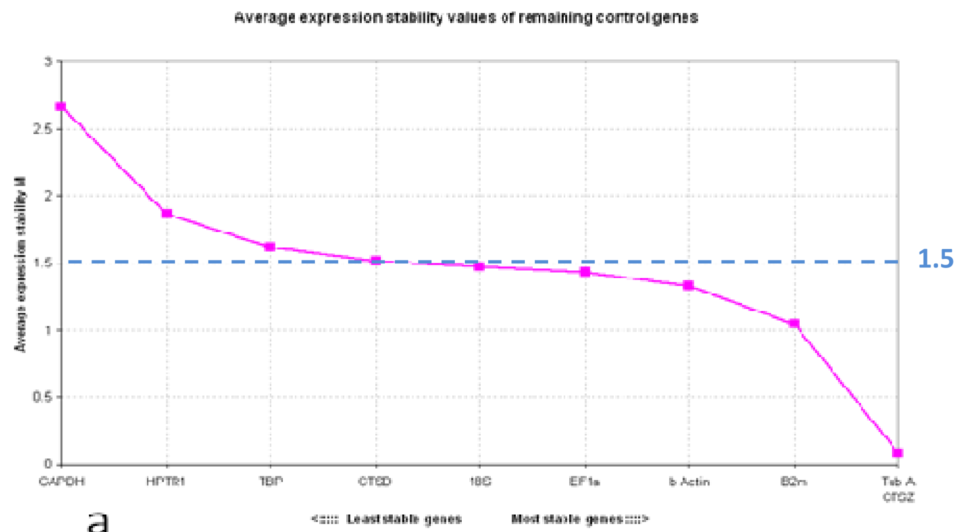
3.2.1.1 GeNorm analysis in adult tissues

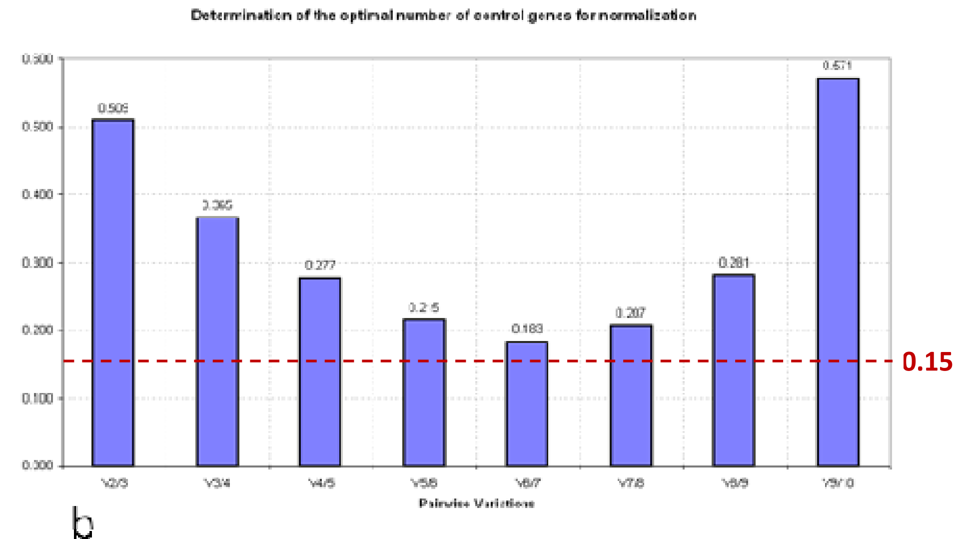
The results of this program for the average expression stability (M values) of all genes in all tissues, indicated that *CTSZ* and *Tub A* were the two most stable genes (**Fig 11 a**). The gene *B2m* was the third most stable gene with an M value around 1.1. *β-actin*, *EF1a* and *18S* were respectively the 4th, 5th and the 6th more stable genes, all with M values near but lower than 1.5. *CTSD* and *TBP* had M values slightly above 1.5 and *GADPH* and *HPRT1* were the two least stable genes. *GADPH* had a distant M value (2.7) comparing to the rest of the genes with M values around 1.5.

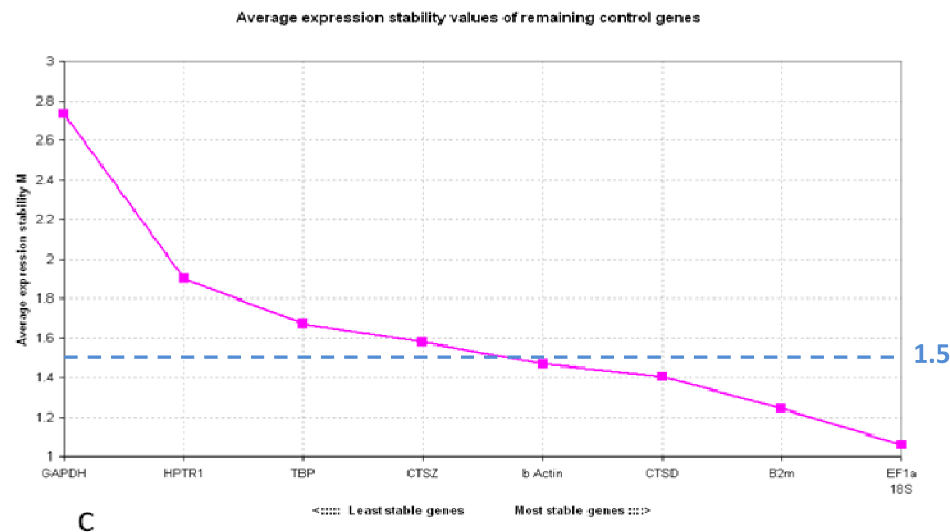
In order to determine the optimal number of reference genes, geNorm also calculated the pairwise variation between the two most stable genes until all genes together by adding in each time the next least stable gene. The determination of the optimal number of control genes, showed on figure 11 (b), revealed that the cut off value of 0.15, recommended by the program, was never achieved. Nevertheless, there is a large pair wise decrease between V2/3 and V3/4, this is indicative that a 4th gene could be included for calculation of a reliable normalization factor. Although simply using the three best reference genes is also a valid normalization strategy as was aforementioned in the last section of the methods (section 2.6.4). Therefore, for adult tissues the most stable reference genes for the accurate calculation of normalization factors should be *Tub A*, *CTSZ* and *B2m*.

Because of the possible bias due to the similar pattern of expression of *Tub A* and CTSZ, we have performed a second analysis excluding *Tub A*. In fact, geNorm analyses the genes pair wise and if two genes have really similar expression pattern, the program considers them as reference of stability. After exclusion of *Tub A*, the analysis revealed different results and is represented in figure 11 (c) and (d). Figure 11 (c) shows that the three genes least stable are the same (*GADPH*, *HPRT1* and *TBP*) as previously and that *B2m* maintained the same position. The remaining genes changed their ranking position and in this case *Efa1* and *18S* were identified as the most stable.

Figure 11 (d) represents the determination of optimal number of genes among all selected genes, excluding *Tub A*. The analysis revealed that the cut off value of 0.15 was not achieved as previously, therefore the most stable 3 genes (*18S*, *Efa1* and *B2m*) should be used for normalization.







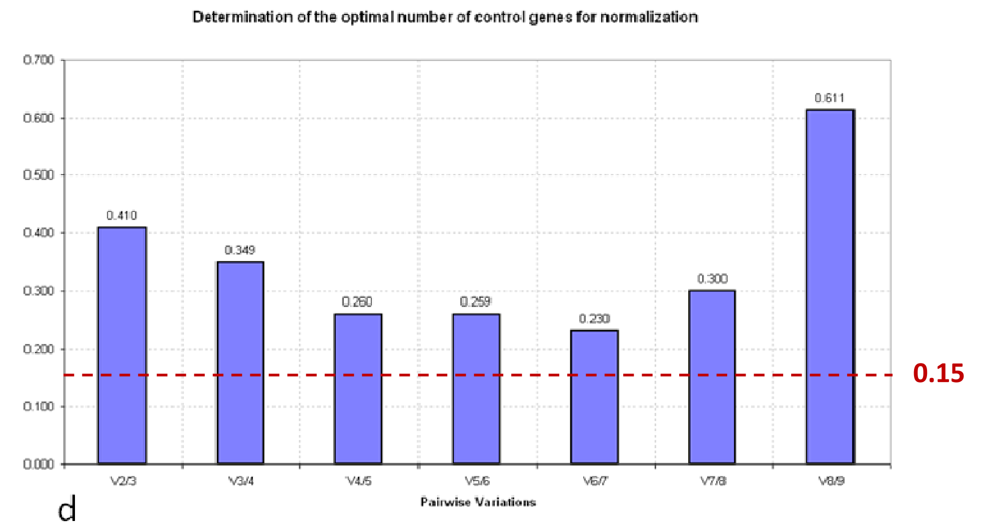
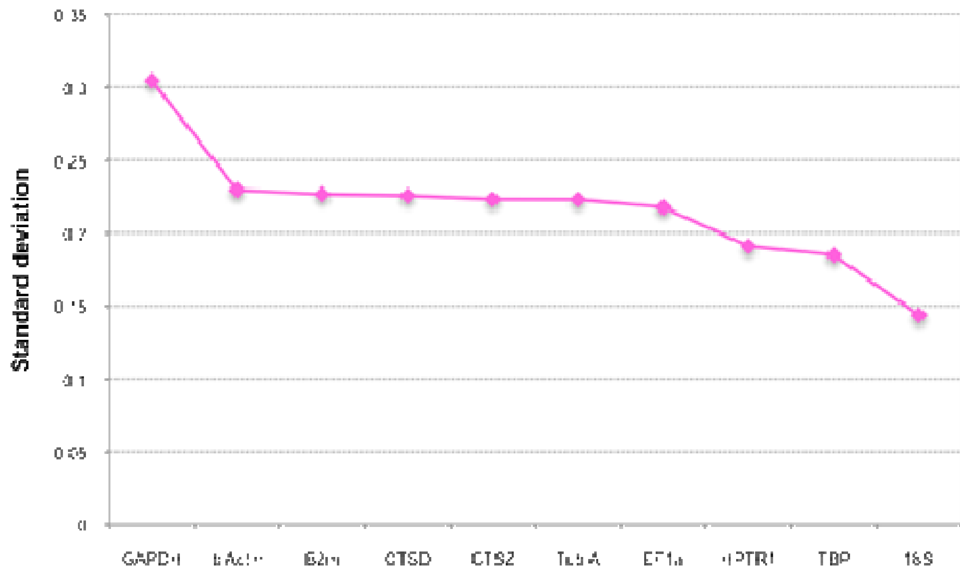


Figure 11 - Selection of the most suitable reference genes for normalization in adult tissues using *geNorm analysis.* In (**a**) and (**b**) all candidate genes were analyze and in (**c**) and (**d**) the gene *Tub A* was excluded. In (**a**) and (**c**)**,** the genes expression stability increases from left to right along with the decreasing of **M values** (y-axis). Determinations of the optimal number of reference genes for normalization are represented in (**b**) and (**d**). **Pair wise variation** of normalization factors after successive inclusion of less stable genes determined the optimal number of reference genes. On the left-most side is the pair wise variation when the number of genes is enlarged from 2 to 3 (V2/3). Stepwise inclusion of less stable genes generates the next data points.

3.2.1.2 Normfinder analysis of reference genes in somatic tissues

The analysis by this program is represented in figure 12 which shows that *18S* is ranked as the most stable gene followed by *TBP* and *HPRT1*. In figure 12 (b) the accumulated standard deviation shows the largest decrease with the two more stable genes, suggesting the use of these genes (*18S* and *TBP*) for normalization.



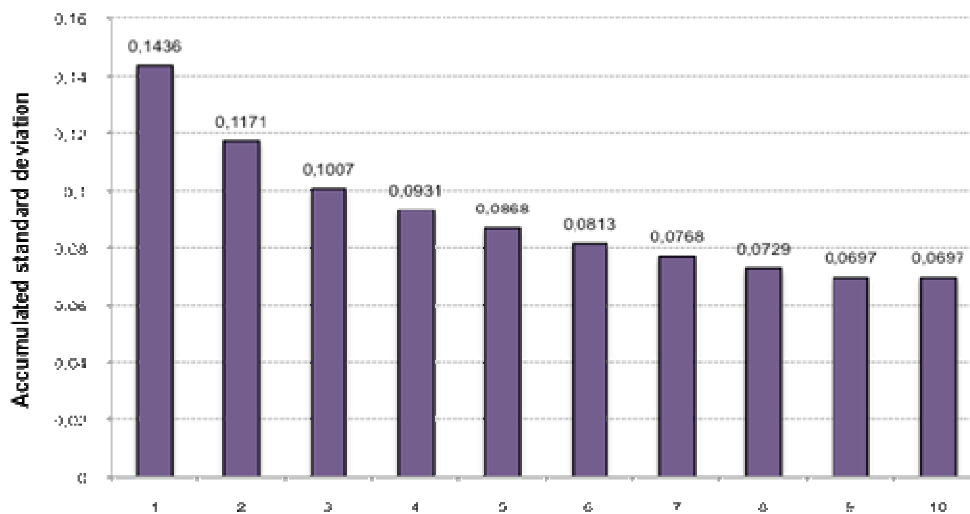


Figure 12 - Selection of the most suitable reference genes for normalization in adult tissues using *Normfinder* analysis. In graph (**a**) it is represented the order of expression stability of all selected genes and in (**b**) is represented the determination of the optimal number of reference genes for normalization. The accumulated standard deviation is a great indicator of the optimal number of reference genes.

Table 9 - Ranking of the candidate reference genes in adult tissues, according to their stability value using *geNorm* and *Normfinder* analysis. The ranking order (GeNorm c) doesn't include Tub A.

	Ranking Order (geNorm a)	Ranking Order (geNorm c)	Ranking Order (Normfinder)
Tub A	1	-	5
CTSZ	1	5	6
B2m	2	2	8
B actin	3	4	9
EF1a	4	1	4
18S	5	1	1
CTSD	6	3	7
TBP	7	6	2
HPRT1	8	7	3
GADPH	9	8	10

In table 9 we have summarized the results obtained with the geNorm and NormFinder software. Both analyses ranked *GADPH* in last place in terms of stability. Considering geNorm c analyzes done by deleting *Tub A* and the NormFinder analyses, and as expected from figures 9 and 10, *18S* is placed first. In geNorm c, *TBP* was ranked as the third least stable gene while in Normfinder *TBP* was the second most stable gene. There is also a large difference in rankings for the gene *HPRT1*, ranked by geNorm as the second least stable gene, while Normfinder ranked it as the third most stable. However, differences in the way rankings are calculated by the two programmes make it difficult to make direct comparisons.

3.2.2 Quantitative RT-PCR on Oocytes during development

To confirm the absence of inhibitors, we have incorporated in the samples an ALIEN® RNA transcript before the cDNA synthesis. The control is a tube without sample RNA but with the same quantity of ALIEN® RNA transcript. The principle is that the transcript is synthesized in the same way that the sample RNA. A PCR is performed on all the tubes with specific primers designed against this ALIEN. The Cq obtained for the control is compared to the Cq of the samples. In case of inhibition the Cq of the ALIEN will be significantly lower than the Cq of the samples. In our case, no inhibition is detected, because the Cq from the control was not significantly different of the samples (**Figure 13**).

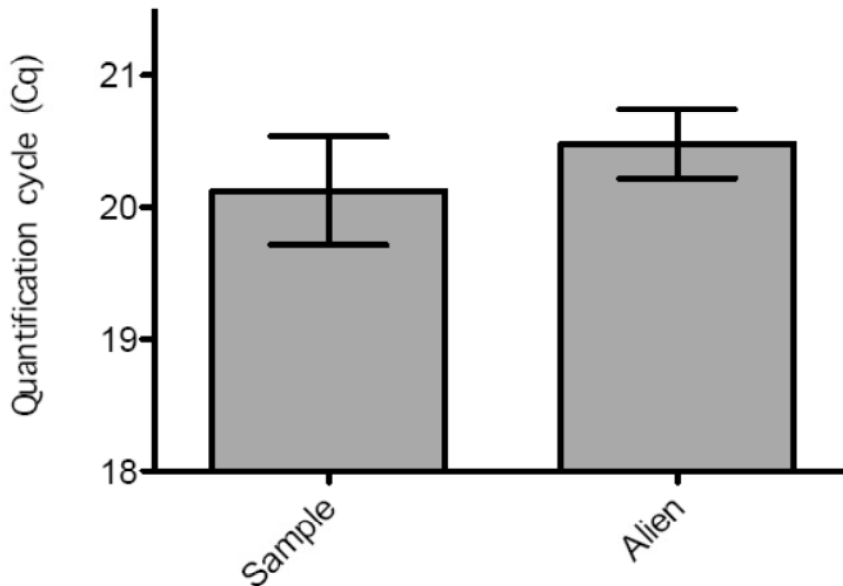


Figure 13 – Control of inhibition in oocytes cDNA synthesis with alien RNA. Comparison between amplification by qRT-PCR of a known amount of Alien RNA in the presence and absence (control) of the experimental RNA samples of interest.

Figure 14 shows the overall expression of the candidate genes in the different stages of maturation. *EF1a* was the gene with less variation among all groups of oocytes, with Cq values around 21 varying less than 1 Cq value. *β-actin* and *CTSZ* where the next two genes with less variation, but with different transcription levels as indicated by the different Cq values. The genes *B2m* and *CTSD* had similar variation among groups but also with different transcription levels. *GADPH*, *HPRT1*, *Tub A* and *TBP* were the genes with more variation of transcription among all groups of oocytes with variations of Cq values around 3 cycles. In general, the Cq values are less variable in this study than in the tissue analysis.

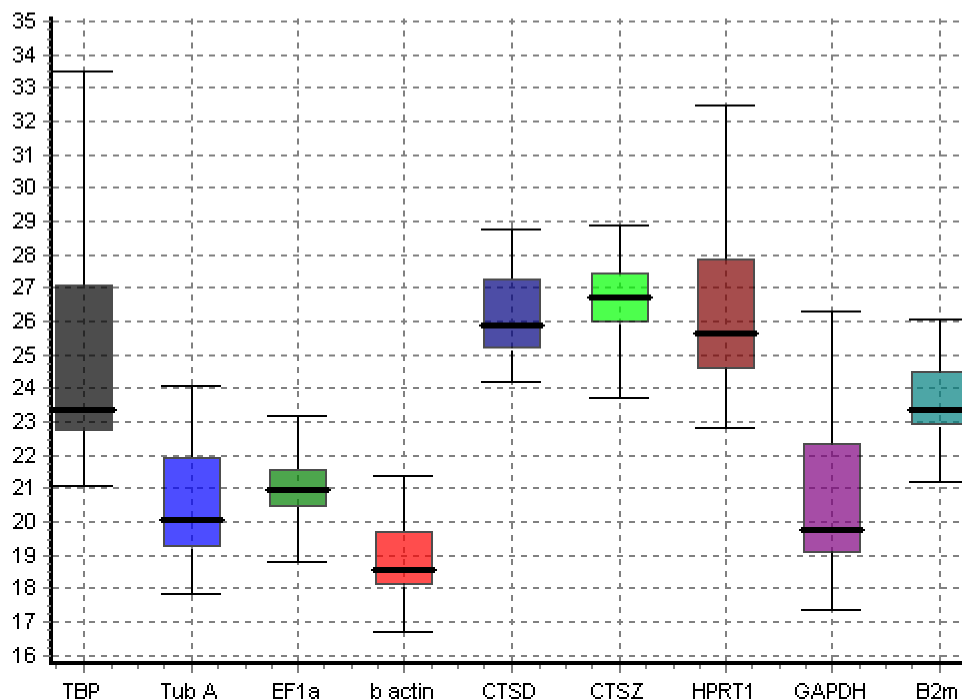


Figure 14 - Overall expression representation of the candidate reference genes in oocytes. For each gene, 5 groups of oocytes in different stages of development were analyzed. The median is marked via a bold line. Bars indicated the 25/75 percentiles.

The individual expression profiles of the candidate reference genes in oocytes were also analyzed and are represented in figure 15. It is possible to see that the genes are grouped in three categories. In the first category *TBP*, *Tub A*, *HPRT1* and *GADPH* show an increase of Cq values as oocytes mature and are presented in figure 15 with the red color. These four genes have a significant increase of Cq values in the last two maturation stages (G4 and G5). In the second

category, presented in blue, the genes *CTSZ*, *Efa1* and *B2m* show a decrease of Cq values with a significant decrease in the last stage of maturation (G5). And in the third category, *β-actin* and *CTSD*, which are presented in green, had alternate transcription levels between all groups (stages) of eggs with no significant differences.

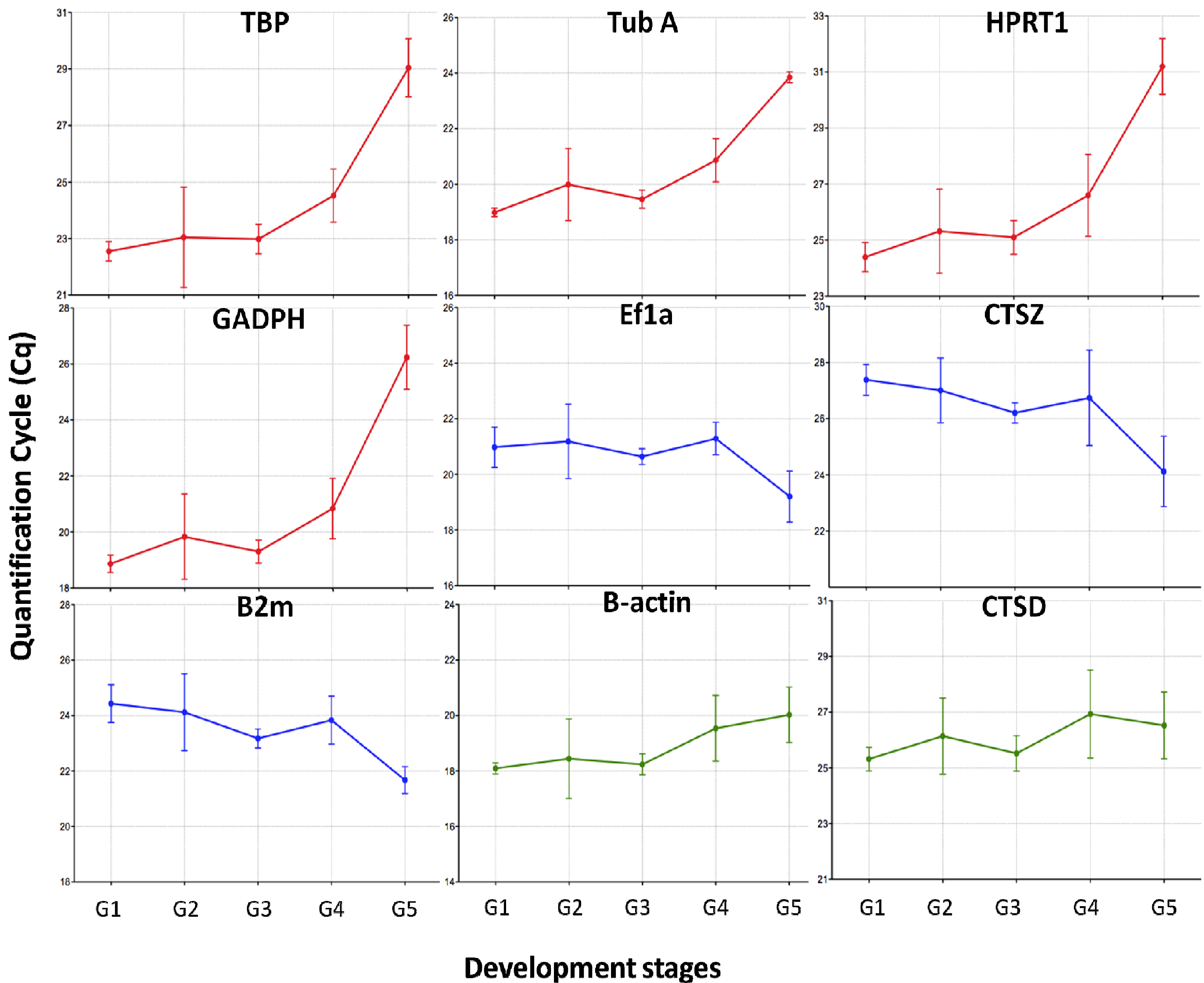
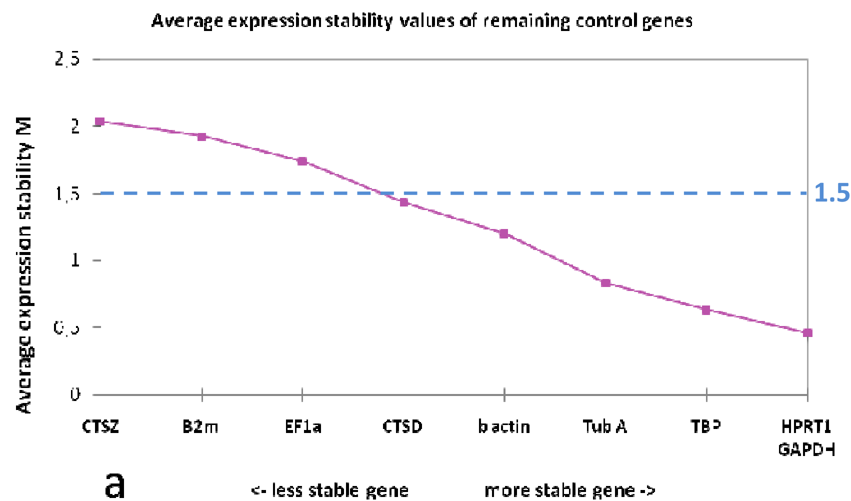


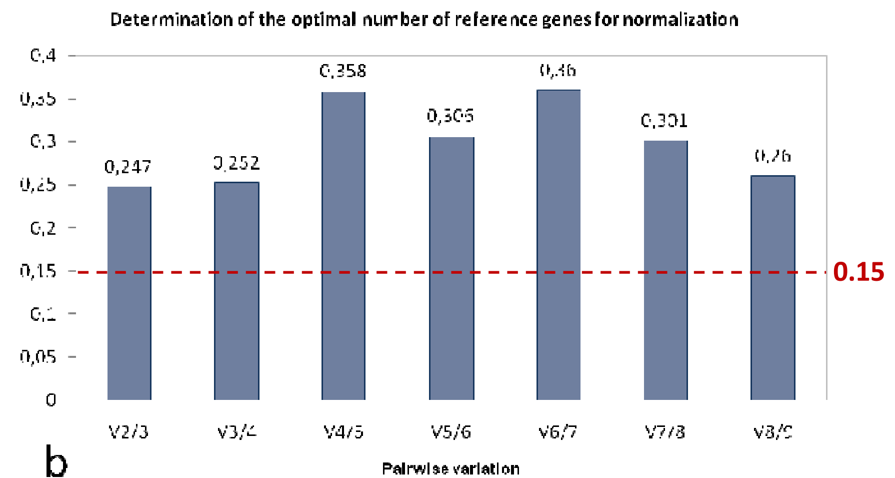
Figure 15 – Individual expression profiles of the candidate reference genes in oocytes. The quantification cycle (Cq) of each candidate reference gene was measured between each of the five groups of oocytes analyzed (G1 – Oogonia; G2 - Primary oocytes; G3 - early vitellogenic oocytes; G 4 - late vitellogenic oocytes; G5 - mature oocytes). The genes with similar patterns are grouped with different colors. The genes *TBP*, *Tub A*, *HPRT1* and *GADPH* have the same pattern of increasing Cq values and are presented in red. The genes *EF1a*, *CTSZ* and *B2m* which are presented in blue have the same pattern of decreasing Cq values, while the two genes *β-actin* and *CTSD* have random pattern of Cq values and are presented in green.

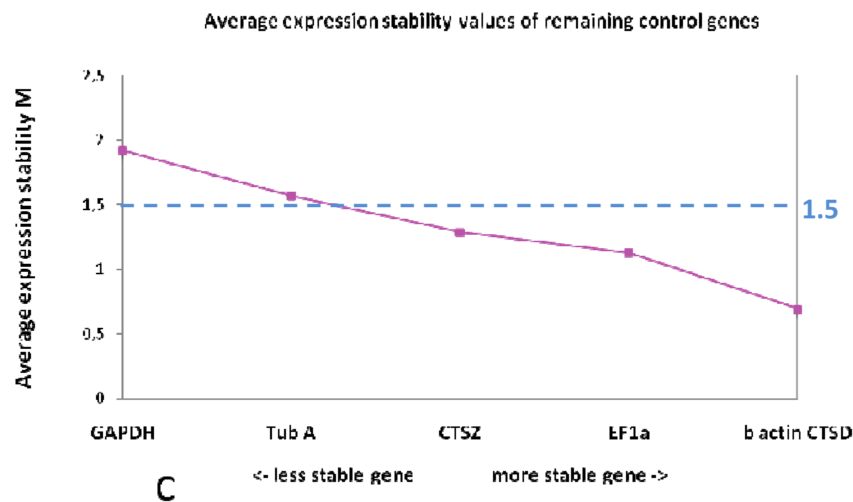
3.2.2.1 GeNorm analysis of Oocytes

The two most stable genes indicated by GeNorm when analyzing all genes were *HPRT1* and *GADPH*, with M values around 0.4 (**Figure 16 (a)**). The next most stable genes were, respectively, *TBP*, *Tub A*, *β-actin* and *CTSD*, all with M values below 1.5. In order of decreasing stability, followed *EF1a*, *B2m* and *CTSZ* all with M values higher than 1.5, with the M value of *CTSZ* of 2. The selection of the optimal number of genes, when analyzing all genes indicated that the inclusion of a 4th gene doesn't have a significant effect on the normalization factor, suggesting that the use of the three most stable genes (*HPRT1*, *GADPH* and *TBP*) is adequate for normalization (**Figure 16 (b)**). This result is not coherent with the observations illustrated in figure 14 and 15, where is shown that these three genes are the ones with higher variation in terms of gene expression, therefore the least stable genes. Nevertheless, it confirms the hypothesis of a bias generated by the genes with similar gene expression pattern. All the genes classified as more stable by geNorm belong to the first category described previously.

A second analysis was performed excluding the genes *HPRT1*, *TBP* and *B2m*, to avoid choosing the genes that have a similar expression variation instead of the most stable expression. In this analysis the pair of genes most stable were *β-actin* and *CTSD* and the least stable genes were *GADPH* and *Tub A* which had the highest M values (**Figure 16 (c)**). In this case, the determination of the optimal number of genes by the program revealed that the cut off value of 0.15 was never achieved (**Figure 16 (d)**) and therefore the best strategy for a reliable normalization is to use the four most stable genes *β-actin*, *CTSD Efa1* and *CTSZ* which had the least pairwise variation value of 0.342.







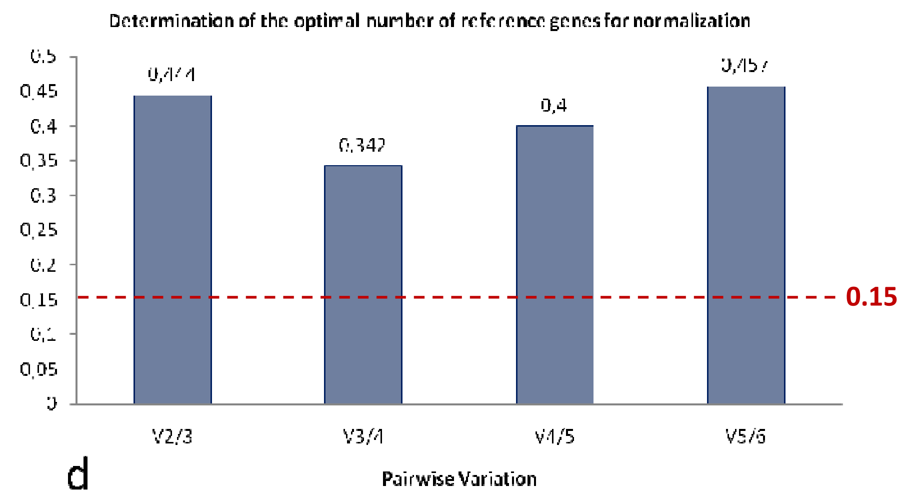


Figure 16 - Selection of the most suitable reference genes for normalization in oocyte samples using *geNorm analysis.* In (**a**) and (**b**) all candidate genes were analyze and in (**c**) and (**d**) the genes *HPRT1*, *TBP* and *B2m* were excluded. In (**a**) and (**c**)**,** the genes expression stability increases from left to right along with the decreasing of **M values** (y-axis). Determinations of the optimal number of reference genes for normalization are represented in (**b**) and (**d**). **Pair wise variation** of normalization factors after successive inclusion of less stable genes determined the optimal number of reference genes. On the left-most side is the pair wise variation when the number of genes is enlarged from 2 to 3 (V2/3). Stepwise inclusion of less stable genes generates the next data points.

3.2.2.2 *NormFinder analysis in Oocytes*

This program indicated *β-actin* as the most stable gene, the only one with a standard deviation bellow 0.5 (**Figure 17 (a)**). The next two most stable genes indicated by Normfinder were *CTSD* and *Tub A* with a standard deviation between 0.5 and 1. The remaining genes had standard deviations above 1.5, with *CTSZ* as the least stable gene.

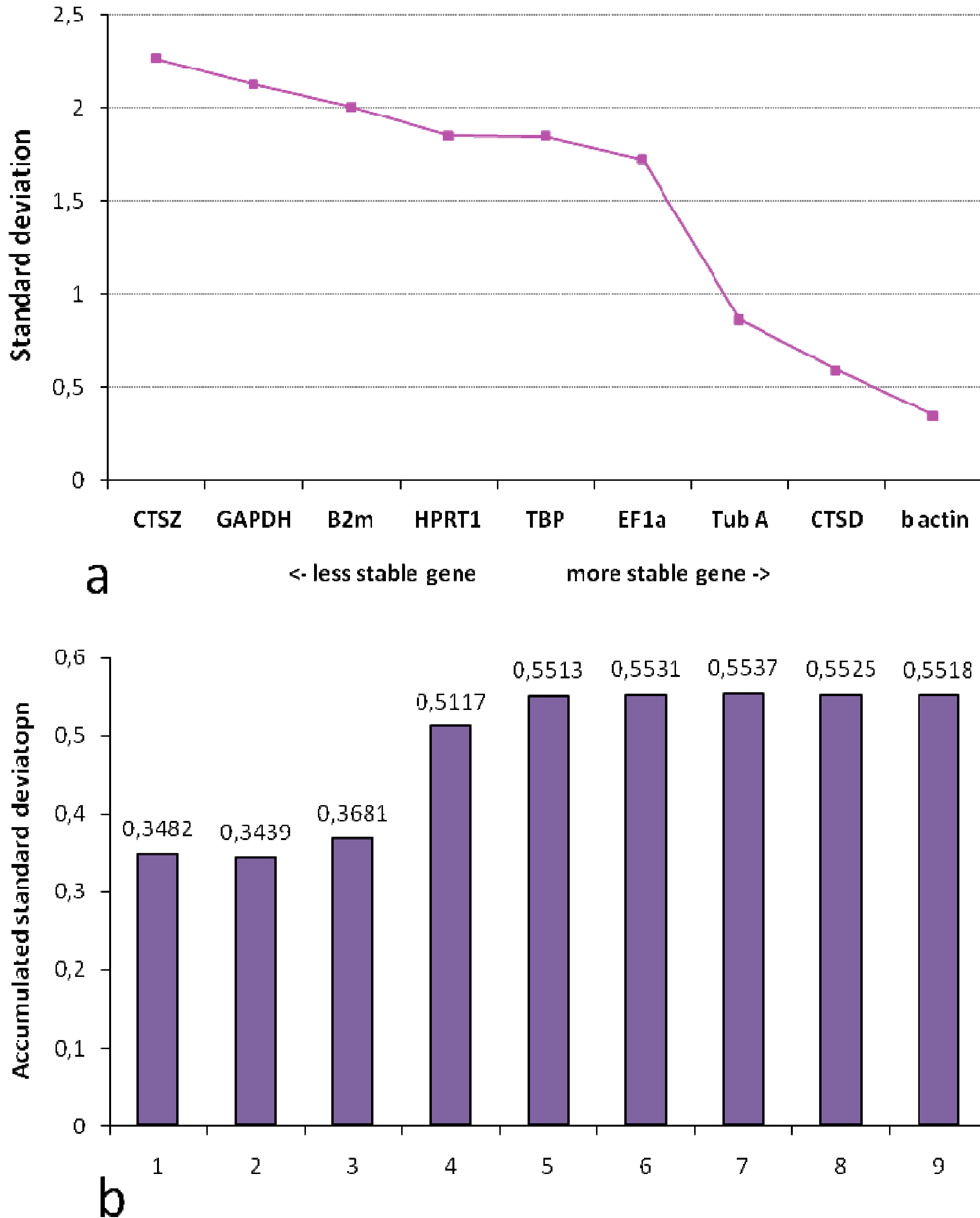


Figure 17 - Selection of the most suitable reference genes for normalization in oocytes using *Normfinder* analysis. In (**a**) it is represented the order of expression stability of all selected genes and in (**b**) it is represented the determination of the optimal number of reference genes for normalization. The accumulated standard deviation is a good indicator of the optimal number of reference genes.

The lowest value of accumulated standard deviation indicates the optimal number of genes to use in this case. The lowest value achieved is when a second gene is added (**Figure 17 (b)**). The value of accumulated standard deviation for the first and second most stable genes is very close, but when the second gene is added this value decreases, suggesting that a second gene could be added for a reliable normalization. The variation of accumulated standard deviation observed determined by this program suggests the use of the most stable gene (*β-actin*) or the two most stable genes (*β-actin* and *CTSD*) for best normalization.

Table 10 summarizes the results obtained which highlight the differences obtained with the two analyses with geNorm. For the reason described above, we will consider the result of the second analysis. The results of geNorm (c) are more in agreement with the analysis of Normfinder when ranking the most stable genes, in both cases *β-actin* and *CTSD* were considered as the more stable genes.

Table 10 - Ranking of the candidate reference genes in *Oreochromis mossambicus* oocytes, according to their stability value using *geNorm* and *Normfinder* analysis. The ranking order (geNorm c) doesn't include the genes *HPRT1*, *TBP* and *B2m*.

	Ranking Order (geNorm a)	Ranking Order (geNorm c)	Ranking Order (Normfinder)
HPRT1	1	-	6
GADPH	1	5	8
TBP	2	-	5
Tub A	3	4	3
β-actin	4	1	1
CTSD	5	1	2
EF1a	6	2	4
B2m	7	-	7
CTSZ	8	3	9

3.2.3 BAMBI and Gremlin expression in Oocytes

To validate the results obtained, we have analyzed the expression the two antagonists of BMPs, *BAMBI* and *Gremlin* during oocyte development. The cDNA used for the qRT-PCR was the same used to test genes for normalization. The results presented in the figure 18 shows the Cq value obtained for the 2 genes during maturation. No normalization was carried for this graph. The amount of poly-A used for the cDNA synthesis was the same and the results reflect the amount of transcripts present in each stage.

The Cq values of *BAMBI* are constant in the first 4 groups (respectively, oogonia, primary oocytes, early vitellogenic oocytes, late vitellogenic oocytes) and increased significantly during the last stage (group 5 - mature oocytes). The Cq for *Gremlin* was constant in the 3 first stages and increased significantly in group 3 and in group 4.

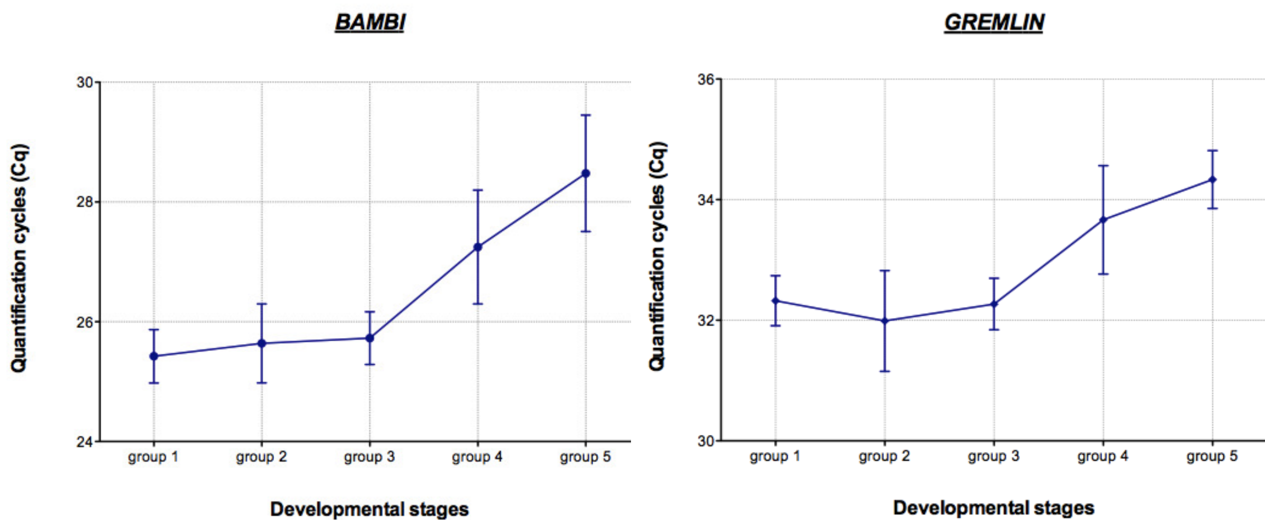


Figure 18 – Individual expression profiles of the genes *BAMBI* and *Gremlin* in oocytes. The quantification cycle (Cq) of each gene was measured between each of the five groups of oocytes analyzed (Group 1 – Oogonia; Group 2 - Primary oocytes; Group 3 - early vitellogenic oocytes ; Group 4 - late vitellogenic oocytes; Group 5 - mature oocytes). The same letter on the top of the bar means that no significant difference was observed. Differences between means were analyzed by one-way ANOVA followed by the Tukey-Kramey multiple comparison test. Significance was set at $P<0.05$.

Four genes (*β-actin*, *CTSD*, *EF1a* and *CTSZ*) have been selected by the geNorm software for the normalization of gene expression during oocyte development. To validate this result, the expression of *BAMBI* and *Gremlin* were normalized against these genes. The normalization results are represented in figure 19 for *BAMBI* and *Gremlin*.

Analyzing these graphs it's possible to observe that for any normalization factor the results show great variations of expression for both genes between the different groups of oocytes. Using the two, three or four most stable genes (NF2, NF3 and NF4) in normalization of *BAMBI* and *Gremlin* results in big differences relative to the their individual expression profiles in figure 18.

Because of the aberrant results given by the normalization with geNorm software, represented in figure 19, we decided to have a different approach by using the classic method of normalization also used by Normfinder software, which is given by the following equation:

$$\textbf{Relative expression} = \frac{\textbf{Concentration of gene of interest}}{\begin{array}{c}\textbf{Geometric mean} \\ (\textbf{concentration of reference gene 1}, \\ \textbf{concentration of reference gene 2}, \dots)\end{array}}$$

This new analysis approach is represented in figure 20 and the results are much more coherent with the concept of normalization which should correct only small variations, especially in this study that every samples were done in triplicates very precisely to reduce at maximum the variation. It's visible for both genes, Bambi and Gremlin, that the normalization by the four genes *β-actin*, *CTSD*, *EF1a* and *CTSZ* give the best and most coherent result, since the differences between before and after normalization are diminished.

To test our reference gene(s), *CTSD* and *β-actin* suggested by Normfinder, were used to normalize *BAMBI* and *Gremlin* expression variation. These genes were used in normalization individually and as a pair of reference genes *CTSD+β-actin* (**Figure 21**). It's possible to see in figure 21 that the normalization with CTSD give the best result for both *BAMBI* and *Gremlin,* because the differences before and after normalization are lower comparing with the normalization only with *β-actin* and with the pair *CTSD+β-actin.* Comparing the results with and without normalization, for *BAMBI* and *Gremlin*, the increase of Cq value in the last two groups seen in figure 21, were attenuated after normalization and that in general the expression of these two genes was constant during oocyte development, especially evident for Gremlin when normalized only with *CTSD*.

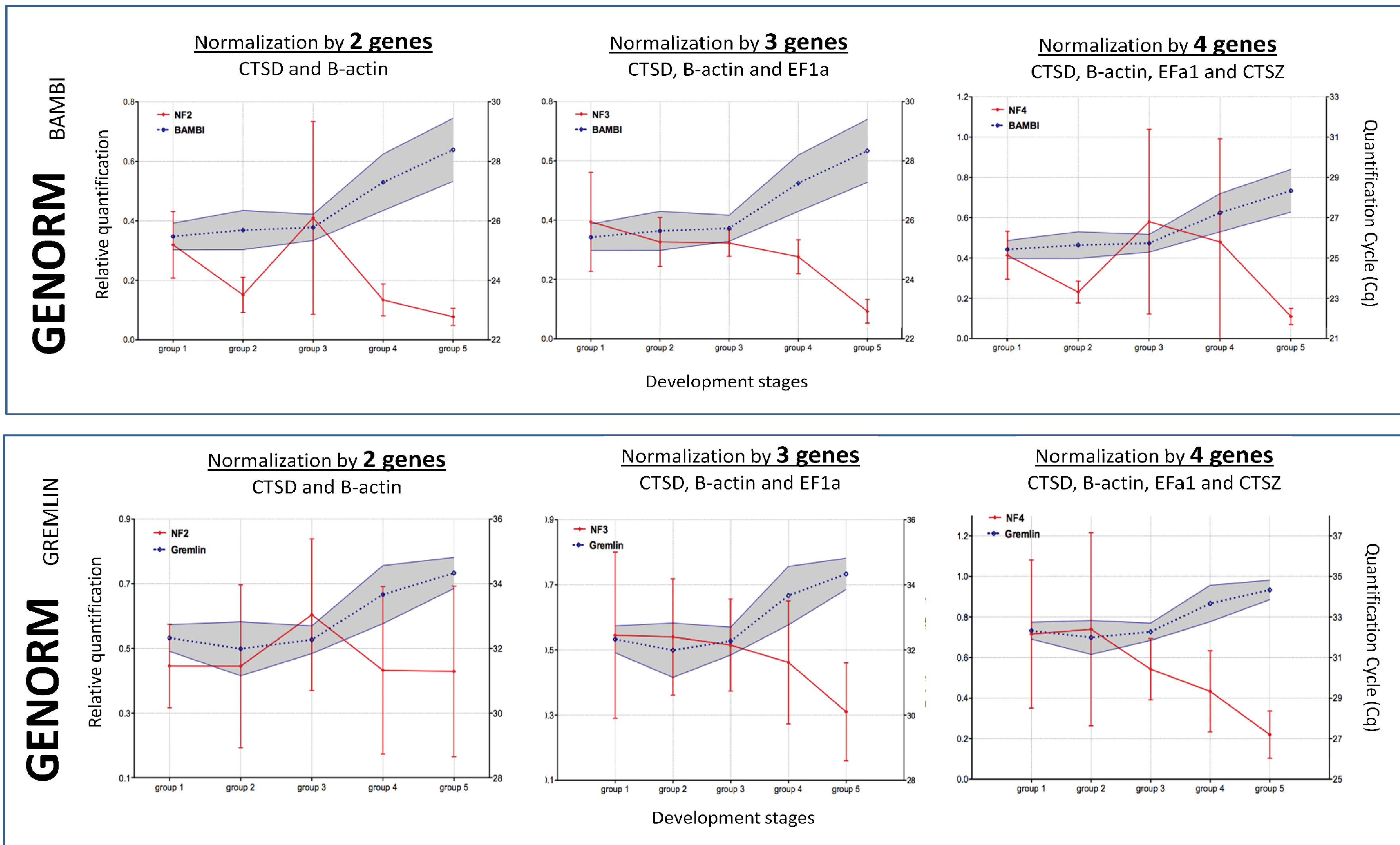


Figure 19 – Normalization of *BAMBI* and *Gremlin* expressions in oocytes against *β-actin*, *CTSD*, *Efa1* and *CTSZ*. The set of three reference genes (*β-actin, CTSD, Efa1* and *CTSZ*) suggested by geNorm software were used. The five groups of oocytes analyzed (Group 1 – Oogonia; Group 2 - Primary oocytes; Group 3 - early vitellogenic oocytes; Group 4 - late vitellogenic oocytes; Group 5 - mature oocytes).

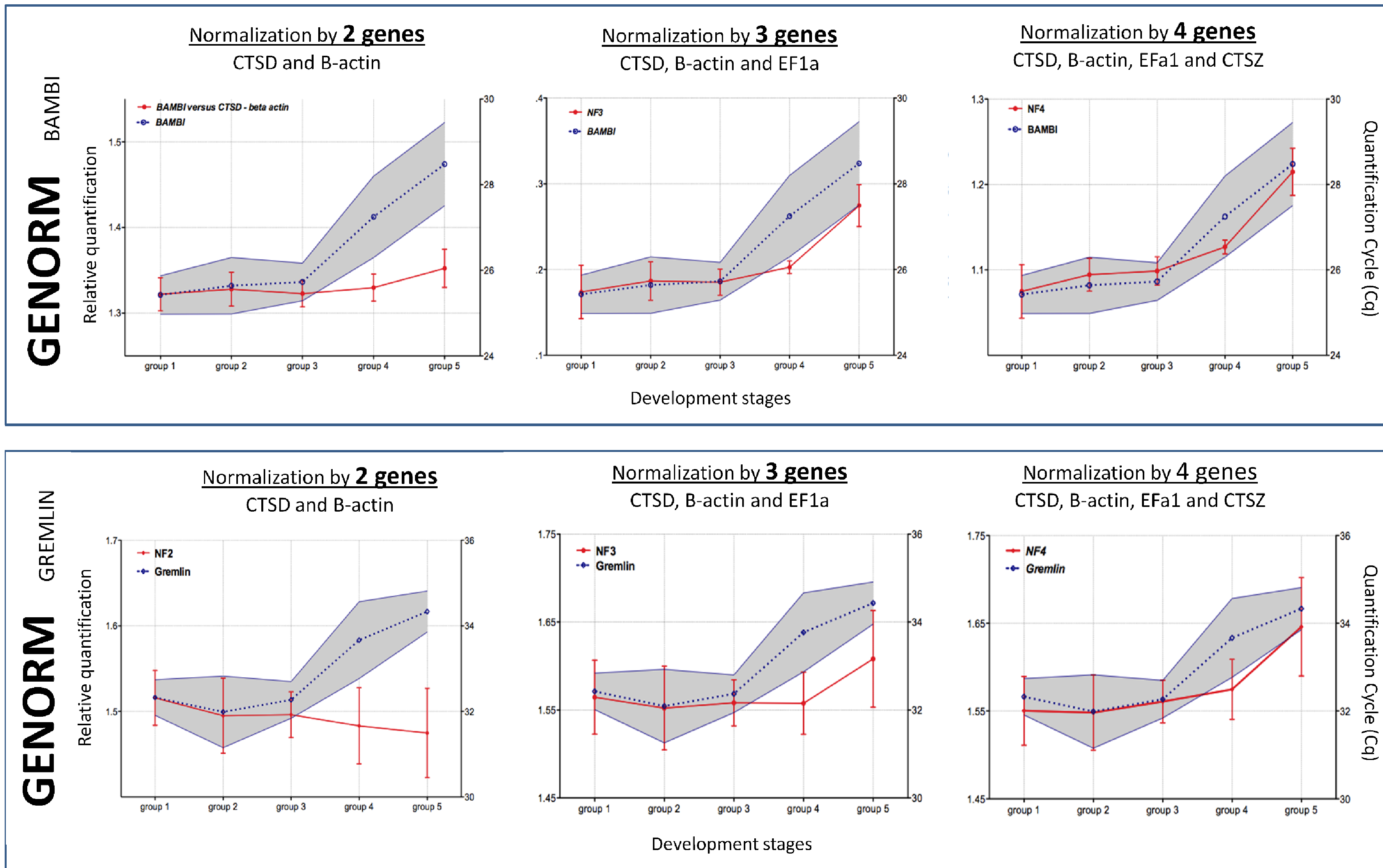


Figure 20 – Normalization of *BAMBI* and *Gremlin* expressions in oocytes against *β-actin*, *CTSD*, *Efa1* and *CTSZ* using Normfinder method. The set of three reference genes (*β-actin, CTSD, Efa1* and *CTSZ*) suggested by geNorm software were used. The five groups of oocytes analyzed (Group 1 – Oogonia; Group 2 - Primary oocytes; Group 3 - early vitellogenic oocytes; Group 4 - late vitellogenic oocytes; Group 5 - mature oocytes).

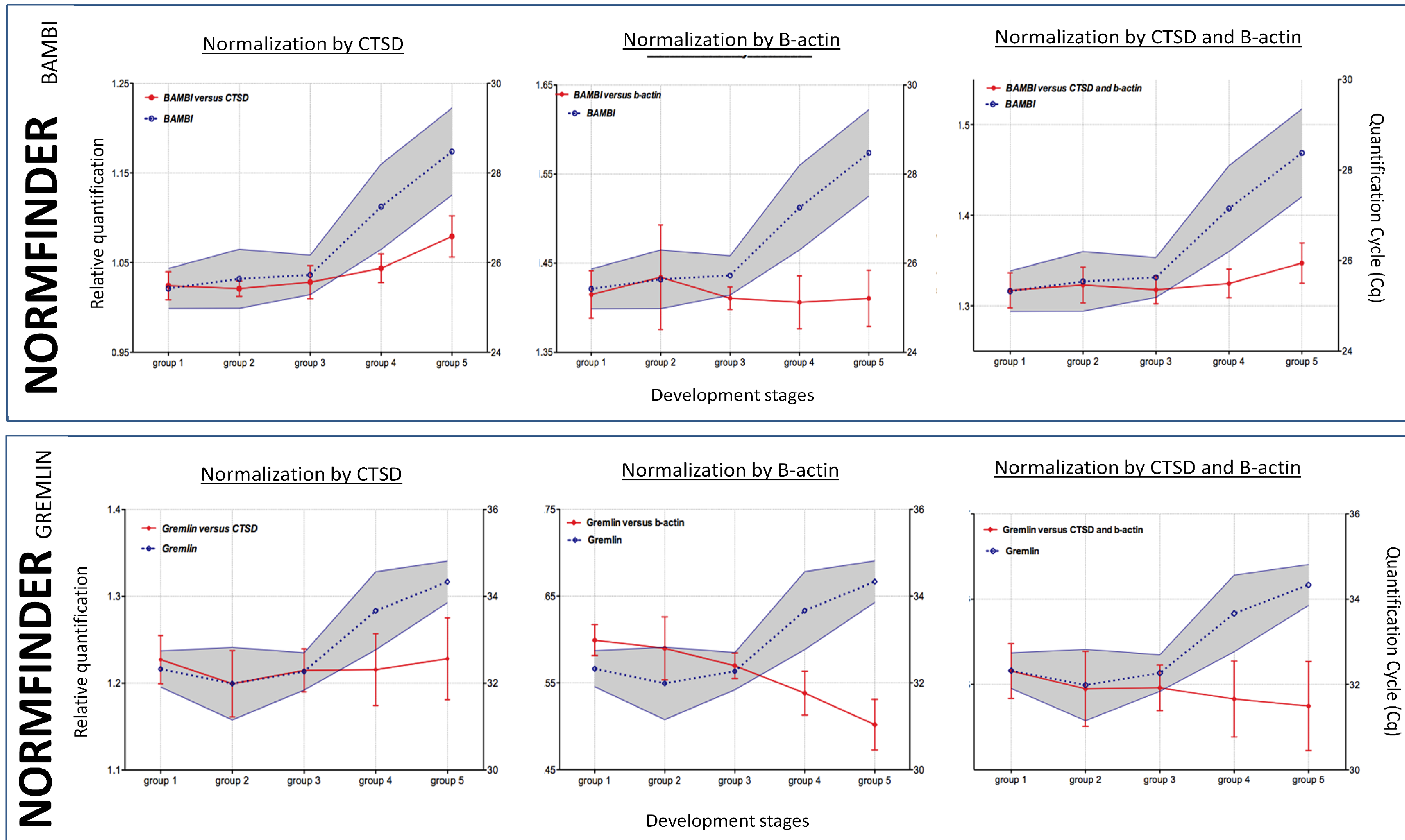


Figure 21 – Normalization of the genes *BAMBI* and *Gremlin* expressions in oocytes with *β-actin* and *CTSD*. The quantification cycle (Cq) of each gene was measured for each of five groups of oocytes (Group 1 – Oogonia; Group 2 - Primary oocytes; Group 3 - early vitellogenic oocytes; Group 4 - late vitellogenic oocytes; Group 5 - mature oocytes).

4. DISCUSSION

In this study the suitability of specific genes as reference genes in the analysis of relative mRNA expression by real-time quantitative PCR was examined, including a calculation of the minimal set of genes needed for 13 different somatic tissues and for different stages of oocytes development. We included reference genes of different functional classes like ribosomal genes, and genes involved in anabolic and catabolic pathways. The optimal reference genes selected by a program can only be as good as their candidates, so the choosing of potential reference genes, that were predicted stable for our case, were based on their “housekeeping” role or by evidence of their stability in previous studies.

Producing a valid and reliable qRT-PCR assay is not an easy task and involves a number of strict requirements, including specific procedures recently outlined in the Minimum Information for Publication of Quantitative Real-Time PCR Experiments (MIQE) guidelines [55]. Strict quality controls were performed in the present study to obtain reliable data, including the RNA integrity, RNA purity, RNA concentration determination, and primer selection. Also the qRT-PCR performance was validated in this study since the efficiencies of our reactions ranged from intervals that are in the accepted range [26].

The data given by the qRT-PCR reactions were first compared by doing overall and individual expression profiles of the candidate reference genes and then their stability were analyzed by using software packages designed for this purpose. Finally these sets of reference genes appointed by the software’s in the case of oocytes development were tested in normalization of the two antagonists of BMPs *BAMBI* and *Gremlin*.

When analyzing the Cq data given by the qRT-PCR reactions in somatic tissues it was observed that *GADPH* was the gene with the highest variation, and that *18S* and *TBP* had the lowest. It was also observed that *Efa1* and *Tub A* had the most similar variation and transcription level. These observations suggest that *18S* and *TBP* are the best genes to use for normalization in these tissues and that *GADPH* is the worst. Also that for oocytes the best reference gene to use could be *Efa1*, *β-actin* or *CTSD*, yet these analyses are based on the Cq values of the genes without normalization. As we have seen in the section 1, selecting suitable

reference genes can be a circular problem, because these genes also need normalization. To normalize our genes, we need to have already the reference genes, although, that's precisely what we don't have and we are trying to get. That's why, to analyze the Cq values of our set of potential reference genes and to find out which are the most stable, it was used geNorm and Normfinder softwares. These programs help us to circumvent the problem by comparing the gene expression among our set of genes and with that indicating the most stables.

When analyzing all the candidate genes in somatic tissues and in oocytes by geNorm, we detected that the program was choosing genes as the most stable, however, when analyzing their individual and overall expression profiles they didn't seem to be. In somatic tissues, the program chose *Tub A* and *CTSZ* as the most stable genes when in overall and individual expression profiles they didn't seem to be the genes with less variability but only the genes with more identical variation. For oocytes analysis this factor is even clearer, because the program appointed *HPRT1* and *GADPH* as the most stable pair of genes, when it was already observed in their individual and overall expression profiles, that this two genes were the ones with the highest but similar variation.

Anderson *et al* [47] demonstrated that the pair-wise comparison approach done by geNorm, tends to select those genes with the highest degree of similarity of the expression profile and that sensitivity to co-regulation is a major weakness of this program. Other studies also refer this problem and took measures to avoid the inclusion of the genes that present this behavior [50] [44, 46 in 50]. To avoid choosing less suitable reference genes due to this problem a second analysis was performed in both cases, excluding those genes that were making pairs of similarity.

The geNorm analysis in adult tissues excluding *Tub A* gave the ranking order from the most stable to the least stable gene as follows: *Efa1*-*18S*; *B2m*; *CTSD*; *β-actin*; *CTSZ*; *TBP*; *HPRT1*; *GADPH*. For Normfinder analysis in tissues the ranking results were, from the most to the less stable gene, as follows: *18S*; *TBP*; *HPRT1*; *Efa1*; *Tub A*; *CTSZ*; *CTSD*; *B2m*; *β-actin*; *GADPH*.

The genes *Efa1* and *18S* were appointed by geNorm as the most stable pair of genes in somatic tissues. Other studies of validation of reference genes in Zebrafish tissues also appointed these two genes as the most stable when

analyzing with this program [48] [45], and in [45]. *B2m* was also ranked in third position in terms of gene stability as in our case. The gene *TBP* was also suggested as reference gene when he was ranked as the third most stable gene with Normfinder in studies of Rhesus monkey tissue [51], but when analyzing with geNorm, it was ranked in sixth position as in our case.

The gene *18S* is one of the most commonly used reference gene, and it was considered to be an ideal internal control in qRT-PCR analysis [53 in 52]. However, there are several arguments against the use of *18S* gene as an internal control. One of the major limitations is that an imbalance of rRNA and mRNA fractions can occur between samples, which make *18S* RNA less suitable as a normalizer in calculating relative mRNA levels [1-3 in 48]. Another drawback is the much greater level of expression relative to target mRNA, what makes it difficult to subtract the baseline value in qRT-PCR data analysis accurately [48]. So it's important to be caution when using *18S* rRNA as a reference gene.

In the geNorm analysis of oocytes, excluding *B2m*, *TBP* and *GADPH* the ranking order from the most stable to the least stable gene was as follows: *β-actin*-*CTSD*; *Efa1*; *CTSZ*; *Tub A*; *HPRT1*. With Normfinder software the ranking order was: *β-actin*; *CTSD*; *Tub A*; *Efa1*; *TBP*; *HPRT1*; *B2m*; *GADPH*; *CTSZ*.

In a study of zebrafish [48] when analyzing embryonic development the gene *β-actin* was also ranked in first position by geNorm and *Efa1* as third. In fact *β-actin* has been one of the most commonly used reference gene for normalizing qRT-PCR data in Zebrafish [48] and other fish studies [54]. Another study of early stage of fish development, more precisely in larvae Senegalese sole and Atlantic halibut [1], suggested the use of *β-actin* and *Efa1* as reference genes since they were ranked in the top positions of stability with geNorm and Normfinder softwares. In the case of study the Atlantic Halibut in different stages of embryonic development, geNorm also identified *β -actin* as one of the most stable genes, ranked as being part of the pair of genes most stable [46].

The gene *CTSD* was appointed in both analysis of oocytes as the second most stable gene. This gene is not a typical reference gene, yet it was reported that it was not differentially expressed across oocyte primary and secondary growth in coho salmon [52]. Also in this study, *CTSZ* had the same result as well as in *Fundulus heteroclitus* study [48 in 52], yet in our study this gene was one of the

least stable across oocytes development. Catephesins are lysosomal enzymes that in the oocyte are responsible for proteolytically cleaving vitellogenin into its constituent yolk proteins and play a role in oocyte reabsortion during atresia [2, 44, 45 in 52]. This role of catephesins and the reporting of previous studies that *CTSD* and *CTSZ* were not differentially expressed in different stages of oocyte growth led us to include them as potential reference genes.

All analysis on adult tissues and oocytes, with both programs, ranked *GADPH* as the least stable gene and it was already seen that it was the gene with the highest expression variation in the individual and overall expressions profiles. The differential regulation of *GADPH* is probably connected to its direct role in S phase-dependent histone H2B transcription [44]. This gene was already reported as one of the least stable genes when analyzing reference genes with geNorm software in Zebra fish tissues [45] and in different embryonic development stages of this fish [45], Atlantic Halibut [46] [1], larvae Senegalense sole[1], wild populations of Atlantic Cod [49] and in chicory [50].

To choose the optimal number of reference genes we excluded the first analysis of geNorm due to the aforementioned reasons and took in consideration the most reliable second analysis in the two cases of study. For adult tissues the program suggests the use of the set of three most stable genes for best normalization (*18S*, *Efa1* and *B2m*) and for oocytes the program suggests the four most stable genes, but in this case *B-actin*, *CTSD, Efa1* and *CTSZ*.

With the Normfinder software it was also possible to choose the optimal number of genes since the accumulated standard deviation is a good indicator. The analysis of this program for somatic tissues suggested the use of the two most stable genes *18S* and *TBP*. For the oocytes, Normfinder indicates *β-actin* and *CTSD* as the most stable genes, and that the optimal number of genes is one or two, so *β-actin* or the pair *β-actin-CTSD* should be used as the reference gene(s) in normalization of gene expression in oocytes.

When analyzing several publications it is difficult to compare results since almost studies never analysis the exact same genes or tissues and that different species, different experimental conditions and different stages of development can influence the gene expression. There are also different approaches when using this two software's and the interpretation sometimes is not the most coherent. For

example when the co-regulated genes are included in the geNorm analysis can have as consequence the choosing of the wrong reference genes. Also when choosing the optimal number of reference genes there are some incongruence's, as in the study of larvae flatfish [1] where they refer to V3/4 as the inclusion of a third gene instead of a fourth. It's also not so clear which cut off value should be used when determining the optimal number of reference genes.

In the study of reference genes in wild populations of Atlantic cod [49] the genes *β-actin* and *Efa1* were considered not suitable as reference genes since they were ranked poorly against all the genes evaluated in this study. Yet, it's visible in their results that for *Efa1* the M values achieved with geNorm in the different tissues were never above 0.6 and that for *β-actin* the values were all less than 1.2. These two M values are considered as stable by the program, but because they were ranked poorly they were considered as not suitable reference genes by the researchers. This demonstrates that there is some relativity when using these programs by the researchers.

It's important to be careful when using this programs and it's crucial to understand the differences between them so we can choose the most adequate or to find that they can be used also as complements. The major differences between the programs is that geNorm will find the two genes whose expression ratios shows least variation with the other genes tested, Normfinder will find both the single gene with most stable expression as well as the best pair of genes with combined most stable expression.

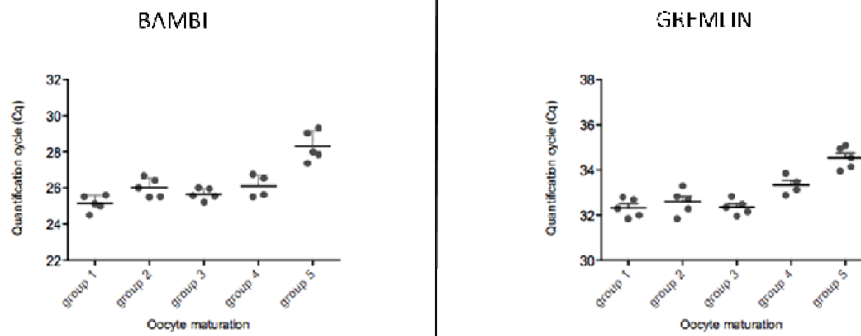
Other issue that is important to refer is that most of the studies regarding the selection of reference genes only mention which genes are considered best by the software without validating if the genes are 'good enough'. In this study, we not only identified the most stable genes true their expression analysis by the two programs but we also test them. To do so and to try to compare the reference genes choose by both programs it was normalized the expression of *BAMBI* and *Gremlin* genes in oocytes. It was verified that the results are very different and that in the case of the genes appointed by Normfinder, normalizing with *CTSD* alone give the best result since the differences before and after normalization are lower comparing with the normalization only with *β-actin* and with the pair *CTSD+β-actin.* After testing we can conclude that, despite Normfinder suggested two genes

to use in normalization, in practice only with one gene a better result is achieve. However it's been clarified that the normalization with 1 gene is risky and sometimes 2 genes are not enough, turning the results given by Normfinder somehow fragile.

In the case of geNorm software we have two different normalization approaches, for the first normalization with the set of genes appointed by geNorm (*β-actin*, *CTSD, Efa1* and *CTSZ*) the results were aberrant for both genes. In fact the results obtained after normalization showed a big alternate variation of both genes along the maturation of oocytes, which contrast significantly with the relative expression of *BAMBI* and *Gremlin* before the normalization.

The differences of normalization by the two programs are evident specially when comparing the results of normalization with the same genes *β-actin* and *CTSD* for the NF2. The reason for this can be explained by the different approach that the programs follow when normalizing genes. To better understand this we will describe the steps that lead to the normalization of *BAMBI* and *Gremlin* with the help of the representation in figure 22. In geNorm, the Cq values were first transform into relative quantification data using the deltaCt method with the formula $Q = E^{(minCt - sampleCt)}$, where E is the amplification efficiency (**Figure 22**). Then the calculation of the normalized genes of interest (GOI) expression levels was performed by dividing the GOI quantities by the appropriate normalization factor for x genes (NFx), producing the final results. It's visible on this figure that in the first transformation to relative quantities the graphs are already very different from the Cq values. This suggests that the variation in expression levels due to normalization could be reflected by variations in the Ct values of the reference genes.

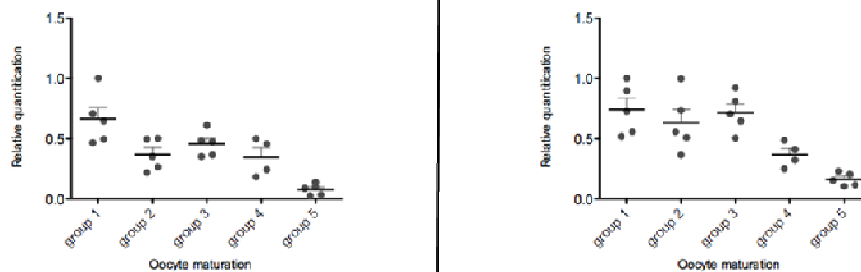
In NormFinder the data is applied in logarithmic scale such as Cq values or fold changes contrarily to geNorm that uses the deltaCt method. If the data is on linear scale, GenEx, an application of Normfinder, will convert the data to logarithmic scale only for the analysis.



⇨ Transform Cq values into relative quantification data using the deltaCT method.

$$Q = E^{(minCt - sampleCt)}$$

Q = sample quantity relative to sample with highest expression
E = amplification efficiency (2 = 100%)
minCt = lowest Ct value = Ct value of sample with highest expression



⇨ Calculation of normalized GOI expression levels by dividing the GOI quantities by the appropriate normalization factor for x genes (NFx)

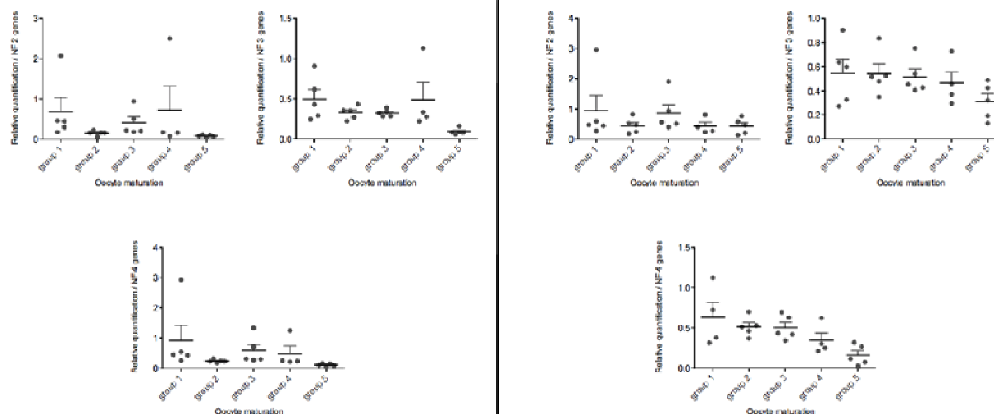


Figure 22– First procedures for normalization of the genes *BAMBI* and *Gremlin* in oocytes with *β-actin*, *CTSD* and *Efa1*. The quantification cycle (Cq) of each gene was measured between each of the five groups of oocytes analyzed (Group 1 – Oogonia; Group 2 - Primary oocytes; Group 3 - early vittelogenic oocytes ; Group 4 - late vittelogenic oocytes; Group 5 - mature oocytes). Cq values were first transform into relative quantification data using the deltaCt method and then divided by the appropriate normalization factor.

This different data treatment of the two programs produces different results of normalization. In the case of geNorm the results are not coherent therefore we choose to use the classical equation to transform Cq values.

In the second normalization approach, we had much more satisfactory results since the normalization in this case didn't affect too much the expression patterns of these two genes, being coherent with the principle that normalization will correct small variations inherited to the experimental design of qRT-PCR. We can conclude that the using of the four most stable genes (*CTSD, B-actin, EFa1* and *CTSZ*) suggested by the software give the best results, and that the using of multiple genes in normalization produced last variation than using individual genes. In fact, Anderson *et al.,* [47] reported that the variation in the average of multiple genes is smaller than the variation in individual gene.

The intent of this study was to provide a database that helps Mozambique tilapia researchers to identify a shortlist of candidate reference genes for specific experiments. In fact, this report represents a clear enrichment for the genomic toolbox of this species.

The results of both programs are not consensual and appointed to use different number of reference genes, explained by the differences of approaches that both have. What the programs had in common is that both appointed to *18S* as one of the most stable gene in somatic tissues, however there are some reservations reported aforementioned when using ribosomal RNA. Both programs also indicated *β-actin* as one of the most stable gene in oocytes.

Although it has been a relatively popular reference gene in fish research, *GADPH* clearly has large variability in its expression in both cases tested in our study and so would not be recommended for normalization. Combining the results of both programs the genes with the least variability across somatic tissues were *18S*, *Efa1*, *B2m* and *TBP* while across all five groups of oocytes in development it were *β-actin*, *Efa1* and *CTSD*.

It was also highlighted that geNorm produced a most suitable normalization when testing *Gremlin* and *BAMBI* gene expression in oocytes, revealing that this program, in this case, was more efficient to select the proper reference genes.

5. REFERENCES

[1] Infante C., Matsuoka M.P., Asensio E., Canavate J.P., Reith M., Manchado M., 2008. Selection of housekeeping genes for gene expression studies in larvae from flatfish using real-time PCR. *BMC Mol. Bio.* 9:28.

[2] Wong, M.L. and Medrano, J., 2005. Real-Time PCR for mRNA quantification. BioTechniques 39(1):75-85.

[3] Jo Vandesompele, Katleen De Preter, Filip Pattyn et al., 2002. Accurate normalization of realtime quantitative RT-PCR data by geometric averaging of multiple internal control genes. Genome Biology. 3(7): research 0034.1-0034.11.

[4] Ambion applied biosystems RT-PCR: The Basics *by Subbu Dharmaraj, MS* http://www.ambion.com/techlib/basics/rtpcr/index.html. (last accessed 29 September 2010).

[5] Bustin S.A., 2002. Quantification of mRNA using real-time reverse transcription (PCR QPCR): trends and problems. *J Mol Endocrinol*. 29:23-39.

[6] Radonić, A., Thulke, S., Mackay, I.M., Landt, O., Siegert, W., Nitsche, A., 2004. Guideline to reference gene selection for quantitative real-time PCR. Biochem. Biophys. Res. Commun. 313, 856–862.

[7] Thellin, O., Zorzi, W., Lakaye, B., De Borman, B., Coumans, B., Hennen, G., Grisar, T., Igout, A., Heinen, E., 1999. Housekeeping genes as internal standards: use and limits. *J. Biotechnol*. 75: 291–295.

[8] S. Derveaux, J. Vandesompele, J. Hellemans, 2009. How to do Successful Gene Expression Analysis Using Real-Time PCR, Methods. doi: 10.1016/j.ymeth.2009.11.001.

[9] Tingaud-Sequeira A., Chauvigné F., Lozano J., Agulleiro M.J., Asensio E., Cerdà J., 2009. New insights into molecular pathways associated with flatfish ovarian development and atresia revealed by transcriptional analysis. *BMC Genomics* 10:434.

[10] McCurley, A. T. and Callard, G. V., 2008. Characterization of housekeeping genes in zebrafish: male-female differences and effects of tissue type, developmental stage and chemical treatment. *BMC Mol. Biol.* 9:10.

[11] Lubzens, E., et al., 2009. Oogenesis in teleosts: How fish eggs are formed. Gen. Comp. Endocrinol. doi:10.1016/j.ygcen.2009.05.022.

[12] Jalabert, B., 2005. Particularities of reproduction and oogenesis in teleost fish compared to mammals. Reprod. Nutr. Dev. 45, 261–279.

[13] Tyler, C.R., Sumpter, J.P., 1996. Oocyte growth and development in teleosts. Rev. Fish Biol. Fish. 6, 287–318.

[14] Canonico, G. C., A. Arthington, J. K. McCrary, and M. L. Thieme., 2005. The effects of introduced tilapias on native biodiversity. Aquatic Conservation – Marine and Freshwater Ecosystems 15:463–83.

[15] Baroiller, J.F., et al., 2009. Tilapia sex determination: Where temperature and genetics meet, Comp. Biochem. Physiol. A. doi:10.1016/j.cbpa.2008.11.018.

[16] Chapman, F.A., Reviewed, 2006. Culture of Hybrid Tilapia: A Reference Profile, Circular 1051, Department of Fisheries and Aquatic Sciences. Institute of Food and Agricultural Sciences, University of Florida.

[17] Bustin, S.A., Benes, V., Nolan, T., Plaff, M.W., 2005. Quantitative real-time RT-PCR- a perspective. J Mol Endocrinol. 34:597-601.

[18] Van Der Velden, V.H.J., Hochhaus, A., Cazzaniga, G., Szczepañski, T., Gabert, J., vanDongen, J.J.M., 2003. Detection of minimal residual disease in hematologic malignancies by real-time quantitative PCR: principles, approaches, and laboratory aspects. Leukemia 17: 1013–1034.

[19] Glick, B.R. and J.J. Pasternak, 2003. Molecular Biotechnology: Principles and Applications of Recombinant DNA. 3rd ed. : ASM Press. 760.

[20 in 11] Wallace, R.A., Selman, K., 1990. Ultrastructural aspects of oogenesis and oocyte growth in fish and amphibians. J. Electron. Microsc. Tech. 16, 175–201.

[21 in 11] Le Menn, F., Cerdà, J., Babin, P.J., 2007. Ultrastructural aspects of the ontogeny and differentiation of ray-finned fish ovarian follicles. In: Babin, P.J. (Ed.), The Fish Oocyte: From Basic Studies to Biotechnological Application. Springer, The Netherlands, pp. 1–37.

[22 in 11] Nagahama, Y., Yamashita, M., 2008. Regulation of oocyte maturation in fish. Dev. Growth Diff. 50, S195–S219.

[23] Pals, G., *et al.,* 1999. A rapid and sensitive approach to mutation detection using real-time polymerase chain reaction and melting curve analyses, using BRCA1 as an example. *Mol Diagn,* 4, 241-246.

[24] Kubista, M., J.M. Andrade and M. Bengtsson *et al.,* 2006. The real-time polymerase chain reaction. Mol. Aspects Med., 27: 95-125.

[25] Suwa, K., Yamashita, M., 2007. Regulatory mechanisms of oocyte maturation and ovulation. In: Babin, P.J., Cerdà, J., Lubzens, E. (Eds.), The Fish Oocyte: From Basic Studies to Biotechnological Applications. Springer, Dordrecht, pp. 323–347.

[26] qPCR Technical Guide Sigma Aldrich http://www.sigmaaldrich.com/etc/medialib/docs/Sigma/General_Information/qpcr_technical_guide.Par.0001.File.tmp/qpcr_technical_guide.pdf (last accessed 29 September 2010). http://www.sigmaaldrich.com/portugal.html (last accessed 29 September 2010).

[27] Real-time PCR. Tevfik Dorak-Medical Department, University of Newcastle upon Tyne, Newcastle, UK **Publish Date:** 27th April 2006 **Imprint:** Taylor & Francis **Pages:** 350 pages. http://dorakmt.tripod.com/genetics/realtime.html (last accessed 29 September 2010).

[28] Redding, J.M., Patiño, R., 2000. Reproductive systems: gross functional anatomy. Bath Academic Press, 261-267 pp.

[29] FAO 2009. The State Of World Fisheries And Aquaculture 2008. FAO Fisheries and Aquaculture Department. Rome. Italy.

[30] Vandesompele, J., Kubista, M., Pfaffl, M. W., 2009. Reference Gene Validation Software for Improved Normalization. Chapter 4 in Real-Time PCR: Current Technology and Applications, Publisher: Caister Academic Press. Editor: Julie Logan, Kirstin Edwards and Nick Saunders ISBN: 978-1-904455-39-4.

[31] Stahlberg, A., Hakansson, J., Xian, X., Semb, H., and Kubista, M., 2004. Properties of the reverse transcription reaction in mRNA quantification. Clin Chem *50*, 509-515.

[32] Clelland, E.S., Tan, Q., Balofsky, A., Lacivita, R., Peng, C., (2007). Inhibition of premature oocyte maturation: a role for bone morphogenetic protein 15 in zebrafish ovarian follicles. Endocrinology 148, 5451-5458.

[33] Chenna,R. *et al*., 2003. Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res.*, 31, 3497–3500.

[34] Tamura, K., Dudley, J., Nei, M. & Kumar, S., 2007. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Molecular Biology and Evolution* 24:1596-1599.

[35] Mikeska, T. and Dobrovic, A., 2009. Validation of a primer optimisation matrix to improve the performance of reverse transcription – quantitative real-time PCR assays. BMC Research Notes 2; 112.

[36] Saitou, N. & Nei, M., 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* 4:406-425.

[37] Felsenstein, J., 1985. Confidence limits on phylogenies: An approach using the bootstrap. *Evolution* 39:783-791.

[38] Zuckerkandl, E. & Pauling, L., 1965. Evolutionary divergence and convergence in proteins, pp. 97-166 in *Evolving Genes and Proteins*, edited by V. Bryson and H.J. Vogel. Academic Press, New York.

[39] Tamura, K., Dudley, J., Nei, M. & Kumar, S., 2007. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Molecular Biology and Evolution* 24:1596-1599.

[40] Zhang, Z., Hu, J., 2007. Development and validation of endogenous reference genes for expression profiling of medaka (*Oryzias latipes*) exposed to endocrine disrupting chemicals by quantitative real-time RT-PCR. *Toxicol Sci.* 95:3563-3568.

[41] Yu, S.H., Chen, X.H., Ao, J.Q., 2009. Molecular characterization and expression analysis of b2-microglobulin in large yellow croaker *Pseudosciaena crocea*. Molecular Biology Reports 36:1715–1723.

[42] Determinants of RNA Integrity and Purity (AMBION – Applied Biosystems). http://www.ambion.com/techlib/tn/112/10.html - http://www.ambion.com/ (last accessed 29 September 2010).

[43] Bench Guide QIAGEN http://www.qiagen.com/selectlocation.aspx?redirect=%2fliterature%2frender.aspx%3fid%3d23576 (last accessed 29 September 2010). http://www.qiagen.com/default.aspx (last accessed 29 September 2010).

[44] Zheng, L., Roeder, R.G., Luo, Y., 2003. S phase activation of the histone H2B promoter by OCA-S, a coactivator complex that contains GAPDH as a key component. Cell 114, 255–266.

[45] McCurley, A.T. and Callard, G.V., 2008. Characterization of housekeeping genes in zebrafish: male-female differences and effects of tissue type, developmental stage and chemical treatment. *B.M.C. Mol. Biol.,* 9, 102.

[46] Fernandes, J.M.O., Mommens, M., Hagen, Ø., Babiak, I., Solberg, C., 2008. Selection of suitable reference genes for real-time PCR studies of Atlantic halibut development. Comp. Biochem. Physiol., B 150, 23–32.

[47] Andersen C.L., Ledet-Jensen J., Ørntoft T., 2004. Normalization of real-time quantitative RT-PCR data: a model based variance estimation approach to identify genes suited for normalization - applied to bladder- and colon-cancer data-sets. Cancer Research. (64): 5245-5250.

[48] Tang, R., Dodd, A., Lai, D., McNabb, W., Love, D., 2007. Validation of Zebrafish (Danio rerio) Reference Genes for Quantitative Real-time RT-PCR Normalization. Acta Biochim Biophys Sinica, 39:384-390.

[49] Olsvik, P.A., Søfteland, L., Lie, K.K., 2008. Selection of reference genes for qRT-PCR examination of wild populations of Atlantic cod Gadus morhua. BMC Research Notes,I:47.

[50] Maroufi, A., Van Bockstaele, E., De Loose, M., 2010. Validation of reference genes for gene expression analysis in chicory (Cichorium intybus) using quantitative real-time PCR. BMC Molecular Biology, vol. 11, no. 15.

[51] Ahn, K., Huh, J.W., Park, S.J., Kim, D.S., Ha, H.S., Kim, Y.J., Lee, J.R., Chang, K.T., Kim, H.S., 2008. Selection of internal reference genes for SYBR green qRT-PCR studies of rhesus monkey (Macaca mulatta) tissues. BMC Mol Biol. 9:78.

[52] Tong, Z.; Gao, Z.; Wang, F.; Zhou, J. and Zhang, Z., 2009. Selection of reliable reference genes for gene expression studies in peach using real-time PCR. BMC Molecular Biology, vol. 10, no. 71.

[53] GeNorm manual http://medgen.ugent.be/~jvdesomp/genorm/geNorm_manual.pdf (last accessed 29 September 2010).

[54] Small, B.C., Murdock, C.A., Bilodeau-Bourgeois, A.L., Peterson, B.C., Waldbieser, G.C., 2008. Stability of reference genes for real-time PCR analyses in channel catfish (Ictalurus punctatus) tissues under varying physiological conditions. Comparative Biochemistry and Physiology B-Biochemistry & Molecular Biology, 151:296-304.

[55] Bustin S.A., et al., 2009. The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments. Clin. Chem. 55(4):611-22.

ANNEX 1: DNA ladder

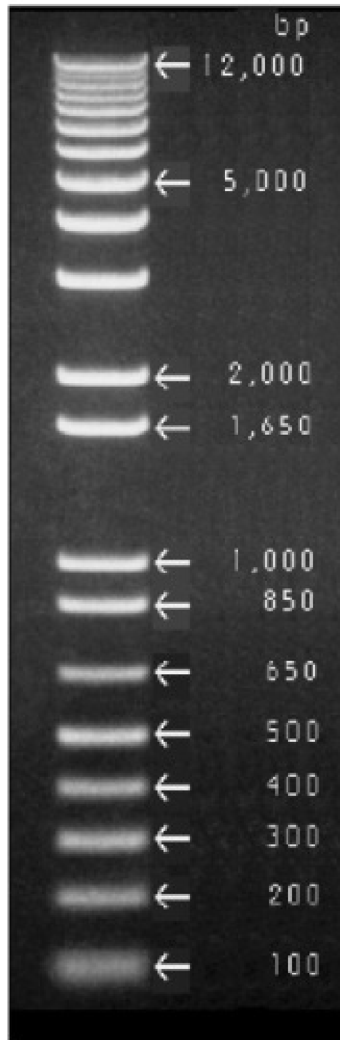


Figure A1 - 1 kb plus DNA ladder
0.7 µg ran in a 0.9% agarose gel stained with ethidium bromide
The numbers in the right size of the ladder are the size in base pair (bp) of the bands
(From Invitrogen life technologies)

ANNEX 2: p-GEM-T Easy vector restriction map

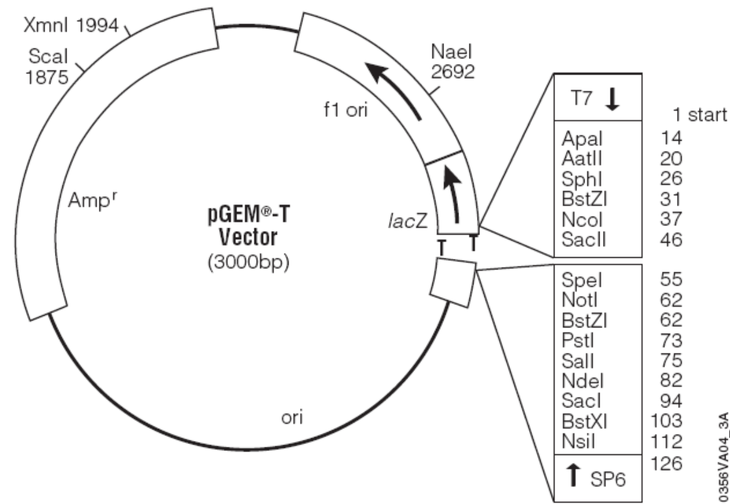


Figure A 2: p-GEM-T Easy vector map and reference points.

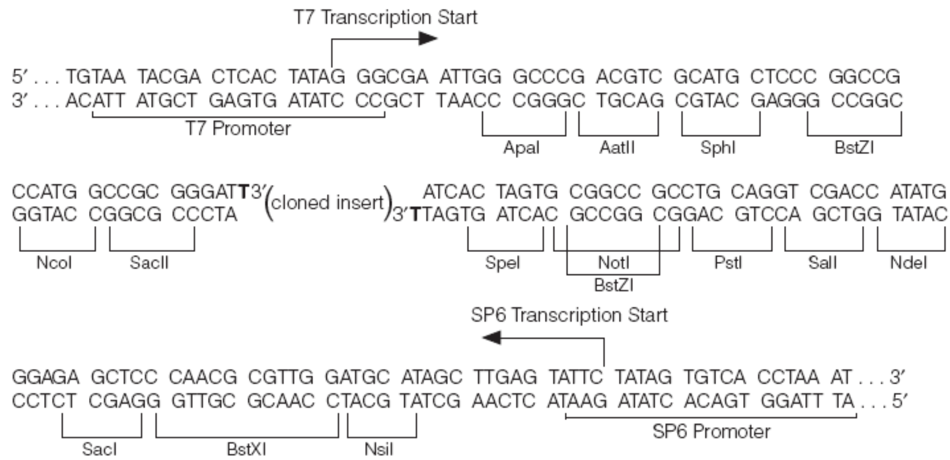


Figure A 3: The promoter and multiple cloning sequence of the pGEM®-T Vector. The top strand corresponds to the RNA synthesized by T7 RNA polymerase. The bottom strand corresponds to the RNA synthesized by SP6 RNA polymerase.

ANNEX 3: Calcium chloride method

To obtain competent bacteria cells, one colony of the strain *Esherichia coli* XL-1 Blue, was isolated from plate of LB medium (30mg/ml tetracycline) and used to inoculate 5 ml of LB medium. Cells were incubated for 12-16 h at 30 ªc and 250 rpm.

1 ml of the culture was inoculated in 250 ml of SOB medium and incubated with agitation at 18 ºC until a DO of 0.6 nm was reached. The bacterial suspension was incubated on ice to stop growth for 10 min and transferred to 50 ml Falcon tubes and centrifuged 1200 g, 10 min, 4 ºC. The supernatant was discarded and the pellet ressuspended in 16 ml of cold transformation buffer (TB), a rich solution of $CaCl_2$, incubated on ice for additional 10 min and centrifuged as described above. DMSO (cryo-preserving agent) was added until a final concentration of 7% was reached and bacteria were incubated for 10 min and aliquoted in 100 µl samples and immediately frozen in liquid nitrogen and stored at –80 ºC until further use.

SOB medium (250 mL):
2% Bacto tryptone (5 g);
1.5 % yeast extract (1.25 g);
10 mM NaCl (500 µL stock 5M);
2.5 mM KCl (625 µL stock 1M);
10 mM $MgCl_2$ (2.5 µL stock 1M);
10 mM $MgSO_4$ (2.5 µL stock 1M);
Autoclave after prepare.

TB buffer (100 mL):
10 mM Pipes (0.30 g);
15 mM $CaCl_2$ (1.5 mL stock 1M);
250 mM KCLO (25 mL stock 1M);
Adjust pH to 6.7 with KOH 1M, otherwise the pipes will not dissolve (take in consideration that the pH changes very fast when near the optimum). After is fully dissolved and the pH is at 6.7 add 55mM MnCl2 (0.69 g). Filter with 0.22 µm filter to sterilize. Do not autoclave the solution. Store it at 4 ºC until use

ANNEX 4: Solutions

1Kb + ladder (1:20)

Sigma water 380 ul
Ladder 20 ul
Loading buffer before use (10 ul)

Bouin

75 ml saturated aqueous solution of piric acid
25 ml formaldehyde
5 ml acetic acid

DEPC-treated water

200 ul of diethylpyrocarbonate (100 ul DEPC/L of water)
Sigma water to 2 L
Agitate and wait 24 hours
Autoclave at 121ºC for at least 40 min

dNTPs

Dilutions of each nucleotide 1:10
For 1 ml:
100 ul dATP
100 ul dTTP
100 ul dCTP
100 ul dGTP
600 ul sigma water

EtBr (500 mg/mL)

EtBr 1% 250 ul
Sterile distilled H2O 4750 ul

GTE

For 100 ml:
1.125 g Glucose
2.5 ml Tris 1M pH 8
2 ml EDTA 0.5 M pH 8
Complete with H2O dd and autoclave

LB

1 pill of LB agar (Sigma-Aldrich)/50 ml of distilled H2O
Autoclave

LB agar

Dissolve 8LB tablets (Sigma) in 400 mL of distilled water, autoclave and allow cooling to 55 º C in a water bath and pouring into Petri dishes. This step should be performed in the laminar flow cabinet. In alternative store the bottle after autoclaving and melt in the microwave when needed, allow to cool to 55ºC and pour into Petri dishes.

LB selective plates

After melting the LB agar and allowing cooling to 55 ºC, add antibiotic or X-Gal to the desired concentration as presented on table A1 before pouring into Petri dishes. This step shoul be performed in the laminar flow cabinet. Store the plates in the fridge for at least 2-3 hours before use.

Table A1: Antibiotic and solutions concentrations for the LB selective plates

Antibiotic/solution	Used at
Ampicilin	75 ug/mL
IPTG	0.5 mM
X-GAL	80ug/mL

Loading Buffer

40% glucose (40 g in 100 mL)
Bromophenol blue (just a few grains)
Sterilize with seringe coupled to an adapter filter (0.2 ul)

Stock IPTG (0.5M)

Dissolve the IPTG in water until achive the desired concentration (0.5M), aliquot and store at -20 ºC.

Mini-prep solutions

Resuspension buffer (P1) – 50 mM Tris HCL pH 8.0, 10 mM EDTA, 100 ug.ml RNase A

5 ml Tris HCL pH 8 1 M
2 ml EDTA 0.5 M
100 ul RNase A 100 mg/ml
Up to 100 ml with H2O

RNase A 100 mg/ml

Dissolve 100 mg RNase A powder in 1 ml of H_2O
Boil 5 min to kill DNase
Aliquot and store at -20ºC

Lysis Buffer – Buffer P2 – 200 mM NaOH, 1% SDS

10 ml NaOH 1 M
5 ml SDS 10 %
35 ml Water

Neutralisation buffer – Buffer P3 – 3.0 M KOAc, pH 5.5

29.5 g KOAc in 100 ml water
Adjust pH with glacial acetic acid

PBS stock solution 10x

79.97 g NaCl
12.46 g $Na_2HPO_4.2H_2O$
4.80 g $NaH_2PO_4.2H_2O$
Dissolve in 800 ml of MilliQ water. Adjust the pH to 7.0. Complete volume to 1 L. Autoclave.

1xPBS is prepared by dilution with double distilled water.

PFA solution 4%

Dilute the 37% paraformaldehyde commercial solution with 1xPBS.

TBE (Tris – Borate – EDTA) Buffer 0.5x

200 ml Tris-HCl 1 M pH 7.5 (freshly prepared)
1.5 ml NaCl 5 M
Complete to 50 ml with double distilled water.

ANNEX 5 – Multisequence alignment for all putative reference genes

Beta-actin

```
Oncorhynchus-mykiss          MEDEIAALVVDNGSGMCKAGFAGDDAPRAVFPSIVGRPRHQGVMVGMGQKDSYVGDEAQSKRGILTLKYPIEHGIVTNWDDMEKIWHHTFYNELRVAPE   99
Poecilia-reticulata          MEDEIAALVVDNGSGMCKAGFAGDDAPRAVFPSIVGRPRHQGVMVGMGQKDSYVGDEAQSKRGILTLKYPIEHGIVTNWDDMEKIWHHTFYNELRVAPE   99
Acanthopagrus-schlegelii     MEDEIAALVVDNGSGMCKAGFAGDDAPRAVFPSIVGRPRHQGVMVGMGQKDSYVGDEAQSKRGILTLKYPIEHGIVTNWDDMEKIWHHTFYNELRVAPE   99
Takifugu-rubripes            MEDEIAALVVDNGSGMCKAGFAGDDAPRAVFPSIVGRPRHQGVMVGMGQKDSYVGDEAQSKRGILTLKYPIEHGIVTNWDDMEKIWHHTFYNELRVAPE   99
Platichthys-flesus           MDDEIAALVVDNGSGMCKAGFAGDDAPRAVFPSIVGRPRHQGVMVGMGQKDSYVGDEAQSKRGILTLKYPIEHGIVTNWDDMEKIWHHTFYNELRVAPE   99
Oreochromis-mossambicus               VDNGSGMCKAGFAGDDAPRAVFPSIVGRPRHQGVMVGMGQKDSYVGDEAQSKRGILTLKYPIEHGIVTNWDDMEKIWHHTFYNELRVAPE   90
Rachycentron-canadum         MEEEIAALVVDNGSGMCKAGFAGDDAPRAVFPSIVGRPRHQGVMVGMGQKDSYVGDEAQSKRGILTLKYPIEHGIVTNWDDMEKIWHHTFYNELRVAPE   99
Osmerus-mordax               MEEEIAALVVDNGSGMCKAGFAGDDAPRAVFPSIVGRPRHQGVMVGMGQKDSYVGDEAQSKRGILTLKYPIEHGIVTNWDDMEKIWHHTFYNELRVAPE   99
Acipenser-transmontanus     MADEEIAALVVDNGSGMCKAGFAGDDAPRAVFPSIVGRPRHQGVMVGMGQKDSYVGDEAQSKRGILTLKYPIEHGIVTNWDDMEKIWHHTFYNELRVAPE  100

Oncorhynchus-mykiss          EHPVLLTEAPLNPKANREKMTQIMFETFNTPAMYVAIQAVLSLYASGRTTGIVMDSGDGVTHTVPIYEGYALPHAILRLDLAGRDLTDYLMKILTERGYS  199
Poecilia-reticulata          EHPVLLTEAPLNPKANREKMTQIMFETFNTPAMYVAIQAVLSLYASGRTTGIVMDSGDGVTHTVPIYEGYALPHAILRLDLAGRDLTDYLMKILTERGYS  199
Acanthopagrus-schlegelii     EHPVLLTEAPLNPKANREKMTQIMFETFNTPAMYVAIQAVLSLYASGRTTGIVMDSGDGVTHTVPIYEGYALPHAILRLDLAGRDLTDYLMKILTERGYS  199
Takifugu-rubripes            EHPVLLTEAPLNPKANREKMTQIMFETFNTPAMYVAIQAVLSLYASGRTTGIVMDSGDGVTHTVPIYEGYALPHAILRLDLAGRDLTDYLMKILTERGYS  199
Platichthys-flesus           EHPVLLTEAPLNPKANREKMTQIMFETFNTPAMYVAIQAVLSLYASGRTTGIVMDSGDGVTHTVPIYEGYALPHAILRLDLAGRDLTDYLMKILTERGYS  199
Oreochromis-mossambicus      EHPVLLTEAPLNPKANREKMTQIMFETFNTPAMYVAIQAVLSLYASGRTTGIVMDSGDGVTHTVPIYEGYALPHAILRLDLAGRDLTDYLMKILTERGYS  190
Rachycentron-canadum         EHPVLLTEAPLNPKANREKMTQIMFETFNTPAMYVAIQAVLSLYASGRTTGIVMDSGDGVTHTVPIYEGYALPHAILRLDLAGRDLTDYLMKILTERGYS  199
Osmerus-mordax               EHPVLLTEAPLNPKANREKMTQIMFETFNTPAMYVAIQAVLSLYASGRTTGIVMDSGDGVTHTVPIYEGYALPHAILRLDLAGRDLTDYLMKILTERGYS  199
Acipenser-transmontanus      EHPVLLTEAPLNPKANREKMTQIMFETFNTPAMYVAIQAVLSLYASGRTTGIVMDSGDGVTHTVPIYEGYALPHAILRLDLAGRDLTDYLMKILTERGYS  200

Oreochromis mossambicus      FTTTAEREIVRDIKEKLCYVALDFEQEMGTAASSSSLEKSYELPDGQVITIGNERFRCPEALFQPSFLGMESCGIHETTYNSIMKCDVDIRKDLYANTVL  290
Oncorhynchus mykiss          FTTTAEREIVRDIKEKLCYVALDFEQEMGTAASSSSLEKSYELPDGQVITIGNERFRCPEALFQPSFLGMESCGIHETTYNSIMKCDVDIRKDLYANTVL  299
Poecilia reticulata          FTTTAEREIVRDIKEKLCYVALDFEQEMGTAASSSSLEKSYELPDGQVITIGNERFRCPEALFQPSFLGMESCGIHETTYNSIMKCDVDIRKDLYANTVL  299
Acanthopagrus schlegelii     FTTTAEREIVRDIKEKLCYVALDFEQEMGTAASSSSLEKSYELPDGQVITIGNERFRCPEALFQPSFLGMESCGIHETTYNSIMKCDVDIRKDLYANTVL  299
Takifugu rubripes            FTTTAEREIVRDIKEKLCYVALDFEQEMGTAASSSSLEKSYELPDGQVITIGNERFRCPEALFQPSFLGMESCGIHETTYNSIMKCDVDIRKDLYANTVL  299
Platichthys flesus           FTTTAEREIVRDIKEKLCYVALDFEQEMGTAASSSSLEKSYELPDGQVITIGNERFRCPEALFQPSFLGMESCGIHETTYNSIMKCDVDIRKDLYANTVL  299
Rachycentron canadum         FTTTAEREIVRDIKEKLCYVALDFEQEMGTAASSSSLEKSYELPDGQVITIGNERFRCPEALFQPSFLGMESCGIHETTYNSIMKCDVDIRKDLYANTVL  299
Osmerus mordax               FTTTAEREIVRDIKEKLCYVALDFEQEMSTAASSSSLEKSYELPDGQVITIGNERFRCPEALFQPSFLGMESCGIHETTYNSIMKCDVDIRKDLYANTVL  299
Acipenser transmontanus      FTTTAEREIVRDIKEKLCYVALDFEQEMATAASSSSLEKSYELPDGQVITIGNERFRCPEALFQPSFLGMESCGIHETTFNSIMKCDVDIRKDLYANTVL  300

Oreochromis mossambicus      SGGTTMYPGIADRMQKEITALAPSTMKIKIIAPPERKYSVWIGGSILASLSTFQQMWISKQEYDESGPS          359
Oncorhynchus mykiss          SGGTTMYPGIADRMQKEITSLAPSTMKIKIIAPPERKYSVWIGGSILASLSTFQQMWISKQEYDESGPSIVHRKCF  375
Poecilia reticulata          SGGTTMYPGIADRMQKEITACAPSTMKIKIIAPPERKYSVWIGGSILASLSTFQQMWISKQEYDESGPSIVHRKCF  375
Acanthopagrus schlegelii     SGGTTMYPGIADRMQKEITALAPSTMKIKVIAPPERKYSVWIGGSILASLSTFQQMWISKQEYDESGPSIVHRKCF  375
Takifugu rubripes            SGGTTMYPGIADRMQKEITALAPSTMKIKIIAPPERKYSVWIGGSILASLSTFQQMWISKQEYDESGPSIVHRKCF  375
Platichthys flesus           SGGTTMYPGIADRMQKEITALAPSTMKIKIIAPPERKYSVWIGGSILASLSTFQQMWISKQEYDESGPSIVHRKCF  375
Rachycentron canadum         SGGTTMYPGIADRMQKEITALAPSTMKIKIIAPPERKYSVWIGGSILASLSTFQQMWISKQEYDESGPSIVHRKCF  375
Osmerus mordax               SGGTTMYPGIADRMQKEITALAPSTMKIKIIAPPERKYSVWIGGSILASLSTFQQMWISKQEYDESGPSIVHRKCF  375
Acipenser transmontanus      SGGTTMYPGIADRMQKEITALAPSTMKIKIIAPPERKYSVWIGGSILASLSTFQQMWISKQEYDESGPSIVHRKCF  376
```

Figure A4 - Multisequence alignment of Beta-actin proteins. Species used were: *Oncorhynchus mykiss* (gi: 185132289); *Poecilia reticulata* (gi: 160693772); *Acanthopagrus schlegelii* (gi: 40362701); *Takifugu rubripes* (gi: 67462093); *Platichthys flesus* (gi: 7546744); *Rachycentron canadum* (gi: 161376754); *Osmerus mordax* (gi: 225707088); *Acipenser transmontanus* (gi: 207298859).

Glyceraldehyde-3-phosphate deshydrogenase

```
Oreochromis mossambicus                                                     ...        KYDSTHGTWKHGEVKSEGGKLVIGNMHIMVFQERDPANIKWGDAGVDYVVESTGVFTTIEKASAHLKGG 69
Oplegnathus fasciatus      MVKVGVNGFGRIGRLVTRAAFTSKK-VEIVAINDPFIDLEYMVYMFKYDSTHGRFKG-EVKMEGDKLVIDGHKITVFHERDPANIKWGDAGAQYVVESTGVFTTIEKASAHLKGG 113
Lates-calcarifer                                                       MVYMFKYDSTHGRYKG-EVKVEGDKLVIDGHKITVFHERDPANIKWGEAGAQYVVESTGVFTTIDKASAHLKGG 73
Sparus aurata              MVKVGINGFGRIGRLVTRAGFTSKK-VEIVAINDPFIDLEYMVYMFKYDSTHGRYHG-EVKIEGDKLVIDGHKITVFHERDPAHIKWGDAGAQYVVESTGVFTTIEKASAHLKGG 113
Tetraodon nigroviridis     MVKVGINGFGRIGRLVTRAAFTSKK-VEIVAINDPFIDLEYMVYMFKYDSTHGRYKG-EVKAEGGKLVIDGHAITVFHERDPANIKWGEAGAQYVVESTGVFTTIEKASAHLKGG 113
Dicentrarchus labrax       MVKVGINGFGRIGRLVTRAAFTSKK-VEIVAINDPFIDLEYMVYMFKYDSTHGRFKG-EVKVEGDKLVVDGHKISVFHERDPANIKWGEAGAQYVVESTGVFTTIEKASAHLKGG 113
Platichthys flesus         MVKVGINGFGRIGRLVTRAAFTSKK-VEIVAINDPFIDLEYMVYMFKYDSTHGRFHG-QVSVEGDKLVIDGHKISVFHERDPANIKWGDAGAAYVVESTGVFTTIEKASAHLKGG 113
Kryptolebias marmoratus    MVKVGINGFGRIGRLVTRAAILSKK-VEIVAINDPFIDLEYMVYMFKYDSTHGRFHG-EVKAEGDKLVIDGHKISVFHERDPTNIKWGEAGAEYVVESTGVFTTIEKASAHLKGG 113
Sebastiscus marmoratus     MVKVGINGFGRIGRLVTRAAFTSKK-VEIVAINDPFIDLEYMVYMFKYDSTHGRFKG-EVKIEGDKLVIDGQKITVFHERDPANIKWGDAGAQYVVESTGVFTTIEKASTHLKGG 113
Gadus-morhua               MVKVGINGFGRIGRLVTRAGFQSKK-VEIVAINDPFIDLEYMVYMFKYDSTHGRFHG-EVKIEGDKLVIDGHKISVFHERDPANIKWGDAGAQYVVESTGVFTTIEKASAHLKGG 113
Misgurnus anguillicaudatus MVKVGINGFGRIGRLVTRAAFQSKK-VEIVAINDPFIDLDYMVYMFKYDSTHGRYKG-EVKAEGGKLVIDGHAITVFSERDPSNIKWGDAGAQYVVESTGVFTTIEKASTHLKGG 113
Oncorhynchus mykiss        MVKVGVNGFGRIGRLVTRAAFHSKKGVEIVAINDPFIDLDYMVYMFKYDSTHGRFHG-EVKAEGGKLVIDGHKITVFHERDPANIKWGDAGATYVVESTGVFTTIEKASTHLKGG 114
Anguilla japonica          MVKVGVNGFGRIGRLVTRAAFHSGK-VQVVAINDPFIDLDYMVYMFKYDSTHGRFHG-EVKAEGGKLHIAGHVITVFHERDPTNIKWGEAGADYVVESTGVFTTIEKASTHLKGG 113
Salmo salar                MVKIGINGFGRIGRLVTRAAAKSGK-VEVVAINDPFIDLDYMVYMFKYDSTHGVWKHSEVKQEGGKLIIGNLHITVFHERDPTAIKWGEAGADYVVESTGVFTTIDKASAHLQGG 114
Esox lucius                MVNIGINGFGRIGRLVTRAAAKSGK-VNVVAINDPFIDLDYMVYMFKYDSTHGVWKDSEVKQEGGKLIIGNLRITVFHERDPSLIKWGEAGADYVVESTGVFTTIEKASAHLKGG 114
Epinephelus coioides       MVKIGINGFGRIGRLVTRAAAQGGK-VEVVAINDPFIDLDYMVYMFKYDSTHGPWKHGEVKAEGGKLVIGKMHIMVFHERDPANIKWSDAGADYVVESTGVFTTIEKASAHLKGG 114

Oreochromis mossambicus    AKRVVISAPSADAPMFVMGVNHEKYDKSLTVVSNASCTTNCLAPLAKVINDSFGIVEGLMSTVHAVTATQKTVDGPSGKMWRDGRGAS ...                          157
Oplegnathus fasciatus      AKRVIISAPSADAPMFVMGVNHEKYDKSLPVVSNASCTTNCLAPLAKVIHDNFGIVEGLMSTVHAITATQKTVDGPSGKLWRDGRGASQNIIPASTGAAKAVGKVIPELNGKLTG 228
Lates calcarifer           AKRVIISAPSADAPMFVMGVNHEKYDKSLSVVSNASCTTNCLAPLAKVIHDNFGIIEGLMSTVHAITATQKTVDGPSGKLWRDGRGASQNIIPASTGAAKAVGKVIPELNGKLTG 188
Sparus aurata              AKRVIISAPSADAPMFVMGVNHQKYDKSLPVVSNASCTTNCLAPLAKVINDNFGIIEGLMSTVHAITATQKTVDGPSGKLWRDGRGASQNIIPASTGAAKAVGKVIPELNGKLTG 228
Tetraodon nigroviridis     AKRVIISAPSADAPMFVMGVNHEKYDNSLAVVSNASCTTNCLAPLAKVINDNFGIIEGLMSTVHAITATQKTVDGPSGKLWRDGRGASQNIIPASTGAAKAVGKVIPELNGKLTG 228
Dicentrarchus labrax       AKRVVISAPSADAPMFVMGVNHEKYSNSLPVVSNASCTTNCLAPLAKVINDNFGIIEGLMSTVHAITATQKTVDGPSGKLWRDGRGASQNIIPASTGAAKAVGKVIPELNGKLTG 228
Platichthys flesus         AKRVIISAPSADAPMFVVGVNHEKYENSLKVVSNASCTTNCLAPLAKVINDNFGIIEGLMSTVHAITATQKTVDGPSGKLWRDGRGASQNIIPASTGAAKAVGKVIPELNGKLTG 228
Kryptolebias marmoratus    AKRVIISAPSADAPMFVMGVNHEKYDKSLQVVSNASCTTNCLASLAKVINDNFGIIEGLMSTVHAITATQKTVDGPSGKLWRDGRGASQNIIPASTGAAKAVGKVIPELNGKLTG 228
Sebastiscus marmoratus     AKRVIISAPSADAPMFVMGVNHEKYDNSLTVVSNASCTTNCLAPLAKVVNDNFGIIEGLMSTVHAITATQKTVDGPSGKLWRDGRGASQNIIPASTGAAKAVGKVIPELNGKLTG 228
Gadus morhua               AKRVIISAPSADAPMFVMGVNHEKYENSLQVVSNASCTTNCLAPLAKVIHDNFGIVEGLMSTVHAITATQKTVDGPSGKLWRDGRGASQNIIPASTGAAKAVGKVIPELNGKITG 228
Misgurnus anguillicaudatus AKRVIISAPSADAPMFVMGVNHEKYDNSLKVVSNASCTTNCLAPLAKVINDNFGIIEGLMSTVHAITATQKTVDGPSGKLWRDGRGASQNIIPASTGAAKAVGKVIPELNGKLTG 228
Oncorhynchus mykiss        AKRVVISAPSADAPMFVMGVNHEKYENSLKVVSNASCTTNCLAPLAKVIHDNYHIIEGLMSTVHAVTATQKTVDGPSGKLWRDGRGASQNIIPASTGAAKAVGKVIPELNGKITG 229
Anguilla japonica          AKRVIISAPSADAPMFVMGVNHEKYDKSLKVVSNASCTTNCLAPLAKVVNDKFVIIEGLMSTVHAITATQKTVDGPSGKLWRDGRGASQNIIPASTGAAKAVGKVIPELNGKLTG 228
Salmo salar                AKRVIISAPSADAPMFVMGVNHEKYDNSLKVVSNASCTTNCLAPIAKVINDNFGIVEGLMSTVHAVTATQKTVDGPSGKLWRDGRGASQNIIPASTGAAKAVGKVIPELNGKLTG 229
Esox lucius                AKRVIISAPSADAPMFVMGVNHEKYENSLKVVSNASCTTNCLAPIAKVINDNFGIVEGLMSTVHSVTATQKTVDGPSGKLWRDGRGAGQNIIPASTGAAKAVGKVIPELNGKLTG 229
Epinephelus coioides       AKRVVISAPSADAPMFVMGVNHEKYDNSMKVVSNASCTTNCLAPLAKVINDNFGIIEGLMSTVHAITATQKTVDGPSGKLWRDGRGASQNIIPASTGAAKAVGKVVPELNGKLTG 229

Oreochromis mossambicus
Oplegnathus fasciatus      MAFRVPTPNVSVVDLTVRLQKPAKYDDIKKVVKAAAEGPMKGILGYTEHQVVSTDFNGDCHSSIFDAGAGIALNEHFVKLVSWYDNEFGYSNRVCDLMAHMAAKE 333
Lates calcarifer           MAFRVPTPNVSVVDLTVRLEKPAKYDDIKKVVKAAADGPMKGILGYTEHQVVSTDF ...                                                 244
Sparus aurata              MAFRVPTPNVSVVDLTVRLEKPAKYDDIKKVVKAAAEGPMKGILGYTEHQVVSTDFNGDSHSSIFDAGAGIALNDHFVKLVTWYDNEFGYSNRVCDLMAHMASKE 333
Tetraodon nigroviridis     MAFRVPTPNVSVVDLTVRLEKPAKYEDIKKVVKAAADGPMKGILGYTEHQVVSTDFNGDSHSSIFDAGAGIALNDHFVKLVTWYDNEFGYSNRVCDLMAHMAGKE 333
Dicentrarchus labrax       MAFRVPTPNVSVVDLTVRLEKPAKYDDIKKVVKAAAEGPMKGILGYTEHQVVSTDFNGDTHSSIFDAGAGIALNDHFVKLVTWYDNEFAYSNRVCDLMAHMASKE 333
Platichthys flesus         MAFRVPTPNVSVVDLTVRLEKPASYENIKKVVKAAADGPMKGYLAYTEHQVVSSDFNGDTHSSIFDAGAGIALNDHFVKLVSWYDNEFAYSQRVCDLMSHMAAKE 333
Kryptolebias marmoratus    MAFRVPTPNVSVVDLTVRLEKPAKYDDIKKVVKAAAEGPMKGILGYTEHQVVSTDFNGDTHSSIFDAGAGIALNDHFVKLVSWYDNEFGYSNRVCDLMAHMASKE 333
Sebastiscus marmoratus     MAFRVPTPNVSVVDLTVRLEKPAKYDDIKKVIKAAAEGPMKGILGYTEHQVVSTDFNGDCHSSIFDAGAGIALNDHFVKLVTWYD ...                  313
Gadus morhua               MAFRVPTPNVSVVDLTVRLEKPAKYDDIKKVIKAASEGPMKGILGYTEDQVVSTDFNGDIRSSIFDAGAGIALNDHFVKLVTWYDNEFGYSNRVIDLMAHMSCKE 333
Misgurnus anguillicaudatus MAFRVPTPNVSVVDLTVRLEKPAKYDDIKKVVKAASEGPMKGILGYTEHQVVSSDFNGDTHSSIFDAGAGIALNDHFVKLVTWYDNEFGYSNRVCDLMAHMATKE 333
Oncorhynchus mykiss        MAFRVPTPNVSVVDLTVRLEKPASYDAIKKVVKAAADGPMKGILGYTEQQVVSSDFNGDTHSSIFDAGAGIALNDHFVKLVTWYDNEFGYSNRVIDLMAHMATKE 334
Anguilla japonica          MAFRVPTPNVSVVDLTVRLEKPAKYDDIKKVIKEAAEGPMKGILGYTEHQVVSTDFNGDTHSSIFDAGAGIALNDHFVKLVSWYDNEFGYSNRVVDLMVHMASKE 333
Salmo salar                MAFRVPTPNVSVVDLTVRLEKAAPYDEIKKVIKAAAEGPMKGILGYTEDQVVSTDFNSDCRSSIFDAGAGIALNDHFVKLVSWYDNEFGYSNRVVDLCLHMASKE 334
Esox lucius                MAFRVPTPNVSVVDLTVRLQKAAPYDQIKKVIK-AAEGPMKGILGYTEDQVVSTDFNGDCRSSIFDAGAGIALNDHFVKLVSWYDNEFGYSNRVVDLCMHMASKE 333
Epinephelus coioides       MAFRVPTPNVSVVDLTVRLEKPAKYDDIK ...                                                                          258
```

Figure A5 - Multisequence alignment of Glyceraldehyde-3-phosphate deshydrogenase proteins from fish species. Species used were: *Oplegnathus fasciatus* (gi: 194241594); *Lates calcarifer* (gi: 259130081); *Sparus aurata* (gi: 108946636); *Tetraodon nigroviridis* (gi: 47221580); *Dicentrarchus labrax* (gi: 57791244); *Platichthys flesus* (gi: 12584632); *Kryptolebias marmoratus* (gi: 215261537); *Sebastiscus marmoratus* (gi: 219808148); *Gadus morhua* (gi: 25989185); *Misgurnus anguillicaudatus* (gi: 119943230); *Oncorhynchus mykiss* (gi: 185135354); *Anguilla japonica* (gi: 21955965); *Salmo salar* (gi: 209737954); *Esox lucius* (gi: 225715600); *Epinephebus coioides* (gi: 221048033).

Hypoxanthine phosphoribosyl transferase 1

```
Oreochromis mossambicus                                                       ... GHHIVALCVLKGGYKFFADLLDYIKSLNQNSDKSVPLTV 39
Anoplopoma fimbria       MAS------YLQIADDEKGYELDLFCVPRHYENDLERVIIPHGLIMDRTERLARDIMRDMGGHHIVALCVLKGGYKFFADLLDYIKALNQNSDQSVPLTV 94
Danio rerio              MAA------FLEISDDSKGHELSLFCVPKHYEDDLDSVIIPNGLIKDRTERLARDIVRDMGGHHIVALCVLKGGYKFFADLMDFIKTLNQHSDKSVPLTV 94
Salmo salar              MAS---SSPCVVISDDEQGYDLDLFCIPKHYADDLDRVYIPHGLILDRTERLAREIMKDMGGHHIVALCVLKGGYKFFADLLDYIKALNRNSDRSIPMTV 97
Osmerus mordax           MAT---NSPCVVISDDEQGYDLDLFCIPKHYAADLERVYIPHGLILDRTERLAREIMKDMGGHHIVALCVLKGGYKFFADLLDYIKALNRNSDRSIPMTV 97
Solea senegalensis       MAT---SSSCVVISDEEQGYDLDLFCIPKHYASDLERVYIPHGLILDRTERLAREIMKEMGGHHIVALCVLKGGYKFFADLLDYIKALNRNSDRSIPMTV 97
Xiphophorus nigrensis                 ... GYDLDLFCIPKHYAADLEKVYIPHGLILDRTERLAREIMKEMGGHHIVALCVLKGGYKFFADLLDYIKALNRNSDRSIPMTV 82
Cairina moschata         MAT---PSPCIVIDDDEQGYDLDLFCIPKHYADDLEKVYIPHGLIMDRTERLAREIMKGMGGHHIVALCVLKGGYKFFADLLDYIKALNRNSDKSIPMTV 97
Gallus gallus            MAT---HSPCIVIGDDEQGYDLDLFCIPKHYADDLEKVYIPHGLIMDRTERLAREIMKGMGGHHIVALCVLKGGYKFFADLLDYIKALNRNSDKSIPMTV 97
Taeniopygia guttata      MSESLLAELQMQIGDDEQGYDLDLFCIPKHYADDLEKVYIPHGLIMDRTERLAREIMKGMGGHHIVALCVLKGGYKFFADLLDYIKALNRNSDKSIPMTV 100
Xenopus laevis           MAS-----PCIVIQDDEQGYDLDLFCIPKHYAADLEKVYIPHGLIMDRTERLARDIMKDMGGHHIVALCVLKGGYKFFADLLDYIKALNRNSDKSIPMTV 95
Mus musculus             MAT---RSPSVVISDDEPGYDLDLFCIPNHYAEDLEKVFIPHGLIMDRTERLARDVMKEMGGHHIVALCVLKGGYKFFADLLDYIKALNRNSDRSIPMTV 97
Homo sapiens             MAG---FG----ISDDEPGYDLDLFCIPNHYAEDLERVFIPHGLIMDRTERLARDVMKEMGGHHIVALCVLKGGYKFFADLLDYIKALNRNSDRSIPMTV 93

Oreochromis mossambicus  DFIRVKSYCNDKSTNNVKVIGGDELSNLSGKNVLIVEDIVETGRTMETLLSLLSECNP ...                                          97
Anoplopoma fimbria       DFIRVKSYCNDKSTNSVKVIGGDELSSLSGKNVLIVEDIVETGRTMQTLLSLLSECKPKMVKVVSLLVKRTPRSSGYRPDYIGFEVPDAFLVGYALDYNE 194
Danio rerio              DFIRLKSYCNDQSTNCVKVIGGDELSALAGKNVLIVEDIVETGKTMETLLKLLHECHPKMVKVVSLLVKRTPRSSGYRPDYTGFEVPDKFLVGYALDYNE 194
Salmo salar              DFIRLKSYCNDQSTGEIKVIGGDILSTLTGKNVLIVEDIIDTGKTMKTLLQLLKQYNPKMVKVASLLVKRTPRSVGYIPDFIGFEVPDKFVVGYALDYNE 197
Osmerus mordax           DFIRLKSYCNDQSTGEIKVIGGDILSTLTGKNVLIVEDIIDTGKTMKTLLQLLQQYKPKMVKVASLLVKRTPRSVGYRPDFVGFEVPDKFVVGYALDYNE 197
Solea senegalensis       DFIRLKSYCNDQSTGEIKVIGGDILSTLTGKNVLIVEDIIDTGKTMKTLLELLKQYNPKMVKVASLLVKRTPRSVGYRPDYVXFEVPDKFVVGYALDYNE 197
Xiphophorus nigrensis    DFIRLKSYCNDQSTGEIKVIGGDILSTLTGKNVLIVEDIIDTGKTMKTLLQLLKQYNPKMVKVASLLVKRTPRSVGYRPDFVGFEVPDKFVVGYALDYNE 182
Cairina moschata         DFIRLKSYCNDQSTGDIKVIGGDILSTLTGKNVLIVEDIIDTGKTMKTLLSLLKQYNPKMVKVASLLVKRTPRSVGYRPDFVGFEVPDKFVVGYALDYNE 197
Gallus gallus            DFIRLKSYCNDQSTGDIKVIGGDILSTLTGKNVLIVEDIIDTGKTMKTLLSLLKQYNPKMVKVASLLVKRTPRSVGYRPDFVGFEVPDKFVVGYALDYNE 197
Taeniopygia guttata      DFIRLKSYCNDQSTGDIKVIGGDILSTLTGKNVLIVEDIIDTGKTMKTLLSLLKQYNPKMVKVASLLVKRTPRSVGYRPDFVGFEVPDKFVVGYALDYNE 200
Xenopus laevis           DFIRLKSYCNDQSTGDIKVIGGDILSTLTGKNVLIVEDIIDTGKTMKTLLAMLKKYNPKMVKVASLLVKRTPRSVGYRPDFVGFEVPDKFVVGYALDYNE 195
Mus musculus             DFIRLKSYCNDQSTGDIKVIGGDILSTLTGKNVLIVEDIIDTGKTMQTLLSLVKQYSPKMVKVASLLVKRTSRSVGYRPDFVGFEIPDKFVVGYALDYNE 197
Homo sapiens             DFIRLKSYCNDQSTGDIKVIGGDILSTLTGKNVLIVEDIIDTGKTMQTLLSLVRQYNPKMVKVASLLVKRTPRSVGYKPDFVGFEIPDKFVVGYALDYNE 193

Oreochromis mossambicus                        97
Anoplopoma fimbria       YFRDLSHICILNDQAKEKYKV 215
Danio rerio              YFRDLSHICILSDRAREKYKV 215
Salmo salar              YFRDLNHICVISETGKEKYKA 218
Osmerus mordax           YFRDLNHICVISETGKEKYKA 218
Solea senegalensis       YFRDLNHICVISETGKGEVQGMKST 222
Xiphophorus nigrensis    YFRDLN ...            188
Cairina moschata         YFRDLNHICVISETGKQKYKA 218
Gallus gallus            YFRDLNHICVISETGKQKYKA 218
Taeniopygia guttata      YFRDLNHICVISETGKQKYKA 221
Xenopus laevis           YFRDLNHICVISEDGKERYKT 216
Mus musculus             YFRDLNHVCVISETGKAKYKA 218
Homo sapiens             YFRDLNHVCVISETGKAKYKA 214
```

Figure A6 - Multisequence alignment of hypoxanthine phosphoribosyl transferase 1 proteins. Species used were: *Anoplopoma fimbria* (gi: 229366178); *Danio rerio* (gi: 50344766); *Salmo salar* (gi: 197631785); *Osmerus mordax* (gi: 225708596); *Solea senegalensis* (gi: 169643691); *Xiphophorus nigrensis* (gi: 113204717); *Cairina moschata* (gi: 125743188); *Gallus gallus* (gi: 45382333); *Taeniopygia guttata* (gi: 224097890); *Xenopus laevis* (gi: 148229801); *Mus musculus* (gi: 1343562); *Homo sapiens* (gi: 119632165).

TATA box binding protein

```
Oreochomis mossambicus       ...               YGSGLTPQPVQNTNSLSILEEQQRQQQAQQQQ------------QQAQQQAQQASSGTSGQTPQLF 54
Salmo salar                  MEQNNSLPPF-QGLASPQGAMTPGMPIFSPMMPYGSGLTPQPVTNTNSLSLLEEQHRQQQ-QQQ---------------ASTQQ-GGVSGGSGQTPQLY 81
Danio rerio                  MEQNNSLPPFAQGLASPQGAMTPGLPIFSPMMPYGTGLTPQPVQNSNSLSLLEEQQRQQQ-QQQ---------------AASQQQGGMVGGSGQTPQLY 83
Oryzias latipes              MDQNNSIQAF-QGLASPPGVMTPTMPVFSPMMPYGSGLTPQPVQNTNSLSILEEQQRQQQ-QQQQ------------QQQTQQLNTGLPGASGQTPQLF 85
Takifugu rubripes            MDQNNSIPGF-QGLASPQGAMTPSMPIFSPMMPYGSGLTPQPVQNTNSLSILEEQQRQQQ-QQQ----------------AQQANTGIPGTSGTTPQLF 81
Gallus gallus                MDQNNSLPPYAQGLASPQGAMTPGIPIFSPMMPYGTGLTPQPVQSTNSLSILEEQQRQQQ-QQQ-------------AAQQSTSQQATQGTSGQTPQLF 85
Taeniopygia guttata          MDQNNSLPPYAQGLASPQGAMTPGIPIFSPMMPYGTGLTPQPVQSTNSLSILEEQQRQQQ-QQQ-------------AAQQSTSQQATQGTSGQTPQLF 85
Mus musculus                 MDQNNSLPPYAQGLASPQGAMTPGIPIFSPMMPYGTGLTPQPIQNTNSLSILEEQQRQQQ-QQQQQQQQQAVATAASVQQSTSQQPTQGASGQTPQLF 99
Homo sapiens                 MDQNNSLPPYAQGLASPQGAMTPGIPIFSPMMPYGTGLTPQPIQNTNSLSILEEQQRQQQ-QQQQQQQQQQAVAAAAVQQSTSQQATQGTSGQAPQLF 99

Oreochomis mossambicus       HSQAVGGSTTTALPGNTPLFT-TPLTPMTPITPATPASETSGIVPQLQNIVSTVNLGCKLDLKTIALRARNAEYDPKRFAAVIMRIREPRTTALIFSSGK 153
Salmo salar                  HS--QTVTTTT-LPGNTPLYTATPLTPMTPITPATPASESSGIVPQLQNIVSTVNLGCKLDLKTIALRARNAEYNPKRFAAVIMRIREPRTTALIFSSGK 178
Danio rerio                  HST-QAVSTTTALPGNTPLYT-TPLTPMTPITPATPASESSGIVPQLQNIVSTVNLGCKLDLKTIALRARNAEYNPKRFAAVIMRIREPRTTALIFSSGK 181
Oryzias latipes              HSQPVPGSTTTALPGNTPLYS-APVTPMTPIAPATPASESSGIVPQLQNIVSTVNLGCKLDLKTIALRARNAEYNPKRFAAVIMRIREPRTTALIFSSGK 184
Takifugu rubripes            HSQAVAGTTTTALPGNTPLYN-TPLTPMTPITPATPASESSGIVPQLQNIVSTVNLGCKLDLKTIALRARNAEYNPKRFAAVIMRIREPRTTALIFSSGK 180
Gallus gallus                HS--QTLTTAP-LPGTTPLYP-SPMTPMTPITPATPASESSGIVPQLQNIVSTVNLGCKLDLKTIALRARNAEYDPKRFAAVIMRIREPRTTALIFSSGK 181
Taeniopygia guttata          HS--QTLTTAP-LPGTTPLYS-SPMTPMTPITPATPASESSGIVPQLQNIVSTVNLGCKLDLKTIALRARNAEYNPKRFAAVIMRIREPRTTALIFSSGK 181
Mus musculus                 HS--QTLTTAP-LPGTTPLYP-SPMTPMTPITPATPASESSGIVPQLQNIVSTVNLGCKLDLKTIALRARNAEYNPKRFAAVIMRIREPRTTALIFSSGK 195
Homo sapiens                 HS--QTLTTAP-LPGTTPLYP-SPMTPMTPITPATPASESSGIVPQLQNIVSTVNLGCKLDLKTIALRARNAEYNPKRFAAVIMRIREPRTTALIFSSGK 195

Oreochomis mossambicus       MVCTGAKSEEQSRLAARKYARVVQKLGFPAKFLDFKIQNMVGNCDVKFPIRLEGLVLTHQQFSSYEPELFPGLIYRMIKPRIVLLIFVSGKVVL       247
Salmo salar                  MVCTGAKSEEQSRLAARKYARVVQKLGFPAKFLDFKIQNMVGSCDVKFPIRLEGLVLTHQQFSSYEPELFPGLIYRMIKPRIVLLIFVSGKVVLTGAKVR 278
Danio rerio                  MVCTGAKSEEQSRLAARKYARVVQKLGFPAKFLDFKIQNMVGSCDVKFPIRLEGLVLTHQQFSSYEPELFPGLIYRMIKPRIVLLIFVSGKVVLTGAKVR 281
Oryzias latipes              MVCTGAKSEEQSRLAARKYARVVQKLGFPAKFLDFKIQNMVGSCDVKFPIRLEGLVLTHQQFSSYEPELFPGLIYRMIKPRIVLLIFVSGKVVLTGAKVR 284
Takifugu rubripes            MVCTGAKSEEQSRLAARKYARVVQKLGFPAKFLDFKIQNMVGSCDVRFPIRLEGLVLTHQQFSSYEPELFPGLIYRMIKPRIVLLIFVSGKVVLTGAKVR 280
Gallus gallus                MVCTGAKSEEQSRLAARKYARVVRKLGFPAKFLDFKIQNMVGSCDVKFPIRLEGLVLTHQQFSSYEPELFPGLIYRMIKPRIVLLIFVSGKVVLTGAKVR 281
Taeniopygia guttata          MVCTGAKSEEQSRLAARKYARVVQKLGFPAKFLDFKIQNMVGSCDVKFPIRLEGLVLTHQQFSSYEPELFPGLIYRMIKPRIVLLIFVSGKVVLTGAKVR 281
Mus musculus                 MVCTGAKSEEQSRLAARKYARVVQKLGFPAKFLDFKIQNMVGSCDVKFPIRLEGLVLTHQQFSSYEPELFPGLIYRMIKPRIVLLIFVSGKVVLTGAKVR 295
Homo sapiens                 MVCTGAKSEEQSRLAARKYARVVQKLGFPAKFLDFKIQNMVGSCDVKFPIRLEGLVLTHQQFSSYEPELFPGLIYRMIKPRIVLLIFVSGKVVLTGAKVR 295

Oreochomis mossambicus         ...
Salmo salar                  GEIYEAFENIYPILKGFRKTT 299
Danio rerio                  GEIYEAFENIYPILKGFRKTS 302
Oryzias latipes              AEIYEAFENIYPILKGFRKTT 305
Takifugu rubripes            AEIYEAFENIYPILKGFRKTT 301
Gallus gallus                AEIYEAFENIYPILKGFRKTT 302
Taeniopygia guttata          AEIYEAFENIYPILKGFRKTT 302
Mus musculus                 AEIYEAFENIYPILKGFRKTT 316
Homo sapiens                 AEIYEAFENIYPILKGFRKTT 316
```

Figure A7 - Multisequence alignment of TATA box binding proteins. Species used were: *Salmo salar* (gi: 223648344); Danio rerio (gi: 41056135); *Oryzias latipes* (gi: 17298076); *Takifugu rubripes* (gi: 153792245); *Gallus gallus* (gi: 164429967); *Taeniopygia guttata* (gi: 224047784); *Mus musculus* (gi: 148688513); *Homo sapiens* (gi: 80474966).

Beta-2 microglobulin

```
Oreochromis mossambicus                                                          ... PDITITLLKNGAEIPDAKQTDLVFNQDWHFHLTKHVAFTPKEGEKYT 47
Oncorhynchus mykiss        MKTVLSVIAFCVFLGFINAKESPPKVQVYSRNPGQHGKDDTLICHVSGFHPPDISIQLLKNGVEIPDAKQTDLAFEQGWQFHLTKSVGFTPASGEEYT 98
Salmo salar                MKSILSIVVLVLIYSAVESKESPPKVQVYSRNPGNFGDKNTLICHVSGFHPPDISIQLLKNGVEIPDAKQTDLAFEQGWQFHLTKSVGFTPARGEEYT 98
Osmerus mordax             MKAFFSFAVIFLIYSTVESKESPPKVQVYGHNPGKYGAPNTLICHVSGFHPPDISIQLMKNGMEIPNAKQTDLAFEKGWKFHLTKSVDFTPQEEDKFT 98
Hypomesus transpacificus   MKAAFSFAVIFLIFSTVESKESPPKVQVYGHNPGKYGTPNTLICHVSGFHPPDISIQLMKNGVEIPNAKQTDLAFEKGWKFHLTKSVDFTPQEEDKFI 98
Esox lucius                MNYILSTVVVALVFCVVESRESPPKVQVYSHNPGKYGQENTLICHVSGFHPPDIDIKLMKNGNEIPGAKQTDLAFEQGWRFHLTKSVAFTPNDGDDYS 98
Sander vitreus             MMLAWCLAALVAVSFAADSKHTSPKVQVYSNHPGEYGKDNTLICHVTGFHPPDITIQVMKNGVELPEAKQTDLAFKQDWHFHLTKTVPFTPMDGDKYT 98
Epinephelus coioides     MDKILVMCLAALVALSCAEETKHSSPKVQVYSFQPGQYGKDNTLICHVSKFHPPDITIQLMKNGQELPNAEQTDLAFENDWHFHLTKHVAFKPMDGDKYS 100
Anoplopoma fimbria         MMLGYCLAALLAVCFAQDAMHTPPKVQVYSHKPGEYGKENTLICRVSGFHPPDITIELMKNDMEIPEAQQTDLAFKKGWHFHLTKNVAFTPNREDRYS 98
Pagrus major               MRIVLCLAALAAVYCSDDAKYTPPKVQVYTAGPAQFGEKNTLICHVSGFHPPDITIQLMKEGEEISNANQTDLAFKPDWHFHLTKNVAFTPTRGQKYS 98
Larimichthys crocea        MKFVLCLAALVAVCYSQDSKHTRPKVQVYSRDPGKFGSKNVLICHVSGFHPPDITIQLMKDEAELPNANQTDLAFKQGWRFHLTKNVAFTPLDGEKYS 98
Seriola quinqueradiata     MNIALCAAVMVAVCLSVDAKVSPPKVQVYSRNPGEYGAKNTLICHVSNFHPPDISIELMKDGVELPDAKQTDLAFKQDWHFHLTKNVEFTPNDGQKYS 98
Lates calcarifer           MKLLMCLAAFAAVCFMVDSKNSPAKVQVYSRNPGEFGKANTLICHVSDFHPPDIKIQLMKDGMEIPNADQTDLAFKQGWLFHLTKNVDFTPTGGEKYS 98
Oryzias latipes            MKELFFIAALAAFCCAAFSKTSPPTVQVYSRDPGQYGKENTLICHVSNFHPPDITIRLMEDDQELPEAKQTDLAFKQNWHFHLTKSAPFTPKQGARYT 98
Danio rerio                MRALITFALLCLLYITVQGKVSTPKVHVYSHFPGEYGKPNTLICYVSSFHPPDISIELLKNGQVMSDTKQTDLAFEKGWQFHLTKSVAFTPEKGDEYT 98
Barbus intermedius         MRAIITFALFCVLYITVQAKTSSPKVQVYSHFPGEYGKENTLICHVSGFHPPDITIELLRNGEILPNTQQTDLAFEKGWQFHLTKSVSFTPERGQDYA 98

Oreochromis mossambicus    CKVTHGTVTKDYG       60
Oncorhynchus mykiss        CRVRHLKNLKTYTWEADM 116
Salmo salar                CRVRHLKNLKTYTWEADM 116
Osmerus mordax             CRVRHLTKETEYMWDPDM 116
Hypomesus transpacificus   CKVRHLTKEQDYMWEPDM 116
Esox lucius                CKVRHLSTTKAYTWESNM 116
Sander vitreus             CKVTHGNTVKYYAWEPNM 116
Epinephelus coioides       CKVTHGKTVKDFAWESNM 118
Anoplopoma fimbria         CKVTHGTTVNNYAWEANM 116
Pagrus major               CKVTHGRNMKEYAWVPNM 116
Larimichthys crocea        CRVTHGQKVSNYAWEPNM 116
Seriola quinqueradiata     CRVTHSGKVNEYAWEPNM 116
Lates calcarifer           CRVTHGGVLKDYAWEPNM 116
Oryzias latipes            CKVTHSGLAKDYTWESNM 116
Danio rerio                CSVRHMKETKKFSWEPNM 116
Barbus intermedius         CKVRHMSNTNAYSWEPNM 116
```

Figure A8 - Multisequence alignment of beta-2 microglobulin proteins from different fish species. Species used were: *Oncorhynchus mykiss* (gi: 225703766); *Salmo salar* (gi: 221222048); *Osmerus mordax* (gi: 225706892); *Hypomesus transpacificus* (gi: 227437335); *Esox lucius* (gi: 225717388); *Sander vitreus* (gi: 58198290); *Epinephelus coioides* (gi: 283776086); *Anoplopoma fimbria* (gi: 229367386); *Pagrus major* (gi: 283046520); *Larimichthys crocea* (gi: 81238387); *Seriola quinqueradiata* (gi: 283046524); *Lates calcarider* (gi: 223976163); *Oryzias latipes* (gi: 157278060); *Danio rerio* (gi: 18858327); *Barbus intermedius* (gi: 45433859).

Tubulin alpha

```
Oreochromis mossambicus                                                      ... SETGAGKHVPRAIFVDLEPTVIDEVRTGTYRQLFHPEQLITGK 43
Tetraodon nigroviridis      MGMLERECISIHVGQAGVQIGNACWELYCLEHGIQPDGQMPSDKTIGGGDDSFNTFFSETGAGKHVPRAVFVDLEPTVIDEVRTGTYRQLFHPEQLITGK 100
Salmo salar                     MRECISMHVGQAGAQMGNACWELYCLEHGIQPDGRMPSDKTIGGGDDSFNTFFSETGAGKHVPRAVFVDLEPTVIDEVRTGTYRQLFHPEQLITGK 96
Danio rerio - tubA6             MRECISMHVGQAGAQMGNACWELYCLEHGIQPDGQMPSDKTIGGGDDSFNTFFSETGAGKHVPRAVFVDLEPTVIDEVRTGTYRQLFHPEQLITGK 96
Osmerus mordax                  MRECISMHVGQAGAQMGNACWELYCLEHGIQPDGQMPSDKTIGGGDDSFNTFFSETGAGKHVPRAVFVDLEPTVIDEVRTGTYRQLFHPEQLITGK 96
Xenopus laevis                  MRECISIHVGQAGVQIGNACWELYCLEHGIQPDGQMPSDKTIGGGDDSFNTFFSETGAGKHVPRAVFVDLEPTVIDEVRTGTYRQLFHPEQLITGK 96
Ornithorhynchus anatinus    M-CRARECISIHVGQAGVQIGNACWELYCLEHGIQPDGQMPSDKTIGGGDDSFNTFFSETGAGKHVPRAVFVDLEPTVIDEVRTGTYRQLFHPEQLITGK 99
Mus musculus                    MRECISIHVGQAGVQIGNACWELYCLEHGIQPDGQMPSDKTIGGGDDSFNTFFSETGAGKHVPRAVFVDLEPTVIDEVRTGTYRQLFHPEQLITGK 96
Chionodraco rastrospinosus      MRECISMHVGQAGVQMGNACWELYCLEHGIQPDGQMPSDKTIGGGDDSFNTFFSETGAGKHVPRVVFVDLEPTVIDEVRTGTYRQLFHPEQLITGK 96
Nematostella vectensis          MRECISIHVGQAGVQIGNACWELYCLEHGIQPDGQMPSDKTIGGGDDSFNTFFSETGAGKHVPRAVFVDLEPTVVDEVRTGTYRQLFHPEQLITGK 96

Oreochromis mossambicus     EDAANNYARGHYTVGKEVIDLVLDRTRKLADQCTGLQGFLIFHSFGGGTGSGFTSLLMERLSVDYGKKSKLEFAVYPAPQVSTAVVEPYNSILTTHTTLE 143
Tetraodon nigroviridis      EDAANNYARGHYTIGKEIIDLVLDRIRKLADQCTGLQGFLVFHSFGGGTGSGFTSLLMERLSVDYGKKSKLEFSIYPAPQVSTAVVEPYNSILTTHTTLE 200
Salmo salar                 EDAANNYARGHYTIGKEIIDLVLDRTRKLADQCTGLQGFLIFHSFGGGTGSGFTSLLMERLSVDYGKKSKLEFAVYPAPQVSTAVVEPYNSILTTHTTLE 196
Danio rerio - tubA6         EDAANNYARGHYTIGKEIIDLVLDRTRKLADQCTGLQGFLIFHSFGGGTGSGFTSLLMERLSVDYGKKSKLEFAVYPAPQVSTAVVEPYNSILTTHTTLE 196
Osmerus mordax              EDAANNYARGHYTIGKEIIDLVLDRTRKLADQCTGLQGFLIFHSFGGGTGSGFTSLLMERLSVDYGKKSKLEFAVYPAPQVSTAVVEPYNSILTTHTTLE 196
Xenopus laevis              EDAANNYARGHYTIGKEIIDLVLDRIRKLADQCTGLQGFLVFHSFGGGTGSGFTSLLMERLSVDYGKKSKLEFAIYPAPQVSTAVVEPYNSILTTHTTLE 196
Ornithorhynchus anatinus    EDAANNYARGHYTIGKEIIDLVLDRIRKLADQCTGLQGFLVFHSFGGGTGSGFTSLLMERLSVDYGKKSKLEFSIYPAPQVSTAVVEPYNSILTTHTTLE 199
Mus musculus                EDAANNYARGHYTIGKEIIDLVLDRIRKLADQCTGLQGFLVFHSFGGGTGSGFTSLLMERLSVDYGKKSKLEFSIYPAPQVSTAVVEPYNSILTTHTTLE 196
Chionodraco rastrospinosus  EDAANNYARGHYTVGKEVIDLVLDRTRKLADQCTGLQGFLIFHSFGGGTGSGFTSLLMERLSVDYGKKSKLEFAVYPAPQVSTAVVEPYNSILTTHTTLE 196
Nematostella vectensis      EDAANNYARGHYTVGKEIVDLVLDRIRKLADQCTGLQGFLIFHSFGGGTGSGFSSLLMERLSVDYGKKSKLEFAIYPAPQVSTAVVEPYNSILTTHTTLE 196

Oreochromis mossambicus     HSDCAFMVDNEAIYDICRRNLDIERPTYTNLNRLIGQIVSSITASLRFDGALNVDLTEFQTNLVPYPRIHFPLATYAPVISAEKAYHEQLTVADITNACF 243
Tetraodon nigroviridis      HSDCAFMVDNEAIYDICRRNLDIERPTYTNLNRLIGQIVSSITASLRFDGALNVDLTEFQTNLVPYPRIHFPLATYAPVISAEKAYHEQLSVADITNACF 300
Salmo salar                 HSDCAFMVDNEAIYDICRRNLDIERPTYTNLNRLIGQIVSSITASLRFDGALNVDLTEFQTNLVPYPRIHFPLATYAPVISAEKAYHEQLSVADITNACF 296
Danio rerio - tubA6         HSDCAFMVDNEAIYDICRRNLDIERPTYTNLNRLIGQIVSSITASLRFDGALNVDLTEFQTNLVPYPRIHFPLATYAPVISAEKAYHEQLSVADITNACF 296
Osmerus mordax              HSDCAFMVDNEAIYDICRRNLDIERPTYTNLNRLIGQIVSSITASLRFDGALNVDLTEFQTNLVPYPRIHFPLATYAPVISAEKAYHEQLSVADITNACF 296
Xenopus laevis              HSDCAFMVDNEAIYDICRRNLDIERPTYTNLNRLISQIVSSITASLRFDGALNVDLTEFQTNLVPYPRIHFPLATYAPVISAEKAYHEQLTVADITNACF 296
Ornithorhynchus anatinus    HSDCAFMVDNEAIYDICRRNLDIERPTYTNLNRLIGQIVSSITASLRFDGALNVDLTEFQTNLVPYPRIHFPLATYAPVISAEKAYHEQLSVAEITNACF 299
Mus musculus                HSDCAFMVDNEAIYDICRRNLDIERPTYTNLNRLIGQIVSSITASLRFDGALNVDLTEFQTNLVPYPRIHFPLATYAPVISAEKAYHEQLSVAEITNACF 296
Chionodraco rastrospinosus  HSDCAFMVDNEAIYDICRRNLDIERPTYTNLNRLIGQIVSSITASLRFDGALNVDLTEFQTNLVPYPRIHFPLATYAPVISAEKAYHEQLSVADITNACF 296
Nematostella vectensis      HSDCAFMVDNEAIYDICRRNLDIERPTYTNLNRLIGQIVSSITASLRFDGALNVDLTEFQTNLVPYPRIHFPLATYAPVISAEKAYHEQLTVAEITNACF 296

Oreochromis mossambicus     EPANQMVKCDPRHGKYMACCLLYRGDVVPKDVNSAIATIKTKRTIQFVDWCPTGFKVGINYQPPTVVPGGDLAKVQRAVCMLSNTTAIAEAWARLDHKFD 342
Tetraodon nigroviridis      EPANQMVKCDPRHGKYMACCLLYRGDVVPKDVNAAIATIKTKRTIQFVDWCPTGFKVGINYQPPTVVPGGDLAKVQRAVCMLSNTTAIAEAWARLDHKFD 400
Salmo salar                 EPANQMVKCDPRHGKYMACCLLYRGDVVPKDVNSAIATIKTKRTIQFVDWCPTGFKVGINYQPPTVVPGGDLAKVQRAVCMLSNTTAIAEAWARLDHKFD 396
Danio rerio - tubA6         EPANQMVKCDPRHGKYMACCLLYRGDVVPKDVNSAIATIKTKRSIQFVDWCPTGFKVGINYQPPTVVPGGDLAKVQRAVCMLSNTTAIAEAWARLDHKFD 396
Osmerus mordax              EPANQMVKCDPRHGKYMACCLLYRGDVVPKDVNSAIATIKTKRTIQFVDWCPTGFKVGINYQPPTVVPGGDLAKVQRAVCMLSNTTAIAEAWARLDHKFD 396
Xenopus laevis              EPANQMVKCDPRHGKYMACCLLYRGDVVPKDVNAAIATIKTKRSIQFVDWCPTGFKVGINYQPPTVVPGGDLAKVQRAVCMLSNTTAIAEAWARLDHKFD 396
Ornithorhynchus anatinus    EPANQMVKCDPRHGKYMACCLLYRGDVVPKDVNSAIATIKTKRTIQFVDWCPTGFKVGINYQPPTVVPGGDLAKVQRAVCMLSNTTAIAEAWARLDHKFD 399
Mus musculus                EPANQMVKCDPRHGKYMACCLLYRGDVVPKDVNAAIATIKTKRTIQFVDWCPTGFKVGINYQPPTVVPGGDLAKVQRAVCMLSNTTAIAEAWARLDHKFD 396
Chionodraco rastrospinosus  EPANQLVKCDPRHGKYMACCLLYRGDVVPKDVNSAIATIKTKRTIQFVDWCPTGFKVGINYQPPTVVPGGDLAKVQRAVCMLSNTTAIAEAWARLDHKFD 396
Nematostella vectensis      EPANQMVKCDPRHGKYMACCLLYRGDVVPKDVNAAIATIKTKRTIQFVDWCPTGFKVGINYQPPTVVPGGDLAKVQRAVCMLSNTTAIAEAWARLDHKFD 396

Oreochromis mossambicus     LMYAKRAFVHWYVGEG ...
Tetraodon nigroviridis      LMYAKRAFVHWYVGEGMEEGEFSEAREDMAALEKDYEEVGVDSIEGEGEEE-GEEY 455
Salmo salar                 LMYAKRAFVHWYVGGGMEEGEFSEAREDMAALEKDYEEVGTDSIGEED-EE-GEEY 450
Danio rerio - tubA6         LMYAKRAFVHWYVGEGMEEGEFSEAREDMAALEKDYEEVGTDSVGEED-EE-GEEY 450
Osmerus mordax              LMYAKRAFVHWYVGEGMEEGEFSEAREDMAALEKDYEEVGTDSVGDED-EE-GEEY 450
Xenopus laevis              LMYAKRAFVHWYVGEGMEEGEFSEAREDMAALEKDYEEVGADSADAE--DE-GEEY 449
Ornithorhynchus anatinus    LMYAKRAFVHWYVGEGMEEGEFSEAREDMAALEKDYEEVGVDSVEGEGEEE-GEEY 451
Mus musculus                LMYAKRAFVHWYVGEGMEEGEFSEAREDMAALEKDYEEVGVDSVEGEGEEE-GEEY 455
Chionodraco rastrospinosus  LMYAKRAFVHWYVGEGMEEGEFSEAREDMAALEKDYEEVGTDTVGDEG-EE-GEEY 450
Nematostella vectensis      LMYAKRAFVHWYVGEGMEEGEFSEAREDLAALEKDYEEVGVDSVEGEGEEEEGEEY 452
```

Figure A9 - Multisequence alignment of Tubulin A proteins from fish species. Species used were: *Tetraodon nigroviridis* (gi 47229230): *Salmo salar* (gi: 197632605); *Danio rerio* (gi: 37595424); *Oamerus mordax* (gi: 225707296); *Xenopus laevis* (gi: 288557288); *Ornithorhynchus anatinus* (gi: 149632561); *Mus musculus* (gi:148672205); *Chionodraco rastrospinosus* (gi:10242284); *Nematostella vectensis* (gi: 156394507).

Elongation factor 1 alpha

```
Oreochromis mossambicus      …              VDSGKSTSTGHLIYKCGGIDKRTIEKFEKEAAEMGKGSFKYAWVLDKLKAERERGITIDIALWKFETSKYYVTIIDAPGHRDFIKNMITGTSQADCAVLIVAVGVGEFEA 110
Oreochromis niloticus        MGKEKTHINIVVIGHVDSGKSTSTGHLIYKCGGIDKRTIEKFEKEAAEMGKGSFKYAWVLDKLKAERERGITIDIALWKFETSKYYVTIIDAPGHRDFIKNMITGTSQADCAVLIVAAGVGEFEA 125
Sparus aurata                MGKEKIHINIVVIGHVDSGKSTSTGHLIYKCGGIDKRTIEKFEKEAAEMGKGSFKYAWVLDKLKAERERGITIDIALWKFETSRYYVTIIDAPGHRDFIKNMITGTSQADCAVLIVAAGVGEFEA 125
Pagrus major                 MGKEKIHINIVVIGHVDSGKSTSTGHLIYKCGGIDKRTIEKFEKEAAEMGKGSFKYAWVLDKLKAERERGITIDIALWKFETAKYYVTIIDAPGHRDFIKNMITGTSQADCAVLIVAAGVGEFEA 125
Takifugu rubripes            MGKEKIHINIVVIGHVDSGKSTSTGHLIYKCGGIDKRTIEKFEKEAAEMGKGSFKYAWVLDKLKAERERGITIDIALWKFETSKYYVTIIDAPGHRDFIKNMITGTSQADCAVLIVAAGVGEFEA 125
Seriola quinqueradiata       MGKEKIHINIVVIGHVDSGKSTSTGHLIYKCGGIDKRTIEKFEKEAAEMGKGSFKYAWVLDKLKAERERGITIDIALWRFETSKYYVTIIDAPGHRDFIKNMITGTSQADCAVLIVAAGVGEFEA 125
Hippoglossus hippoglossus    MGKAKTHNNIVVIGHVDSGKSTSTGHLIYKCGGIDKRTIEKFEKEAAEMGKGSFKYAWVLDKLKAERERGITIDIALWKFETTRYCVTIIDAPGHRDFIKNMITGTSQADCAVLIVAAGVGEFEA 125
Oryzias latipes              MGKEKIHINIVVIGHVDSGKSTSTGHLIYKCGGIDKRTIEKFEKEAAEMGKGSFKYAWVLDKLKAERERGITIDIALWKFETSKYYVTIIDAPGHRDFIKNMITGTSQADCAVLIVAAGVGEFEA 125
Carassius auratus            MGKEKTHINIVVIGHVDSGKSTTTGHLIYKCGGIDKRTIEKFEKEAAEMGKGSFKYAWVLDKLKAERERGITIDIALWKFETSKYYVTIIDAPGHRDFIKNMITGTSQADCAVLIVAGGVGEFEA 125
Cyprinus carpio              MGKEKTHINIVVIGHVDSGKSTTTGHLIYKCGGIDKRTIEKFEKEAAEMGKGSFKYAWVLDKLKAERERGITIDIALWKFETSKYYVTIIDAPGHRDFIKNMITGTSQADCAVLIVAGGVGEFEA 125
Danio rerio                  MGKEKTHINIVVIGHVDSGKSTTTGHLIYKCGGIDKRTIEKFEKEAAEMGKGSFKYAWVLDKLKAERERGITIDIALWKFEISKYYVTIIDAPGHRDFIKMNITGTSQADCAVLIVAGGVGEFEA 125
Pimephales promelas          MGKEKIHINIVVIGHVDSGKSTTTGHLIYKCGGIDKRTIEKFEKEAAEMGKGSFKYAWVLDKLKAERERGITIDIALWKFETSKYYVTIIDAPGHRDFIKNMITGTSQADCAVLIVAGGVGEFEA 125
Gadus morhua                 MGKEKTHINIVVIGHVDSGKSTTTGHLIYKCGGIDKRTIEKFEKEAAEMGKGSFKYAWVLDKLKAERERGITIDIALWKFESGKYYVTIIDAPGHRDFIKNMVTGTSQADCAVLIVAGGVGEFEA 125
Solea senegalensis           MGKEKIHINIVVIGHVDSGKSTTTGHLIYKCGGIDKRTIEKFEKEAAEMGKGSFKYAWVLDKLKAERERGITIDIALWKFETNKYYVTIIDAPGHRDFIKNMITGTSQADCAVLIVAAGVGEFEA 125
Lethenteron japonicum        MGKEKIHINIVVIGHVDSGKSTSTGHLIYKCGGIDKRTIEKFEKEAAEMGKGSFKYAWVLDKLKAERERGITIDISLWKFETQKYYVTIIDAPGHRDFIKNMITGTSQADCAVLIVAAGVGEFEA 125

Oreochromis mossambicus      GISKNGQIREHALLAFTLGVKQLIVGVNKMDSTEPPYSQARFEEITKEVSAYIKKIGYNPAAVAFVPISGWHGDNMLETSDKMSWFKGWKIERKDGNASGTTLLEALDAILPPSRPTDKPLRLPL 235
Oreochromis niloticus        GISKNGQTREHALLAFTLGVKQLIVGVNKMDSTEPPYSQARFEEITKEVSAYIKKIGYNPAAVAFVPISGWHGDNMLETSDKMSWFKGWKIERKEGNASGTTLLEALDAILPPSRPTDKPLRLPL 250
Sparus aurata                GISKNGQTREHALLAFTLGVKQLIVGVNKMDSTEPPYSQARFEEITKEVSTYIKKIGYNPASVAFVPISGWHGDNMLETSEKMSWFKGWKVERKEGNANGTTLLEALDAIVPPSRPTDKPLRLPL 250
Pagrus major                 GISKNGQTREHALLAFTLGVKQLIVGVNKMDSTEPPYSQARFEEITKEVSTYIKKIGYNPASVAFVPISGWHGDNMLETSEKMGWFKGWKVERKEGNGSGTTLLEALDAILPPARPTDKPLRLPL 250
Takifugu rubripes            GISKNGQTREHALLAFTLGVKQLIVGVNKMDSTEPPYSEARFNEITKEVSSYIKKIGYNPAAVAFVPISGWHGDNMLEASTKMSWFKGWKVERKEGNASGTTLLDALDAILPPARPTDKPLRLPL 250
Seriola quinqueradiata       GISKNGQTREHALLAFTLGVKQLIVGVNKMDSTEPPYSQARFEEITKEVSTYIKKIGYNPATVAFVPISGWHGDNMLETSEKMGWFKGWKIERKEGNASGTTLLEALDAILPPSRPTEKPLRLPL 250
Hippoglossus hippoglossus    GISKNGQTREHALLAFTLGVKQLIVGINKMDSTEPPYSQPRFEEITKEVSTYIKKIGYNPAAVAFVPISGWHGDNMIETSEKMTWYKGWKVERKEGNGSGTTLLEALDSILPPSRPTDKALRLPL 250
Oryzias latipes              GISKNGQTREHALLAFTLGVKQLIVGVNKMDSTEPPYSQARFEEIQKEVSTYIKKIGYNPAAVAFVPISGWHGDNMLEASDKMSWFKGWKIERKDGNASGTTLLEALDAILPPSRPTDKPLRLPL 250
Carassius auratus            GISKNGQTREHALLAFTLGVKQLIVGVNKMDSTEPPYSQARFEEITKEVSAYIKKIGYNPASVAFVPISGWHGDNMLEASSNMGWFKGWKIERKEGNASGVTLLDALDAILPPSRPTDKPLRLPL 250
Cyprinus carpio              GISKNGQTREHALLAFTLGVKQLIVGVNKMDSTEPPYSQARFEEITKEVSAYIKKIGYNPASVAFVPISGWHGDNMLEPSTNMGWFKGWKIERKEGNANGVTLLDALDAILPPSRPTDKPLRLPL 250
Danio rerio                  GISKNGQTREHALLAFTLGVKQLIVGVNKMDSTEPPYSQARFEEITKEVSAYIKKIGYNPASVAFVPISGWHGDNMLEASSNMGWFKGWKIERKEGNASGTTLLDALDAILPPSRPTDKPLRLPL 250
Pimephales promelas          GISKNGQTREHALLAFTLGVKQLIVGVNKMDSTEPPYSQARFEEITKEVSAYIKKIGYNPASVAFVPISGWHGDNMLEASSNMGWFKGWKVERKEGNASGVTLLDALDAILPPSRPTDKPLRLPL 250
Gadus morhua                 GISSNGQTREHALLAFTLGVKQLIVGINKMDNTEPPYSQSRYEEITKEVSQYIKKIGYNPAAVPFVPISGWHGDNMLEASSKMAWFKGWKVERKDGNASGTTLLEALDAILPPSRPTDKALRLPL 250
Solea senegalensis           GISKNGQTREHALLAYTLGVKQLIVGVNKMDSTEPPYSQKRFEEITKEVSAYIKKIGYNPATVGFVPISGWHGDNMLEASDKMSWFKGWKIERKEGGATGTTLLEALDSILPPSRPTDKPLRLPL 250
Lethenteron japonicum        GISKNGQTREHALLAYTLGVKQLIVGINKMDSTEPPYSADRYNEIVKEVGTYIKKIGYNPAAVGFVPISGWHGDNMLEASTNMPWFKGWKVERKDGNASGVTLLEALDAILPPSRPTDKPLRLPL 250

Oreochromis mossambicus      QDVYKIGGIGTVPVGRVETGILKPGTVVTFAPVNLTTEVKSVEMHHESLPEAVPGDNVGFNVKNVSVKEIRRGYVAGDSKNDPPKGAESFNA        …                          299
Oreochromis niloticus        QDVYKIGGIGTVPVGRVETGILKPGTVVTFAPVNLTTEVKSVEMHHESLPEAVPGDNVGFNVKNVSVKEIRRGYVAGDSKNDPPKGAESFNAQVIILNHPGQIAAGYAPVLDCHTAHIACKFSEL 375
Sparus aurata                QDVYKIGGIGTVPVGRVETGVLKPGMVVTFAPPQLTTEVKSVEMHHESLPEAVPGDNVGFNIKNVSVKEIRRGYVAGDSKNDPPKGADNFNAQVIILNHPGQINARYAPVLDCHTAHIACKFTEL 375
Pagrus major                 QDVYKIGGIGTVPVGRVETGVLKPGMVVTFAPPQLTTEVKSVEMHHESLPEAVPGDNVGFNIKNVSVKEIRRGYVAGDSKNDPPKGADNFNAQVIILNHPGQINAGYAPVLDCHTAHIACKFTEL 375
Takifugu rubripes            QDVYKIGGIGTVPVGRVETGLLKPSMVVTFAPVNLTTEVKSVETHHESLTEAVPGDNVGFNVKDVSVKEIRRGYVAGDSKNDPPKGADNFNAQVIILNHPGQINAGYAPVLDCHTAHIACKFSEL 375
Seriola quinqueradiata       QDVYKIGGIGTVPVGRVETGVLKPGMVVTFAPPNLTTEVKSVEMHHESLPEAVPGDNVGFNVKNVSVKEIRRGYVAGDSKNDPPKAADSFNAQVIILNHPGQINEGYAPVLDCHTAHIACKFKEL 375
Hippoglossus hippoglossus    QDVYKIGGIGTVPVGRVETGLLKPGMLVTFAPPNLTTEVKSVEMHHESMPEAVPGDNVGFNIKNVSVKEIRRGYVAGDSKNDPPKAAASFNAQVIILNHPGQINEGYAPVLDCHTAHIACKFKEL 375
Oryzias latipes              QDVYKIGGIGTVPVGRVETGVLKPGMVVTFAPPNLTTEVKSVEMHHESLPEAVPGDNVGFNIKNVSVKEIRRGYVAGDSKNDPPKAAASFNAQVIILNHPGQINQGYAPVLDCHTAHIACKFNEL 375
Carassius auratus            QDVYKIGGIGTVPVGRVETGVLKPGMVVTFAPANLTTEVKSVEMHHESLAEATPGDNVGFNVKNVSVKDIRRGNVAGDSKNDPPMEAGSFNAQVIILNHPGQISQGYAPVLDCHTAHIACKFAEL 375
Cyprinus carpio              QDVYKIGGIGTVPVGRVETGVLKPGMVVTFAPANLTTEVKSVEMHHESLAEATPGDNVGFNVKNVSVKDIRRGNVAGDSKNDPPMEAGSFNAQVIILNHPGQISQGYAPVLDCHTAHIACKFAEL 375
Danio rerio                  QDVYKIGGIGTVPVGRVETGVLKPGMVVTFAPANVTTEVKSVEMHHESLTEATPGDNVGFNVKNVSVKDIRRGNVAGDSKNDPPMEAANFNAQVIILNHPGQISQGYAPVLDCHTAHIACKFAEL 375
Pimephales promelas          QDVYKIGGIGTVPVGRVETGILKPGMVVTFAPANVTTEVKSVEMHHESLSEASPGDNVGFNVKNVSVKDIRRGNVAGDSKNDPPMEAGSFLAQVIILNHPGQISQGYAPVLDCHTAHIACKFAEL 375
Gadus morhua                 QDVYKIGGIGTVPVGRVETGILKPNMVVTFAPANVTTEVKSVEMHHESLTEALPGDNVGFNVKNVSIKDIRRGNVAGDSKNDPPMQADSFNAQVIILNHPGTISQGYAPVLDCHTAHIACKFNEL 375
Solea senegalensis           QDVYKIGGIGTVPVGRVETGILKPGMVVTFAPPNLTTEVKSVEMHHESLTEALPGDNVGFNIKNVSVKEIRRGNVAGDSKNDPPMAADNFTAQVIILNHPGQIAQGYAPVLDCHTAHIACKFSEL 375
Lethenteron japonicum        QDVYKIGGIGTVPVGRVETGVLKPGMVVTFAPVNVTTEVKSVEMHHEALTEAVPGDNVGFNVKNVSVKDVRRGNVAGDSKNDPPMEAGVFTAQVIVLNHPGQINAGYSPVVDCHTAHIACKFAEL 375

Oreochromis mossambicus
Oreochromis niloticus        VEKIDRRSGKKLEDSPKFVKSGDAAIVKLIPQKPMVVESFSNYPPLGRFAVRDMRQTVAVGVIKSVIPKEVS-GKTTKAAEKAQKGKK 462
Sparus aurata                IEKIDRRSGKKLEDAPKFVKSGDAAIVKLHPQKPMVVEPFSSYPPLGRFAVRDMRQTVAVGVIKAVELKEVS-GKTTKAAEKAQKKK- 461
Pagrus major                 IEKIDRRSGKKLEDAPKFVKSGDAAIVKLHPQKPMVVEPFSSYPPLGRFAVRDMRQTVAVGVIKAVETKEVS-GKTTKAAEKAQKKK- 461
Takifugu rubripes            IEKIDRRSGKKLEDQPKFVKSGDAAIVKLIPQKPMVVEPFSNYPPLGRFAVRDMRQTVAVGVIKSVETKEVS-GKTTKAAEKAQKKK- 461
Seriola quinqueradiata       IEKIDRRSGKKLEDHPKFVKSGDAAIVKLIPQKPMVVEPFSNYAPLGRFAVREMKQTVAVGVIKAVDVKEIS-GKTTKAAEKAQKKK- 461
Hippoglossus hippoglossus    IEKIDRRSGKKLEDSPKFVKSGDAAIVKLVPQKPMVVESFSNYAPLGRFAVRDMRQTVAVGVIKSVDNKDIS-GKTTKAAEKAQKKK- 461
Oryzias latipes              IEKIDRRSGKKLEDNPKFVKSGDAAIVKLIPQKPMVVEPFSNYPPLGRFAVRDMRQTVAVGVIKAVDTKEIS-GKTTKAAEKAQKKK- 461
Carassius auratus            KEKIDRRFGKKLEDNPKALKSGDAAIVEMIPGKPMCVESFSTYPPLGRFAVRDMRQTVAVGVIKSVEKKIGGSGKVTKSAQKAAKTK- 462
Cyprinus carpio              KEKIDRRSGKKLEDNPKALKSGDAAIVEMIPGKPMCVESFSTYPPLGRFAVRDMRQTVAVGVIKSVEKKVGGSGKVTKSAQKAAKTK- 462
Danio rerio                  KEKIDRRSGKELEDNPKALKSGDAAIVEMVPGKPMCVESFSTYPPLGRFAVRDMRQTVAVGVIKSVEKKIGGAGKVTKSAQKAAKTK- 462
Pimephales promelas          KEKIDRRSGKKLEDNPKALKSGDAAIVDMIPGKPMCVESFSTYPPLGRFAVRDMRQTVAVGVIKSVEKKAAGSGKVTKSAQKAAKTK- 462
Gadus morhua                 KEKIDRRSGKKIEDLPKSVKSGDAAIVNMIPSKPMCVESFQSYPPLGRFAVRDMRQTVAVGVIKSVDKKIGGSAKVTKSAKVATKK-- 461
Solea senegalensis           KEKIDRRSGKKLEDNPKALKSGDAAIITMVPGKPMCVESFSQYPPLGRFAVRDMRQTVAVGVIKSVEKKVASGGKVTKSAQKADKKK- 462
Lethenteron japonicum        KEKIDRRSGKKLEDNPKFLKSGDAAMVDMVPGKPMCVESFSDYPPLGRFAVRDMRQTVAVGVIKAVEKKAATTGKVTKSAQKATGKAK 463
```

Figure A10 - Multisequence alignment of Elongation factor 1 alpha proteins from fish species. Species used were: *Oreochromis niloticus* (gi 18181927): *Sparus aurata* (gi: 5923899); *Pagrus major* (gi: 37779018); *Takifugu rubripes* (gi: 83776612); *Seriola quinqueradiata* (gi: 6012184); *Hippoglossus hippoglossus* (gi: 190683996); *Oryzias latipes* (gi: 157278064); *Carassius auratus* (gi: 15528537); *Cyprinus carpio* (gi: 28628941); *Danio rerio* (gi: 68136140); *Pimephales promelas* (gi: 50980344); *Gadus moruha* (gi: 89152378); *Solea senegalensis* (gi: 149166263); *Lethenteron japonicum (gi:* 50284525).

Cathepsin D

```
Oreochromis mossambicus
Danio rerio               MRIAFLLLVAAFFCTSDAIVRIPLKKFRTLRRTLSDSGRSLEELVSSSNSLKYN-LGFPAS-NDPTPETLKNYLDAQYYGEIGLGTPVQTFTVVFDTGSS   98
Clupea harengus           MKFLYLFLFAVFAWTSDAIVRIPLKKFRSIRRTLSDSGLNVEQLLAGTNSLQHN-QGFPSS-NAPTPETLKNYMDAQYYGEIGLGTPVQMFTVVFDTGSS   98
Paralichthys olivaceus    MRNLLLLVLAALALNGDALIRIPLKKFRSIRRELTDSGRPVEELLANEHSLKYN-TGFPSS-NGPTPETLKNYLDAQYYGDIALGTPPQTFSVVFDTGSS   98
Lates calcarifer          MRSLFLVVFAALALSSDALVRIPLKKFRSIRRELTDSGTRLEELLADKHSLKYN-FGFPSS-NGPTPETLKNYLDAQYYGDISLGTPPQTFSVVFDTGSS   98
Takifugu rubripes         MKLLILCVFAALALTNDALVRIPLKKFRSIRRELTDSGRKIEELLADRRINKYN-YGFPTA-GAPTPETLKNYLDAQYYGEIGLGTPPQPFTVVFDTGSS   98
Salmo salar               MKVLYLCLFAALALASDALVRIPLRKFRSIRRTLTDSGRAAEELLAGKEHTKYNNLGFPSSSNGPTPETLKNFMDAQYYGEIGLGTPAQTFTVVFDTGSS  100
Oncorhynchus mykiss       MKVLYLCLFAALALASDALVRIPLRKFRSIRRTLTDSGRAAEELLAGQEHTKYNNLGFPSSSNGPTPETLKNFMDAQYYGEIGLGTPVQTFTVVFDTGSS  100
Chionodraco hamatus       MKMLLLCVFSALALTNDAACSNSLKEIP-FHQTSADRLWEESRGAPGRPSLPEVQLSFPAS-NAPTPETLKNYLDAQYYGEIGLGTPPQPFTVVFDTGSS   98

Oreochromis mossambicus                                           GTSFAIQYGSGSLSGYLSQDTCSIGDISVEKQIFGEAIKQPGVAFIAAKFDGILGMAYPSISVDGVV   67
Danio rerio               NLWVPSVHCSLTDIACLLHHKYNGGKSSTYVKNGTQFAIQYGSGSLSGYLSQDTCTIGDIAVEKQIFGEAIKQPGVAFIAAKFDGILGMAYPRISVDGVP  198
Clupea harengus           NLWLPSIHCSFTDIACLLHHKYNGAKSSTYVKNGTEFAIQYGSGSLSGYLSQDSCTIGDIVVEKQLFGEAIKQPGVAFIAAKFDGILGMAYPRISVDGVP  198
Paralichthys olivaceus    NLWVPSVHCSILDIACWLHHKYNSAKSSTYVKNGTTFAIQYGSGSLSGFLSQDTCTIGDLTVEKQVFGEATKQPGVAFIAAKFDGILGMAYPRISVDGVA  198
Lates calcarifer          NLWVPSVHCSLLDIACLLHHKYNSAKSSTYVKNGTAFAIQYGSGSLSGYLSEDTCTIGDISVEKQLFGEAIKQPGVAFIAAKFDGILGMAYPRISVDGVV  198
Takifugu rubripes         NLWVPSVHCSLLDIACLLHHKYNSAKSSSYVKNGTAFAIRYGSGSLSGYLSQDTCTLGDLAVEKQLFGEAIKQPGIAFIAAKFDGILGMAYPRISVDGVT  198
Salmo salar               NLWVPSVHCSFTDIACLLHHKYNGAKSSTYVKNGTAFAIQYGSGSLSGYLSQDTCTIGGLSIEEQVFGEAIKQPGVAFIAAKFDGILGMAYPRISVDGVA  200
Oncorhynchus mykiss       NLWVPSVHCSFTDIACLLHHKYNGAKSSTYVKNGTAFAIQYGSGSLSGYLSQDTCTIGGLSIEDQGFGEAIKQPGVAFIAAKFDGILGMAYPRISVDGVA  200
Chionodraco hamatus       NLWVPSIHCSLLDIACLLHHKYNSGKSSTYVKNGTAFAIQYGSGSLSGYLSQDTCTIGDLAIDSQLFGEAIKQPGVAFIAAKFDGILGMAYPRISVDGVA  198

Oreochromis mossambicus   PVFDNMMNQKKVEKNVFSFYLNRNPDTEPGGELLLGGTDPKYYDGDFHYANISRQAYWQVHMDGMTVGSQLSLCKGGCEAIVDTGTSLITGPAAEVKALQ  167
Danio rerio               PVFDMMMSQKKVEKNVFSFYLNRNPDTQPGGELLLGGTDPKYYTGDFNYVDISRQAYWQIHMDGMSIGSGLSLCKGGCEAIVDTGTSLITGPAAEVKALQ  298
Clupea harengus           PVFDMMMSQKKVEQNVFSFYLNRNPDTEPGGELLLGGTDPKYYTGDFNYVPVTRQAYWQIHMDGMSIGSQLTLCKDGCEAIVDTGTSLITGPPAEVRALQ  298
Paralichthys olivaceus    PVFDNIMSQKKVEENVFSFYLNRNPDMAPGGELLLGGTDPKYYSGDFNYVNVTRQAYWQIHMGGMGAGSQLTLCKDGCEAIVDTGTSLITGPSAEVKALQ  298
Lates calcarifer          PVFDNIMSQKKVEQNVFSFYLNRNPDTAPGGELLLGGTDPKYYTGDFNYVNITRQAYWQIHMDELVVGTQLSLCKGGCEAIVDTGTSLITGPSAEVKALQ  298
Takifugu rubripes         PVFDNIMSQKKVEKNVFSFYLNRNPDTQPGGELLLGGTDPKYYTGDFDYVNVTRQAYWQIHMDGMSVGSQLSLCKSGCEAIVDTGTSLLTGPSEEVKALQ  298
Salmo salar               PPFDNIMSQKKVEQNVFSFYLNRNPESEPGGELLLGGTDPKYYSGDFQYLNVSRQAYWQVHMDGMGVGSQLSLCKGGCEAIVDTGTSLITGPTAEVKALQ  300
Oncorhynchus mykiss       PPFDNIMSQKKVEQNVFSFYLNRNPDSEPGGELLLGGTDPKYYSGDFQYLDVSRQAYWQIHMDGMGVGSQLSLCKGGCEAIVDTGTSLITGPAAEVKALQ  300
Chionodraco hamatus       PVFDNIMSQKKVEQNVFSFYLNRNPDTEPGGELLLGGTDPKYYTGDFNYVNVTRQAYWQIRVDSMAVGDQLSLCTGGCEAIVDSGTSLITGPSVEVKALQ  298

Oreochromis mossambicus   KAIGAIPLIQGEYLVNCSKIPSLPVITFNVGGQSYTLTGEQYVLLESQAGKTICLSGFMGLDIPPPAGP                                239
Danio rerio               KAIGAIPLVQGEYMVDCKKVPTLPTISFSLGGKVYSLTGEQYILKESQGGHDICLSGFMGLDIPPPAGPLWILGDVFIGQYYTVFDRENNRVGFAKAKSV  398
Clupea harengus           KAIGAIPLIQGEYMIDCKKVPTLPTISFNVGGKTYSLTGEQYVLKESQGGKTICLSGLMGLEIPPPAGPLWILGDVFIGQYYTVFDRESNRVGFAKST    396
Paralichthys olivaceus    KAIGAVPLIQGEYMVSCDKIPSLPVITFNLGGQSYSLTGDQYVLKVSQAGKVICLSGFMGLDIPAPAGPLWILGDVFIGQYYTVFDRENNRVGFAKSK    396
Lates calcarifer          KAIGAIPLIQGEYMVNCDKVPSLPVITFNVGGQSYSLTGEQYILKESQAGKTICLSGFMGLDIPAPAGPLWILGDVFIGQYYTVFDRDNNRVGFAKSK    396
Takifugu rubripes         KAIGAMPLIQGEYMVSCDKIPSLPVITFNIGGKPFSLSGDQYVLKVSQAGKTICLSGFMALDIPAPAGPLWILGDVFIGQYYTVFDRDNNRVGFAKAK    396
Salmo salar               KAIGATPLIQGEYMVNCDKIPTMPDITFNLGGQSYSLTAEQYVLKESQAGKTICLSGFMGLDIPAPAGPLWILGDVFIGQYYTVFDRDNNRVGFAKSK    398
Oncorhynchus mykiss       RAIGATPLIQGEYMVNCDKIPTMPVITFNLGGQSYSLTAEQYVLKESQAGKTICLSGFMGLDIPAPAGPLWILGDVFIGQYYTVFDRDNNRVGFAKSK    398
Chionodraco hamatus       KAIGAFPLIQGEYMVNCDTVPSLPVISFTVGGQVYTLTGEQYILKVTQAGKTMCLSGFMGLDIPAPAGPLWILGDVFMGQYYTVFDRDANRVGFAKAK    396
```

Figure A11 - Multisequence alignment of Cathepsin D proteins. Species used were: *Danio rerio* (gi: 94732449); *Clupea harengus* (gi: 25452827); *Paralichthys olivaceus* (gi: 224460527); *Lates calcarider* (gi: 157644743); *Takifugu rubripes* (gi: 118344558); *Salmo salar* (gi: 197631813); *Oncorhynchus mykiss* (gi: 185132376); Chionodraco hamatus (gi: 3378161).

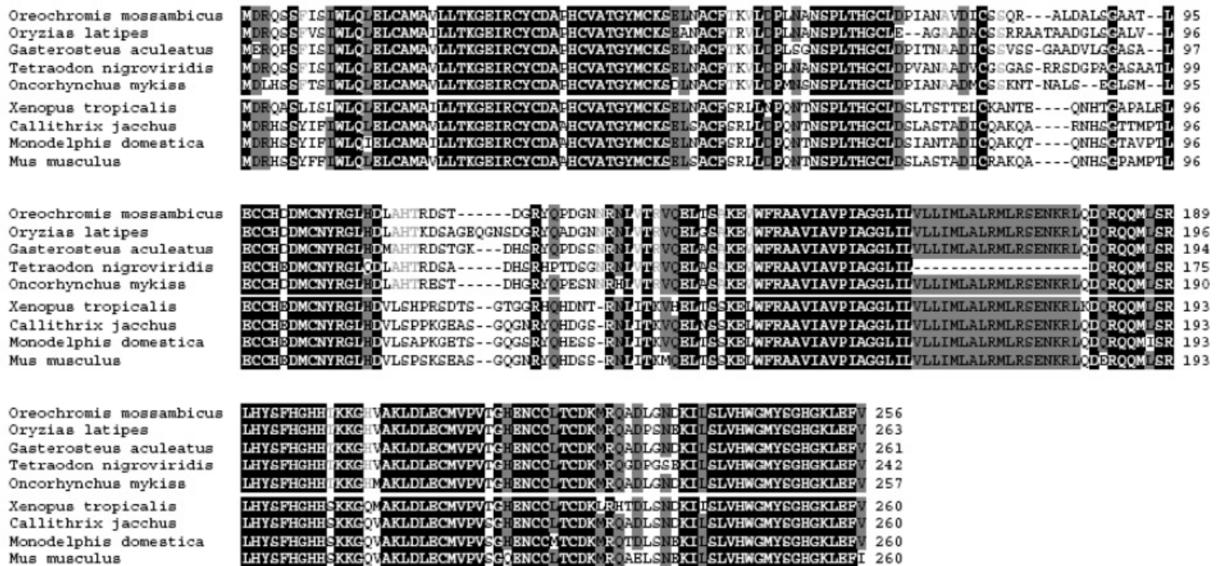
Cathepsin Z

```
Oreochromis mossambicus                                                                                          ...       H 1
Fundulus heteroclitus       MARTAAVCVLLSVLFSAVS--------AGVFFSQKQPCYVRT--PRRNNFGLRTGPLPHEYLNISELPKVWDWRNINGANFVSTTRNQHIPQYCGSCWAH 90
Salmo salar                 --R---ALLLFVPMLSGVF--------AGRYFNEKQPCYRPK--LTK-HNGVRTLPRPHKFLKLNELPKTWDWRNINGTNYVSTTRNQHIPQYCGSCWAH 84
Oncorhynchus mykiss         MAR---SLLLFVPMVSGVF--------AGRYFNEKQPCYGPK--LNK-HNGFRTMPRPYEFLELNELPKAWDWRNINGTNYVSTTRNQHIPQYCGSCWAH 86
Osmerus mordax              MAR---TLVLVLPMLSGVL--------AGRYFNKTKPCYQPT--PRKTHPGVKTTVRPHEPLNLEELPKTWDWRNVNGTNYVSTTRNQHIPQYCGSCWAH 87
Cyprinus carpio             MAR---VVLPYLPALYGVF--------AGRYFNENEPCYRPQ--PRRNLQGVKTGPRPYEYMNLKDLPKAWDWRNIKGVNYVSTTRNQHIPQYCGSCWAH 87
Danio rerio                 MAR---VLFFYLFTLSGVF--------AGRYFNENEPCYKPQ--PRRNLQGVKTGPRPYESMNLKELPKEWDWRNIKGVNYVSTTRNQHIPQYCGSCWAH 87
Tetraodon nigroviridis      -----------MLLLPVVV--------PRHFFSNKPSCYQPV--QRKDDFGVKTSARPHELLNLAQLPKSWDWRNVNGVNYASTTRNQHIPQYCGSCWAH 79
Gallus gallus               MAGPVRRLAALLLVLCGCLY----PCRAGLYVREGQHCYKPA--PRR-APGLRTYPRPHEYLDMAELPQSWDWRNVNGVNYASTTRNQHIPQYCGSCWAH 93
Mus musculus                MASSGSVQQLPLVLLMLLLA---SAARARLYFRSGQTCYHPIRGDQLALLGRRTYPRPHEYLSPADLPKNWDWRNVNGVNYASVTRNQHIPQYCGSCWAH 97
Homo sapiens                MARRG--PGWRPLLLLVLLA---GAAQGGLYFRRGQTCYRPLRGDGLAPLGRSTYPRPHEYLSPADLPKSWDWRNVDGVNYASITRNQHIPQYCGSCWAH 95
Monodelphis domestica       MAPPRGLLTLQLQLLLLVLWGCWSPSSAGIYYQKGQSCYRPLEGLGLSTCRLGTYPRPHEYMARSSLPKAWDWRNVNGVNYASITRNQHIPQYCGSCWAH 100

Oreochromis mossambicus     GSTSAMADRINIKRKGAWPSAYLSVQNVIDCGDAGSCHGGDHTGVWEYANSHGIPDETCNNYQAKDQDCNKFNQCGTCTTFGVCNTVSNYTLWKVGDYGS 101
Fundulus heteroclitus       GSTSAMADRINIKRKGAWPSAYLSVQHVIDCAGAGTCHGGDHSGVWEYASTHGIPDETCNNYQAKDQACKPFNQCGTCTTFGQCNIVKNYTMWKVGDYGS 190
Salmo salar                 GSTSAMADRINIKRKGAWPSAYLSVQNVVDCGEAGSCHGGDHSGVWEYANKHGIPDETCNNYQAKDQKCKPFNQCGTCTTFGVCNIIQNYTLWKVGDFGA 184
Oncorhynchus mykiss         GSTSAMADRINIKRKGAWPSAYLSVQNVVDCGEAGSCHGGDHTGVWEYANKHGIPDETRNNYQAIDQTCKLFNQCGTCTTFGVCNVIQNYILWKVADFGP 186
Osmerus mordax              GSTSAMADRINIKRKGAWPSAYLSVQNVVDCGDAGSCHGGDHSGVWEYANKHGIPDETCNNYQAKDQKCKPFNECGTCTTFGVCNVIQNYTLWKVGDFGS 187
Cyprinus carpio             GSTSAMADRINIKRKAVWPSAYLSVQNVIDCGNAGSCNGGDHSGVWEYAHSNGIPDETCNNYQAKDQDCNSFNQCGTCTTFGVCNVVKNFTLWKVGDYGS 187
Danio rerio                 GSTSALADRINIKRKAAWPSAYLSVQNVIDCGDAGSCSGGDHSGVWEYAHNKGIPDETCNNYQAKDQDCKPFNQCGTCTTFGVCNIVKNFTLWKVGDYGS 187
Tetraodon nigroviridis      GSTSAMSDRINIKRRGAWPSAYLSVQHVLDCSDSGTCHGGDHGGVWTYAHKHGIPDETCNNYQARDQKCKPFNACGTCTTFKKCNTLKNYTLWKVGDFGS 179
Gallus gallus               GSTSALADRINIKRKGAWPSAYLSVQNVIDCANAGSCEGGDHTGVWMYAHDHGIPDETCNNYQAKNQKCKKFNQCGTCVTFGECHVIKNYTLWKVADYGA 193
Monodelphis domestica       GTTSALADRINIKRKGAWPSTLLSVQHVIDCGNAGSCEGGMDIPVWEYAHMHGIPDETCNNYQAKDQECDKFNECGTCNEFHNCYSIKNYTLWKVGDYGT 200
Mus musculus                GSTSAMADRINIKRKGAWPSILLSVQNVIDCGNAGSCEGGNDLPVWEYAHKHGIPDETCNNYQAKDQDCDKFNQCGTCTEFKECHTIQNYTLWRVGDYGS 197
Homo sapiens                ASTSAMADRINIKRKGAWPSTLLSVQNVIDCGNAGSCEGGNDLSVWDYAHQHGIPDETCNNYQAKDQECDKFNQCGTCNEFKECHAIRNYTLWRVGDYGS 195

Oreochromis mossambicus     VSGREKMMAEIYANGPISCGIMATDKLDAYTGGLYSEYQDEAYINHIVSVAGWGV-EDDVEYWIVRNSWGEPWGEKGWLRIVTSAYKGGSGSKYNLAVEE 199
Fundulus heteroclitus       VSGREKMMAEIYARGPISCGIMATDQLDAYTGGLYSEYQEEAFINHIVSVAGWGV-EDGVEYWVVRNSWGEPWGEKGWLRIVTSAYKGGSGSKYNLALEE 289
Salmo salar                 INGREKMMAEIYAGGPISCGIMATDKLDAYTGGLYSEYIEEPFINHIVSVAGWGMDENGVEEWIVRNSWGEPWGEKGWLRIVTSAYKGGSGSQYNLALEE 284
Oncorhynchus mykiss         VSGREKMMAEIYAGGPISCGIMATDKLDAYTGGLYSEYIQEPYINHIVSVAGWGLDENGVEEWIVRNSWGEPWGEKGWLRIVTSAYKGGSGSQYNLALEE 286
Osmerus mordax              VSGREKMMAELYAGGPISCGIMATEKLDAYTGGLYSEYVESPSINHIVSVAGWGV-ENGVEYWIVRNSWGEPWGEKGWLRIVTSAYKGGSGSQYNLALEE 286
Cyprinus carpio             VSGLDKMKAEIYSGGPISCGIMATDKLDAYTGGVYSEYVQDPYINHIISVAGWGVDENGVEYWVVRNSWGEPWGEKGWLRIVTSAYKGGSGSQYNLAIEE 287
Danio rerio                 ASGLDKMKAEIYSGGPISCGIMATDKLDAYTGGLYSEYVQEPYINHIVSVAGWGVDENGVEEWVVRNSWGEPWGEKGWLRIVTSAYKGGSGSQYNLAIEE 287
Tetraodon nigroviridis      ISGRANMMAEIHSGGPISCGIMATEKLDDYTGGLYSEYVESPEINHIVSVAGWGV-DNGTEYWIVRNSWGEPWGERGWLRIVTSLYKGGSGSKYNLALED 278
Gallus gallus               VSGREKMMAEIYANGPISCGIMATEKLDAYTGGLYTEYNPSPTVNHIVSVAGWGV-ENGTEYWIVRNSWGEPWGERGWLRIVTSAYKGGRGAEYNLAVEE 292
Monodelphis domestica       LSGREKMMAEIYANGPISCGIMATEALDNYTGGIYFEYNPQPMINHIISVAGWGVSDNGTEYWIVRNSWGEPWGEKGWLRIVTSRYKDGQGSNYNLAIEE 300
Mus musculus                LSGREKMMAEMYANGPISCGIMATEMMSNYTGGIYAEHQDQAVINHIISVAGWGVSNDGIEYWIVRNSWGEPWGEKGWMRIVTSTYKGGTGDSYNLAIES 297
Homo sapiens                LSGREKMMAEIYANGPISCGIMATERLANYTGGIYAEYQDTTYINHVVSVAGWGIS-DGTEYWIVRNSWGEPWGERGWLRIVTSTYKDGKGARYNIAIEE 294

Oreochromis mossambicus                ...  199
Fundulus heteroclitus       DCMYGDPIVPKDYL 303
Salmo salar                 DCMYGDPIVPKTYM 298
Oncorhynchus mykiss         DCMYGDPIVPKTYM 300
Osmerus mordax              DCMYGDPIVPKSYL 300
Cyprinus carpio             NCMYGDPILPKNYL 301
Danio rerio                 DCMYGDPILPKNYL 301
Tetraodon nigroviridis      DCMYGDVIIP     288
Gallus gallus               DCAYGDVILP     302
Monodelphis domestica       TCSFGDPIV      309
Mus musculus                ACTFGDPLFI     307
Homo sapiens                HCTFGDPIV      303
```

Figure A12- Multisequence alignment of cathepsin Z proteins. Species used were: *Fundulus heteroclitus* (gi: 37907340); *Salmo salar* (gi: 224613522); *Oncorhynchus mykiss* (gi: 185135804); *Osmerus mordax* (gi: 225706784); *Cyprinus carpio* (gi: 61661416); *Danio rerio* (gi: 54400588); *Tetraodon nigroviridis* (gi: 47222865); *Gallus gallus* (gi: 118100844); *Monodelphis domestica* (gi: 126303150); *Mus musculus* (gi: 11066226); *Homo sapiens* (gi: 3294548).

ANNEX 6 – Multisequence alignment, matrix of identities and philogenetic trees for *BAMBI* and *Gremlin* genes

Bone morphogenetic protein and activin membrane-bound inhibitor



<u>Figure A13 - Multisequence alignment of *BAMBI* proteins</u>. Species used were: *Oryzias latipes* (ENSORLT00000005643); *Gasterosteus aculeatus* (ENSGACT00000002180); *Tetraodon nigroviridis* (gi: 47225123); *Oncorhynchus mykiss* (gi: 238231547); *Xenopus tropicalis* (gi: 56118430); *Callithrix jacchus* (gi: 296206382); *Monodelphis domestica* (gi: 126341210); *Mus musculus* (gi: 33468995).

	O. mossambicus	*G. aculeatus*	*T. nigroviridis*	*O. mykiss*	*D. rerio*	*O. latipes*	*X. tropicalis*	*A. carolinensis*	*G. gallus*	*M. musculus*
O. mossambicus										
G. aculeatus	92									
T.nigroviridis	87	87								
O. mykiss	90	89	83							
D. rerio	78	78	76	78						
O. latipes	91	84	83	85	75					
X. tropicalis	76	73	71	73	71	73				
A. carolinensis	76	72	71	75	71	75	83			
G. gallus	77	77	72	75	72	75	83	89		
M. musculus	77	73	72	75	72	73	81	87	91	

Figure A14 – Matrix of identity percentages between *BAMBI* proteins. Species used were: O. mossambicus; *Oryzias latipes* (ENSORLT00000005643); *Gasterosteus aculeatus* (ENSGACT00000002180); *Tetraodon nigroviridis* (gi: 47225123); *Oncorhynchus mykiss* (gi: 238231547); *Danio rerio* (gi: 18858339); *Xenopus tropicalis* (gi: 56118430); *Anolis carolinensis* (ENSACAG00000007580); *Gallus gallus* (gi: 306482572); *Mus musculus* (gi: 33468995). The bold line separates the fish species from the other vertebrates.

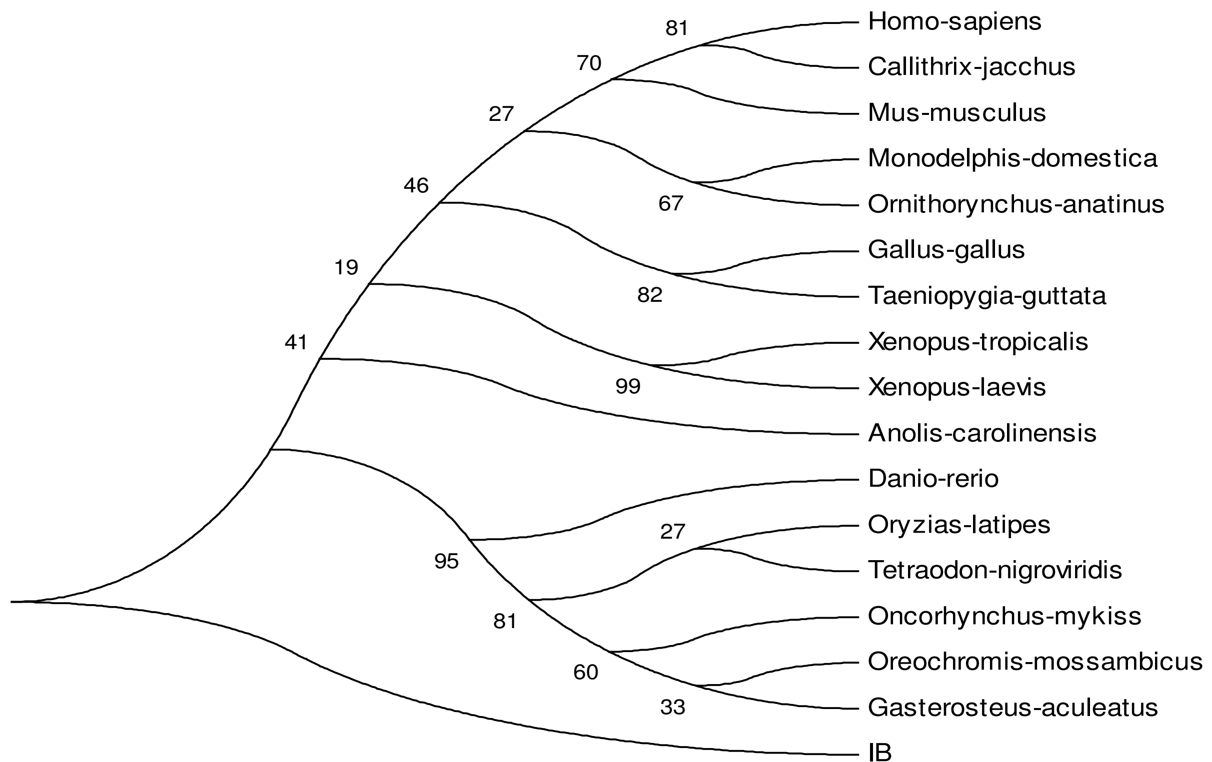


Figure A15 – Phylogenetic tree representing evolutionary relationships of taxa for *BAMBI* protein.
The evolutionary history was inferred using the Neighbor-Joining method [36]. The optimal tree with the sum of branch length = 2.23290180 is shown. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (1000 replicates) is shown next to the branches [37]. The evolutionary distances were computed using the Poisson correction method [38] and are in the units of the number of amino acid substitutions per site. The analysis involved 17 amino acid sequences. All positions containing gaps and missing data were eliminated. There were a total of 203 positions in the final dataset. Evolutionary analyses were conducted in MEGA4 [39].
Species used were: *O. mossambicus*; *Oryzias latipes* (ENSORLT00000005643); *Gasterosteus aculeatus* (ENSGACT00000002180); *Tetraodon nigroviridis* (gi: 47225123); *Oncorhynchus mykiss* (gi: 238231547); *Xenopus tropicalis* (gi: 56118430); *Callithrix jacchus* (gi: 296206382); *Monodelphis domestica* (gi: 126341210); *Mus musculus* (gi: 33468995); *Homo sapiens* (gi: 6912534); *Gallus gallus* (gi: 306482572); *Danio rerio* (gi: 18858339); *Xenopus laevis* (gi:148226467); *Taeniopygia guttata* (gi:197128229); *Ornithorhynchus anatinus* (gi:149634738); *Anolis carolinensis* (ENSACAG00000007580). The BMP IB protein of the fish species *O. mossambicus* (gi: 317415913) was used as an out group in this analysis

GREMLIN

```
Oreochromis mossambicus
Takifugu rubripes                                                              MAN--STR---IFCSMVFIIGLLFSPVD-CKRNRGSQGAIPHPDKNNPN 43
Tetraodon nigroviridis                                                         MAN--TTR---IFCSVAFIIGLLSFPVD-CKRNRGSQGAIPHPDKNNPN 43
Gasterosteus aculeatus                                                         MAG--STRIFCIFCSVLFAVGLLSLPVD-SKRNRGSQGAIPHPDKTNPN 46
Danio rerio                                MGRFTDRFFSKISPGRSSSTLTVMPASCASSSGWILCAVLLSSALLWCPTE-SKRNR---GAIPHPDKNTPN 68

Ambystoma mexicanum                                                                MFRLVYAVGALLLLYGLMLPVAEGEKKTRGSQGAIPPPDKAQPN 44
Xenopus laevis                                                                     MNCLVYALGSLFLLSGLLLPSSEGKKKVSGSQGAIPPPDKGQPN 44
Callithrix jacchus                                                                 MSRTAYTVGALLLLLGTLLPAAEGKKK--GSQGAIPPPDKAQHN 42
Mus musculus             MPDTRCALRRPGDPEYRGSGALLRLPPASSPSGSSRGLTQRHSSTAKNLRTHGSDRMNRTAYTVGALLLLLGTLLPTAEGKKK--GSQGAIPPPDKAQHN 98


Oreochromis mossambicus                                  QEALHVTERQYLKRDWCKTQPLKQTIHEEGCVSRTIINRFCYGQCNSFYIPRHIRREEGAFQSCS  65
Takifugu rubripes        ESEQHPQPGRAGPAPRQRH---GSTSPADEVLESSQEALHVTERQYLKRDWCKTQPLKQTIHEEGCVSRTIINRFCYGQCNSFYIPRHIRREEGAFQSCS 140
Tetraodon nigroviridis   ESEQHPQPGRAGPAPRQRH---GSTSPADEVLESSQEALHVTERQYLKRDWCKTQPLKQTIHEEGCVSRTIINRFCYGQCNSFYIPRHIRREEGAFQSCS 140
Gasterosteus aculeatus   ESEQQPQTPQAGPGPRQRQ---GSSSPADEVLESSQEALHVTERQYLKRDWCKTQPLKQTIHEEGCVSRTIINRFCYGQCNSFYIPRHIRREEGAFQSCS 143
Danio rerio              ESSRSPQQQESASGSRAR------SRGSEEVLESSQEALHVTERRYLKRDWCKTQPLKQTIHEEGCISRTIINRFCYGQCNSFYIPRHVRREEGAFQSCS 162

Ambystoma mexicanum      DSEQTQQQP-PGGRARGR--GKGTAMPAEEVLESSQEALHVTERKYLKRDWCKTQPLKQTIHEEGCNSRTILNRFCYGQCNSFYIPRHVRREEGSFQSCS 141
Xenopus laevis           DSEQGQAQP--GDRVRGK--GKGQALAAEEVLESSQEALHVTERKYLKRDWCKTQPLKQTIHEIGCNSRTIINRFCYGQCNSFYIPRHIRREEGSFQSCS 140
Callithrix jacchus       DSEQTQSPQQPGSRNRGRGQGRGTAVPGEEVLESSQEALHVTERKYLKRDWCKTQPLKQTIHEEGCNSRTIINRFCYGQCNSFYIPRHIRKEEGSFQSCS 142
Mus musculus             DSEQTQSPPQPGSRTRGRGQGRGTAMPGEEVLESSQEALHVTERKYLKRDWCKTQPLKQTIHEEGCNSRTIINRFCYGQCNSFYIPRHIRKEEGSFQSCS 198


Oreochromis mossambicus  FCKPKRFTTMTPTLNCPDQQPPTKKKRI                 93
Takifugu rubripes        FCKPKRFTTMTPTLNCPDQQPPTKKKRIQRVKQCRCISIDLD 182
Tetraodon nigroviridis   FCKPKRFTTMTPTLNCPDQQPPTKKKRIQRVKQCRCISIDLD 182
Gasterosteus aculeatus   FCKPKRFTTMTPTLNCPDQQPPTKKKRIQRVKQCRCISIDLD 185
Danio rerio              FCKPKRFTTMSPTLSCPDQQPPTRKKRVQRVKQCRCISIELD 204

Ambystoma mexicanum      FCKPKKFTTMTVTLNCPELQPPTKKKRITRVKQCRCISIDLE 183
Xenopus laevis           FCKPKKFTTMVVTLNCPELQPPTKKKRITRVKQCRCISIDLD 182
Callithrix jacchus       FCKPKKFTTMMVTLNCPELQPPTKKKRVTRVKQCRCISIDLD 184
Mus musculus             FCKPKKFTTMMVTLNCPELQPPTKKKRVTRVKQCRCISIDLD 240
```

Figure A16 - Multisequence alignment of gremlin proteins. Species used were**:** *Takifugu rubripes* (ENSTRUT00000036344); *Tetraodon nigroviridis* (gi: 47230263); *Gasterosteus aculeatus* (ENSGACG00000006103); *Danio rerio* (gi: 189531067); *Ambystoma mexicanum* (gi: 242277418); *Xenopus laevis* (gi: 148225953); *Callithrix jacchus* (gi: 296214280); *Mus musculus* (gi: 74149177).

	O. mossambicus	*T. rubripes*	*T.nigroviridis*	*G. aculeatus*	*D. rerio*	*A. mexicanum*	*X. laevis*	*C. jacchus*	*M. musculus*	*G. gallus*
O. mossambicus										
T. rubripes	100									
T.nigroviridis	100	97								
G. aculeatus	100	88	89							
D. rerio	92	72	72	74						
A. mexicanum	90	68	69	69	64					
X. laevis	90	71	70	68	64	85				
C. jacchus	89	69	67	67	65	84	81			
M. musculus	89	69	67	67	64	85	82	97		
G. gallus	88	69	69	71	65	85	80	83	83	

Figure A17 – Matrix of identity percentages between GREMLIN proteins. Species used were: *Takifugu rubripes* (ENSTRUT00000036344); *Tetraodon nigroviridis* (gi: 47230263); *Gasterosteus aculeatus* (ENSGACG00000006103); *Danio rerio* (gi: 189531067); *Ambystoma mexicanum* (gi: 242277418); *Xenopus laevis* (gi: 148225953); *Callithrix jacchus* (gi: 296214280); *Mus musculus* (gi: 74149177); Gallus gallus (gi: 45384474). The bold line separates the fish species from the other vertebrates.

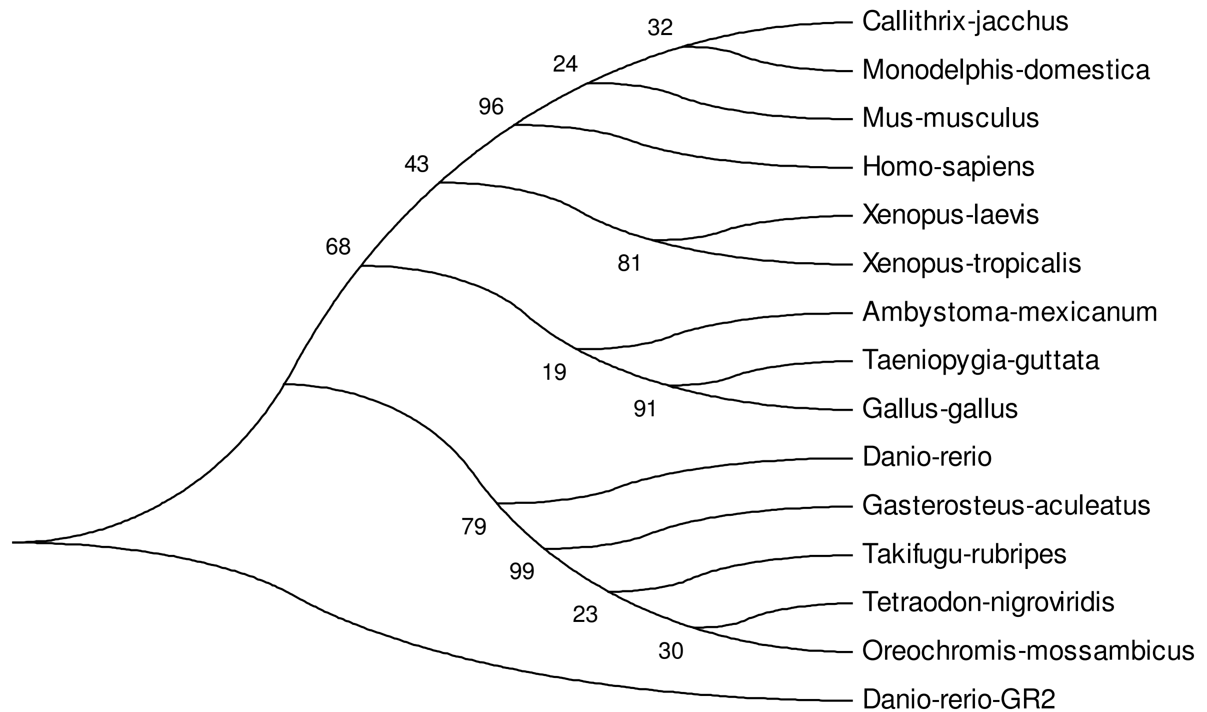


Figure A18 – Phylogenetic tree representing evolutionary relationships of taxa for GREMLIN protein. The evolutionary history was inferred using the Neighbor-Joining method [36]. The optimal tree with the sum of branch length = 2.28517694 is shown. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (500 replicates) are shown next to the branches [37]. The evolutionary distances were computed using the Poisson correction method [38] and are in the units of the number of amino acid substitutions per site. All positions containing gaps and missing data were eliminated from the dataset (Complete deletion option). There were a total of 93 positions in the final dataset. Phylogenetic analyses were conducted in MEGA4 [39].

Species used were: *Oreochromis mossambicus*; *Takifugu rubripes* (ENSTRUT00000036344); *Tetraodon nigroviridis* (gi: 47230263); *Gasterosteus aculeatus* (ENSGACG00000006103); *Danio rerio* (gi: 189531067); *Ambystoma mexicanum* (gi: 242277418); *Xenopus laevis* (gi: 148225953); *Callithrix jacchus* (gi: 296214280); *Mus musculus* (gi: 74149177); Gallus gallus (gi: 45384474); *Taeniopygia guttata* (gi:224051346); *Monodelphis domestica* (gi: 126278318); *Xenopus tropicalis (*gi: 154147688) ; *Homo sapiens* (gi: 7019349); The *GREMLIN 2* protein of the fish species *Danio rerio (gi:* 62955385) was used as an out group in this analysis.

ANNEX 7: Standard and Melting curves for all genes

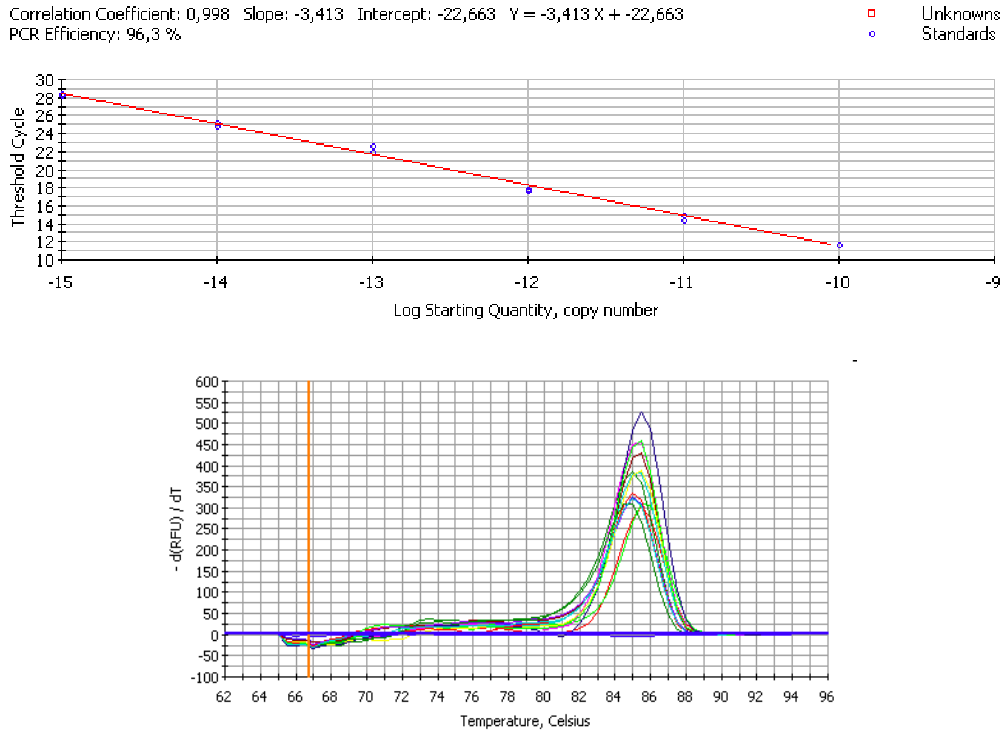


Figure A19 – Standard curve and Melting curve of the gene ***CTSD***

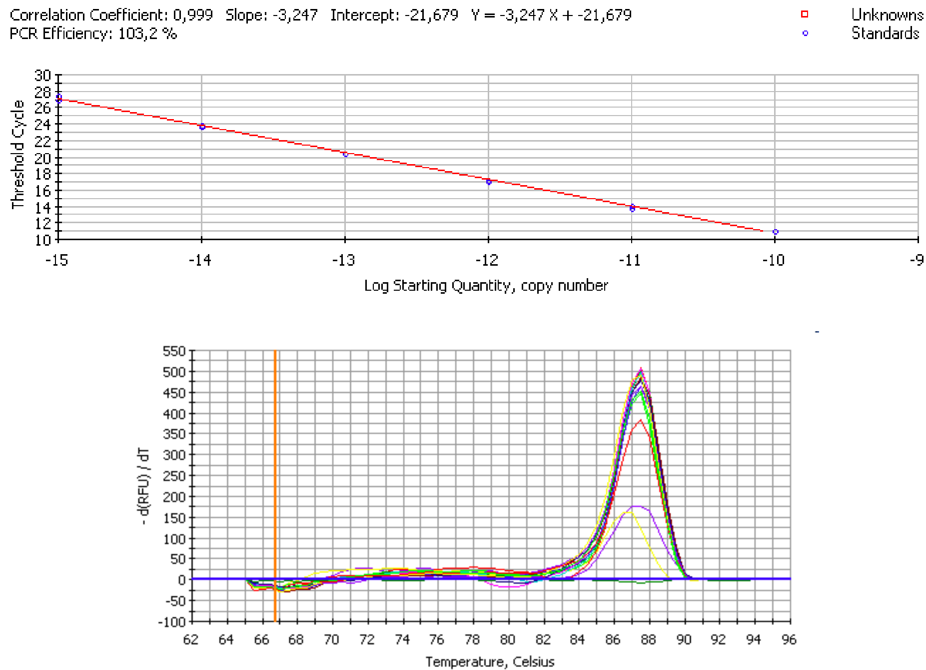


Figure A20 – Standard curve and Melting curve of the gene ***CTSZ***

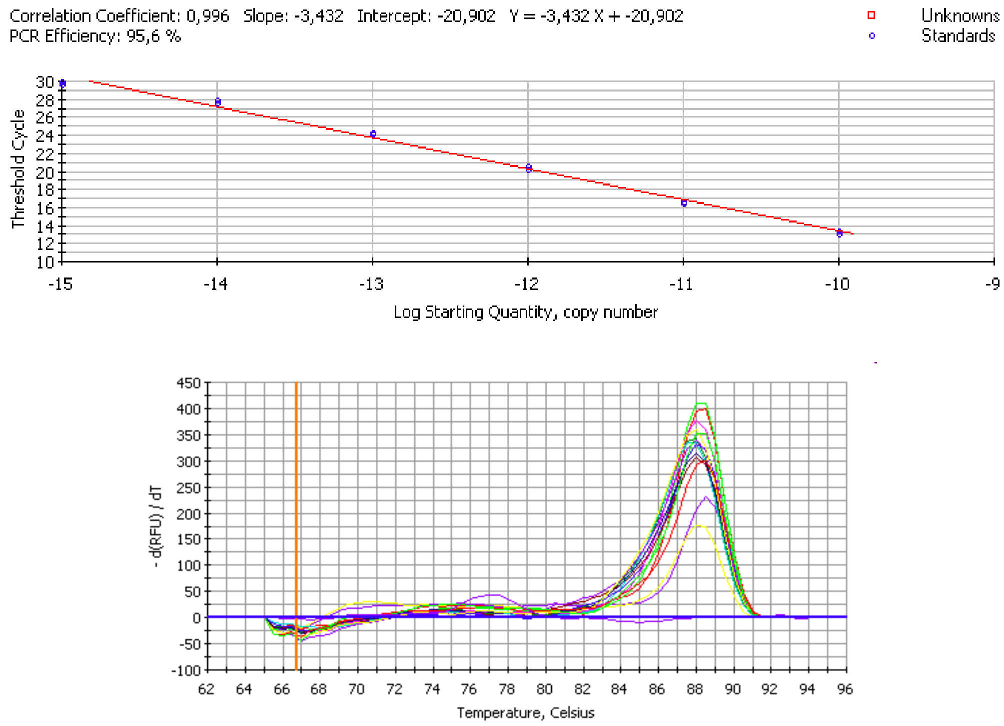


Figure A21 – Standard curve and Melting curve of the gene ***β-actin***

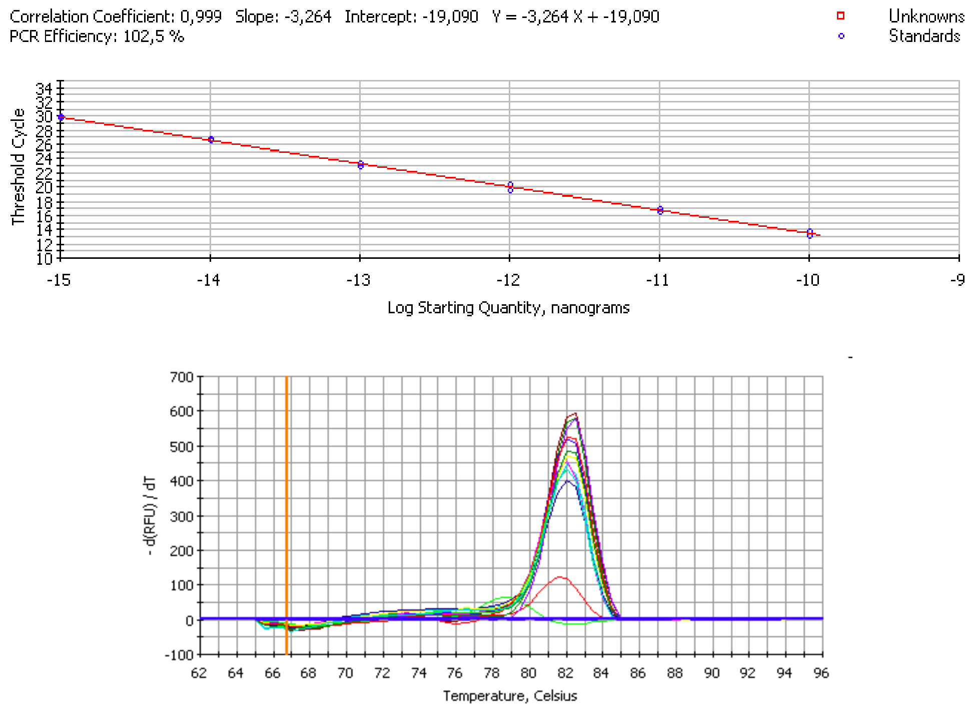


Figure A22 – Standard curve and Melting curve of the gene ***HPRT1***

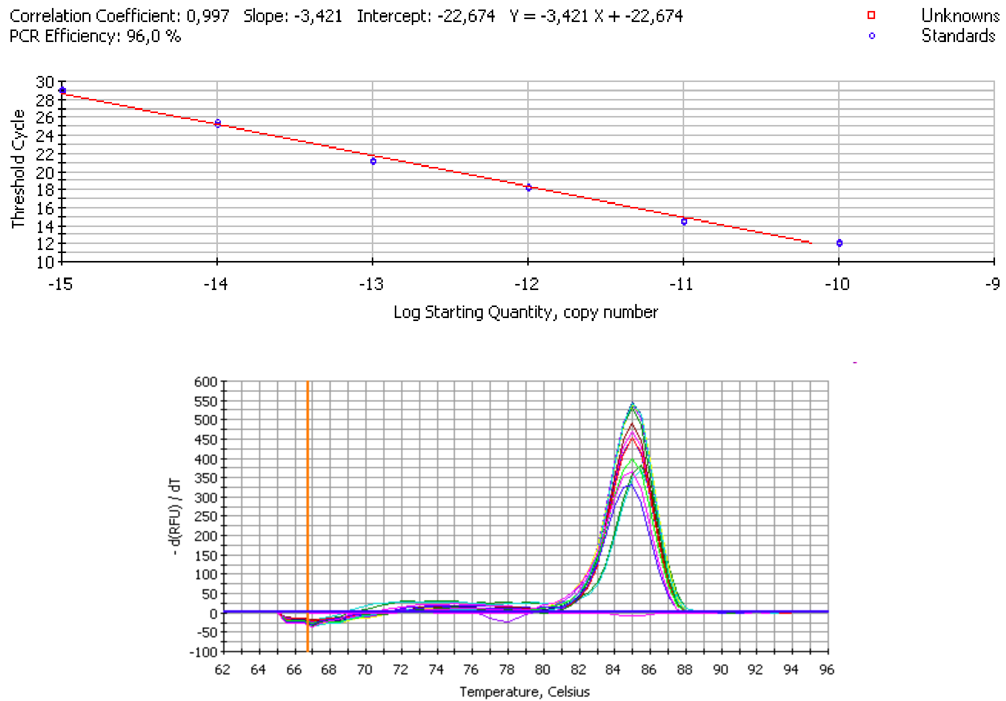


Figure A23 – Standard curve and Melting curve of the gene ***TBP***

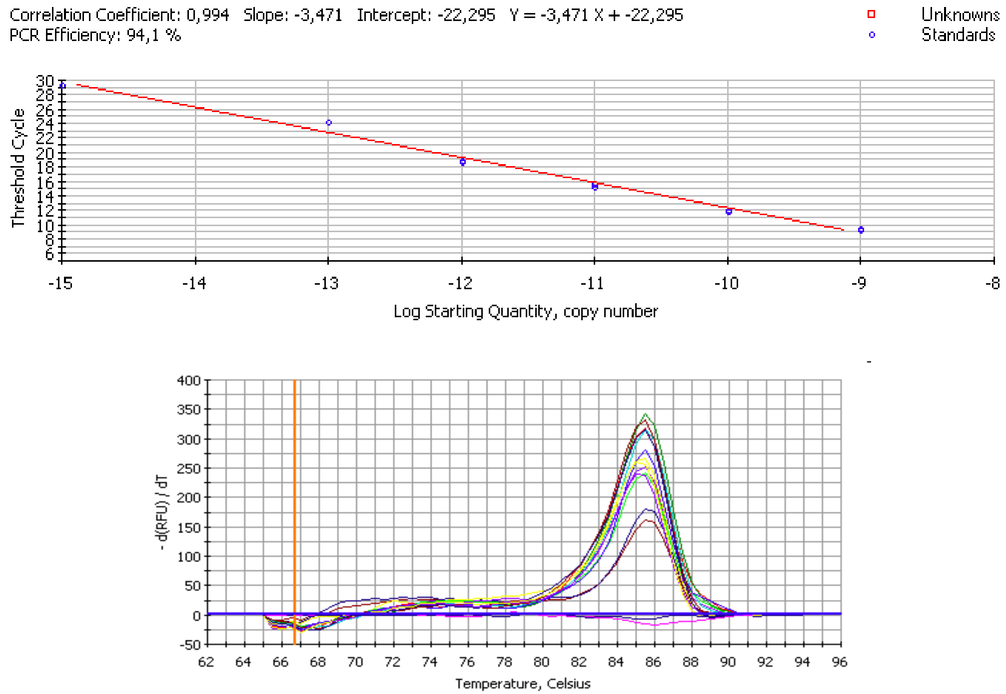


Figure A24 – Standard curve and Melting curve of the gene ***Tub A***

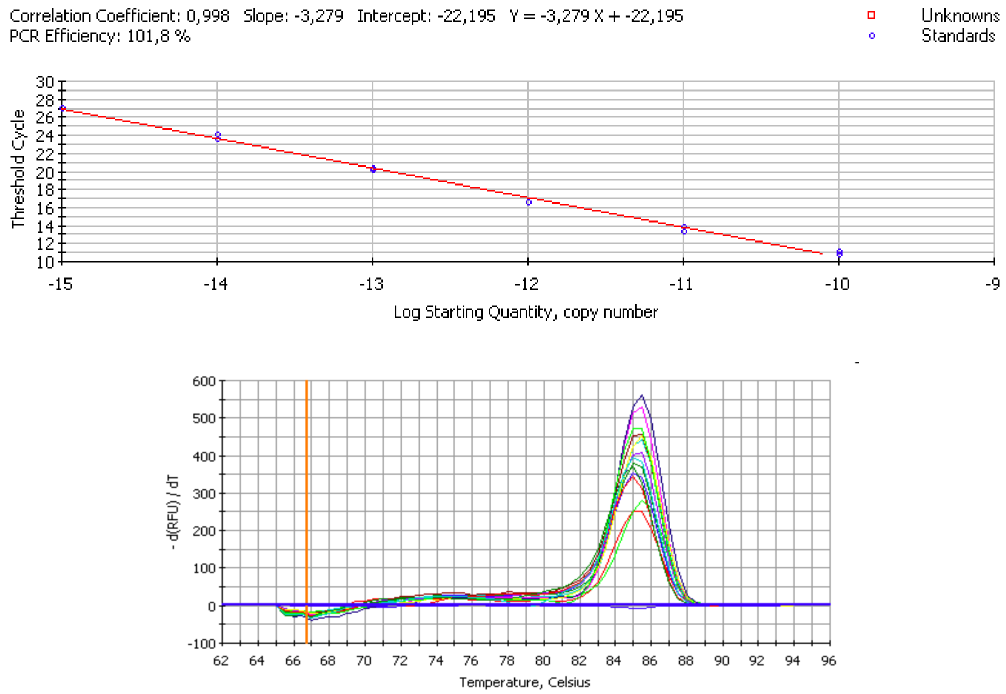


Figure A25 – Standard curve and Melting curve of the gene ***GADPH***

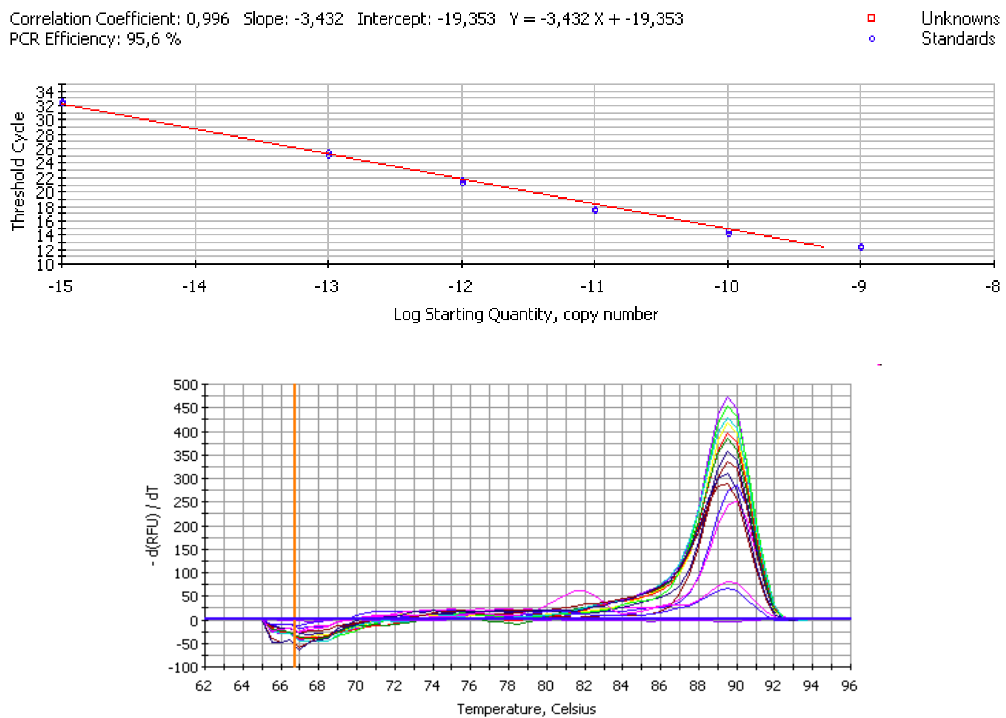


Figure A26 – Standard curve and Melting curve of the gene ***Efa1***

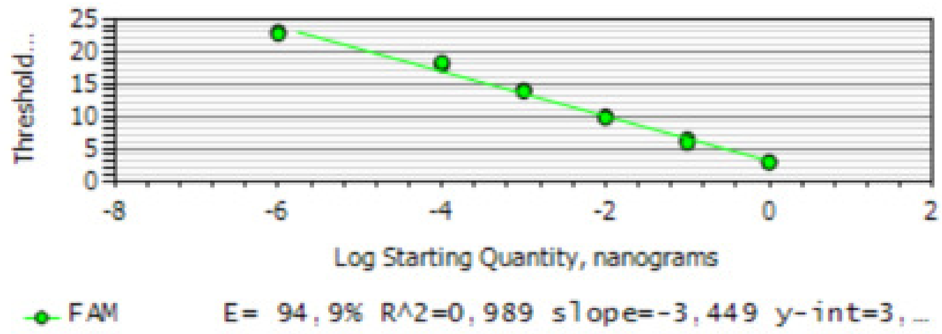


Figure A27 – Standard curve of the gene ***18S***

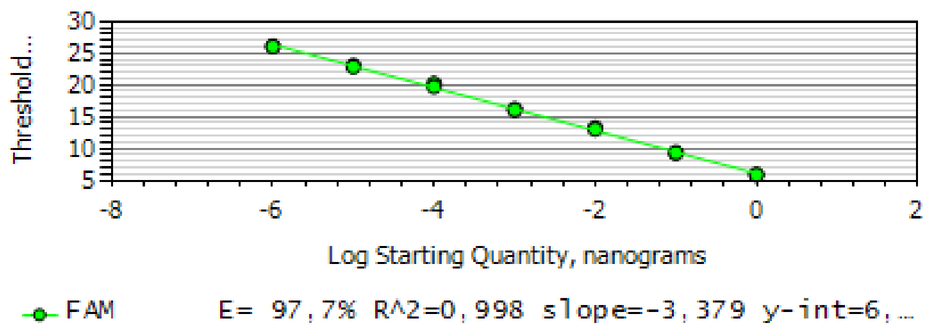


Figure A28 – Standard curve of the gene ***B2m***

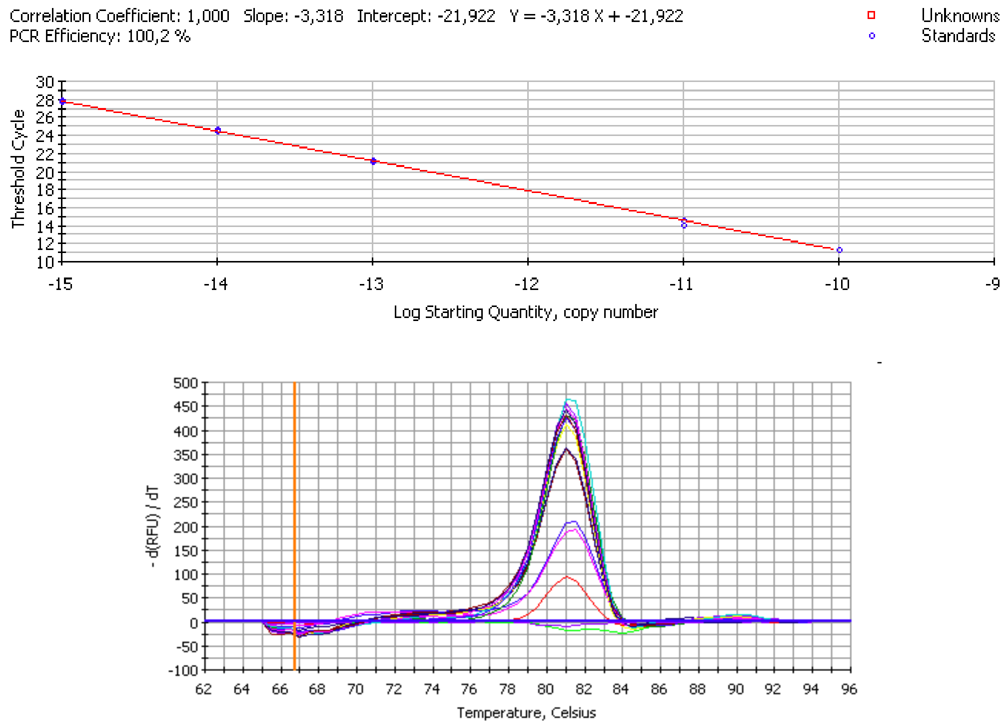


Figure A29 – Standard curve and Melting curve of the gene ***BAMBI***

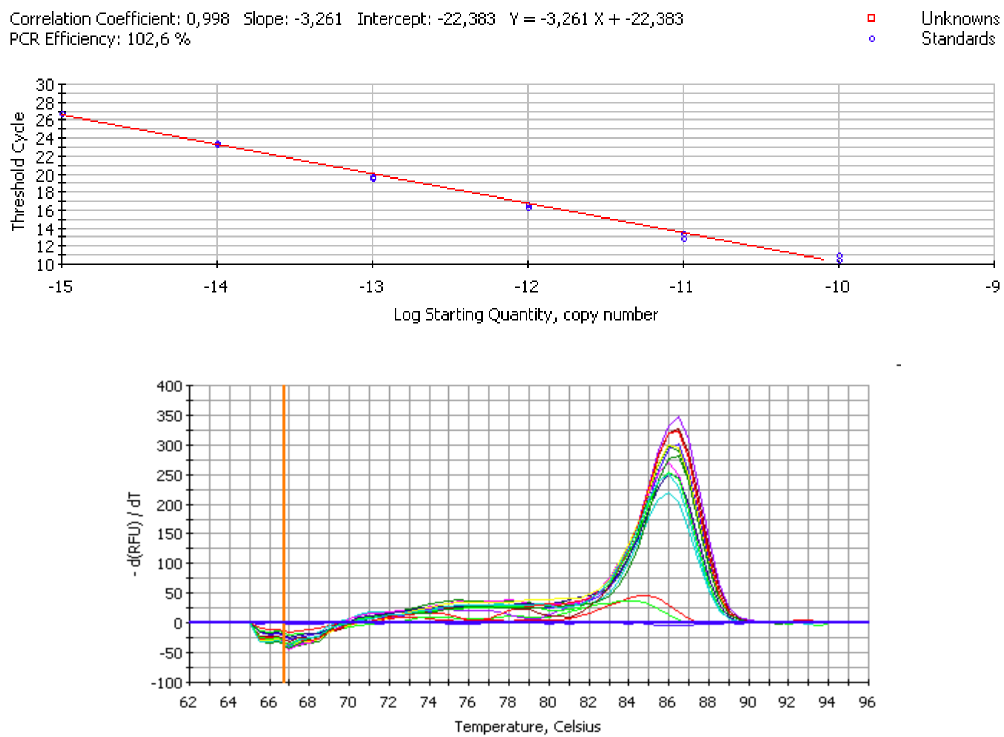


Figure A30 – Standard curve and Melting curve of the gene ***Gremlin***