



HAL
open science

Evolution of sex hormone binding globulins reveals early gene duplication at the root of vertebrates

Yann Guiguen, Jérémy Pasquier, Alexis Fostier, Julien Bobe

► To cite this version:

Yann Guiguen, Jérémy Pasquier, Alexis Fostier, Julien Bobe. Evolution of sex hormone binding globulins reveals early gene duplication at the root of vertebrates. 2020. hal-03048977

HAL Id: hal-03048977

<https://hal.inrae.fr/hal-03048977>

Preprint submitted on 9 Dec 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

1 **Evolution of sex hormone binding globulins reveals early**
2 **gene duplication at the root of vertebrates.**

3 Yann Guiguen¹, Jeremy Pasquier¹, Alexis Fostier¹, Julien Bobe^{1*}

4

5 ¹ INRAE, LPGP, 35000 Rennes, France

6 Correspondence to: Julien Bobe, julien.bobe@inrae.fr

7

8 **Highlights:**

9

- 10 • Phylogeny, synteny and expression analyses shed new light on Shbg evolution in
11 vertebrates.
12 • Shbg diversity originates from a duplication event at the root of vertebrate evolution.
13 • This duplication was followed by many independent losses of *Shbg* paralogs in
14 vertebrates.
15 • *Shbg* paralogs have acquired different tissue expression patterns.
16

17 **Abstract**

18 Sex hormone-binding globulin (Shbg) is an important vertebrate blood carrier protein
19 synthesized in the liver and involved in the transport and local regulation of sex steroids in
20 target tissues. A novel *shbg* gene (*shbgb*) with a predominant ovarian expression was
21 recently characterized. Being initially found only in salmonids, this *shbgb* was originally
22 thought to result from the Salmonid-specific whole genome duplication. Using updated
23 transcriptomic and genomic resources we identified *Shbgb* orthologs in non-salmonid
24 teleosts (European eel, arowana), holosteans (spotted gar, bowfin), polypteriformes
25 (reedfish), agnatha (sea lamprey) and in amphibians, and found that the classical *Shbg* gene
26 (*Shbga*) displays a predominant hepatic expression whereas *Shbgb* has a predominant
27 gonadal expression. Together, these results indicate that these two *Shbg* genes most likely
28 originate from a whole genome duplication event at the root of vertebrate evolution, followed
29 by numerous and independent losses and by tissue expression specialization of *Shbga* and
30 *Shbgb* paralogs.

31

32 **Keywords:** evolution, Sex hormone-binding globulin, teleosts, vertebrates, gene expression

33

34 **1 Introduction**

35 Sex hormone-binding globulin (Shbg) is mainly known as a blood-protein carrier involved in
36 the transport of sex steroids in the plasma and in the regulation of their bioavailability to
37 target organs. Shbg proteins present sequence similarities with the other LamG domain-
38 containing proteins growth arrest-specific 6 (Gas6) and protein S alpha (Pros1) (Joseph,
39 1997; Joseph and Baker, 1992). By transporting and regulating androgens and estrogens
40 access to the gonads, Shbg plays important roles in vertebrates reproduction (Hammond,
41 2011). The Shbg protein was originally identified in the beta-globulin fraction of the human
42 serum (Rosner et al., 1969) and has previously been known as Androgen-binding protein

43 (Abp). *Shbg* genes and Shbg proteins have been characterized in a variety of tetrapod
44 species with the notable exception of birds (Westphal, 1986; Wingfield et al., 1984). In
45 aquatic vertebrates, Shbg was originally found in the plasma of an elasmobranch, the skate
46 (*Raja radiata*) (Freeman and Idler, 1969), and of a teleost, the rainbow trout (*Oncorhynchus*
47 *mykiss*) (Fostier and Breton, 1975). Since that moment on, Shbg have been subsequently
48 identified and studied in many fish species (see for review (Bobe et al., 2010)). Two different
49 *shbg* genes i.e., *shbga* and *shbgb*, have been characterized in teleosts with *shbga* being the
50 ortholog of the mammalian *Shbg*, which has been conserved from chondrichthyes to
51 tetrapods and *shbgb* that has only been reported up to now in salmonids. In contrast to
52 *shbga* that is mainly expressed in the liver (Bobe et al., 2008), the *shbgb* transcript was
53 mainly found in the ovary, suggesting a local mediation of the sex steroids effects by the
54 Shbgb protein (Bobe et al., 2008). These two salmonid Shbg proteins share very low identity
55 percentages, with for instance 26% identity between Shbga and Shgbb at the amino acid
56 level in the rainbow trout (Bobe et al., 2010, 2008). In comparison to other vertebrates and
57 teleost fishes, the salmonid ancestor has experienced an additional (4R) whole genome
58 duplication known as the salmonid specific whole genome duplication, or SaGD (Berthelot et
59 al., 2014). For this reason, and because no *shbgb* gene had ever been reported in a non-
60 salmonid species, it was first hypothesized that *shbga* and *shbgb* were ohnologous genes
61 resulting from the SaGD (Bobe et al., 2008). This hypothesis was subsequently challenged in
62 another study and the possibility of an ancient duplication followed by a lineage-specific
63 retention in salmonids was suggested (Miguel-Queralt et al., 2009).

64 The evolutionary history of *Shbg* genes in vertebrates thus remained unclear and deserved
65 further investigations. Using the increasing amount of genomic and transcriptomic data
66 available for many vertebrate species we revisited the evolutionary history of *Shbg* genes.
67 The transcriptomes of 24 actinopterygian species (including 22 teleosts) and vertebrate
68 genomes were included in the analysis, which led to the identification of previously non-
69 characterized *Shbgb* genes in several non-salmonid vertebrate lineages. Using
70 phylogenomic analyses, we identified several *Shbgb* orthologs in a variety of non-salmonid

71 vertebrate species, including teleosts, non-teleost actinopterygians, amphibians and one
72 agnatha. Combined with synteny reconstruction analysis, we demonstrated that *Shbg*
73 diversity results from a duplication event much older than the SaGD. To gain new information
74 on the functional evolution of *shbg* genes, we also used quantitative PCR and next
75 generation sequencing approaches, to characterize the expression profiles of *shbga* and
76 *shbgb* transcripts for several actinopterygian species. This showed that the paralogous *shbg*
77 genes have acquired different expression profiles with *shbgb* having a predominant gonadal
78 expression contrasting with a predominant liver expression of *shbga*.

79 **2 Material and Methods**

80 **2.1 Genomic and transcriptomic databases**

81 The genomes of the following species, human, *Homo sapiens*; tropical clawed frog, *Xenopus*
82 *tropicalis*; coelacanth, *Latimeria chalumnae*; spotted gar, *Lepisosteus oculatus* and zebrafish,
83 *Danio rerio* were explored using the Ensembl genome browser
84 (<http://www.ensembl.org/index.html>). The rainbow trout (*Oncorhynchus mykiss*) genomic
85 database was searched using the Genoscope trout genome browser
86 (<http://www.genoscope.cns.fr/trout/>). The European eel (*Anguilla Anguilla*) genomic database
87 was investigated using the European eel assembly available at ZF-Genomics
88 (<http://www.zfgenomics.org/sub/eel>). Transcriptomes of holostean and teleostean species
89 were investigated using the PhyloFish project resource (Pasquier et al., 2016) available at
90 <http://phylofish.sigenae.org>. The protein sequences of Human SHBG, *Xenopus* Shbg,
91 zebrafish Shbg, and rainbow trout Shbga and Shbgb were used as queries to identify
92 homologs of Shbga and Shbgb in the different genomic and transcriptomic databases
93 investigated. A similar methodology was used for Gas6 and Pros1 proteins that were
94 relevant to study due to their phylogenetical proximity and structural similarity.

95 2.2 Phylogenetic and synteny analyses

96 Amino-acid sequences of 126 predicted Shbg (a, b), Gas6 and Pros1 proteins were first
97 aligned using ClustalW (Thompson et al., 1994), then alignments were manually adjusted, to
98 improve the quality of the multiple sequence alignments. The JTT (Jones, Taylor and
99 Thornton) protein substitution matrix of the resulting alignment was determined using
100 ProtTest software (Darriba et al., 2011). Phylogenetic analysis of the proteins presenting
101 LamG domains (*i.e.* Shbga, Shbgb, Gas6 and Pros1) was performed using the neighbour
102 joining (NJM) method (MEGA 5.1 software), with 1000 bootstrap replicates (Tamura et al.,
103 2011). Trees were edited online with iTOL (Letunic and Bork, 2016) and exported as
104 Scalable Vector Graphics.

105 Synteny maps of the conserved genomic regions in human, *Xenopus*, coelacanth, spotted
106 gar and zebrafish were constructed based on information available within the Genomicus
107 (Muffato et al., 2010) v75.01 website ([http://www.genomicus.biologie.ens.fr/genomicus-
108 75.01/cgi-bin/search.pl](http://www.genomicus.biologie.ens.fr/genomicus-75.01/cgi-bin/search.pl)). Synteny map of the conserved genomic regions in the rainbow trout
109 was performed using the Rainbow Trout Genomicus Server
110 (<http://www.genomicus.biologie.ens.fr/genomicus-trout-01.01/cgi-bin/search.pl>). The synteny
111 analyses of European eel conserved genomic regions were obtained performing TBLASTN
112 searches in the corresponding genomic database. For each studied gene, the protein
113 sequences of human and zebrafish were used as queries.

114 Multiple alignments plots of *shbgb* genes in salmonids were processed online
115 (<http://genome.lbl.gov/vista/>) with mVISTA (Dubchak and Ryaboy, 2006; Poliakov et al.,
116 2014) using genomic *shbgb* sequences of rainbow trout, *Oncorhynchus mykiss*, Atlantic
117 salmon, *Salmon salar*, and Coho salmon, *Oncorhynchus kisutch*. Putative *shbgb*
118 pseudogenes were retrieved by TBLASTN searches on whole genome sequences using as
119 query the protein sequence of rainbow trout Shbgb.

120

121 **2.3 RNA-seq *shbg* and *shbg* tissue expression in holosteans and teleosts.**

122 RNA-seq and *de novo* assembly were performed for all studied species as previously
123 described (Berthelot et al., 2014; Braasch et al., 2016; Pasquier et al., 2016). In order to
124 study the expression patterns and levels of *shbg* transcripts for each actinopterygian species
125 with two *shbg* genes, we mapped RNA-seq reads on the corresponding *shbg* coding
126 sequence (CDS) using BWA-Bowtie (Langmead and Salzberg, 2012) with stringent mapping
127 parameters (maximum number of allowed mismatches $-aln\ 2$). Mapped reads were counted
128 using SAMtools (Li et al., 2009) `idxstat` command, with a minimum alignment quality value ($-q\ 30$) to discard ambiguous mapping reads. For each species, the numbers of mapped reads
129 were then normalized for each *shbg* gene across the eleven tissues using the reads per kilo
130 base per million mapped reads (RPKM) normalization. All RNA-seq data are available here:
131 (<http://phylofish.sigenae.org/index.html>)
132

133

134 **2.4 Quantitative PCR analysis (QPCR).**

135 QPCR was performed using the RNA collections of the PhyloFish RNA-seq project as
136 previously described (Braasch et al., 2016; Pasquier et al., 2016). Briefly, tissues were
137 sampled from the same female individual and testis from a male individual, when possible. In
138 some species and depending on the tissues, RNA samples from different individuals were
139 pooled to obtain sufficient amounts of RNA. Total RNA was extracted using Tri-Reagent
140 (Molecular Research Center, Cincinnati, OH, USA) according to the manufacturer's
141 instructions. Reverse transcription (RT) was performed using 1 μ g of RNA for each sample
142 with M-MLV reverse transcriptase and random hexamers (Promega, Madison, WI, USA).
143 Briefly, RNA and dNTPs were denatured for 6 min at 70°C, chilled on ice for 5 min before the
144 RT master mix was added. RT was performed at 37°C for 1 h and 15 min followed by a 15-
145 min incubation step at 70°C. Control reactions were run without reverse transcriptase and
146 used as negative control in the real-time PCR study. Quantitative RT-PCR (QPCR)
147 experiments were performed using an Applied Biosystems StepOne Plus. RT products,

148 including control reactions, were diluted to 1/25, and 4 μ l were used for each PCR. All QPCR
149 were performed in triplicates. QPCR was performed using a real-time PCR kit provided with
150 a Fast-SYBR Green fluorophore (Applied Biosystems) with 200 nM of each primer in order to
151 keep PCR efficiency between 80% and 100% for all target *shbg* genes. The relative
152 abundance of target cDNA within a sample set was calculated from serially diluted cDNA
153 pool (standard curve) using Applied Biosystem StepOne V.2.0 software. After amplification, a
154 fusion curve was obtained to validate the amplification of a single PCR product. The fusion
155 curves obtained showed that each primer pair used was specific of a single *shbg* transcript.
156 The negative control reactions were used to estimate background level. Genes were
157 considered significantly expressed when measured level was significantly above background
158 at $p < 0.05$ and within the range of the standard curve. For each studied tissue, cDNA
159 originating from three individual fish were pooled and subsequently used for real-time PCR.
160 Before further analysis, real-time PCR data were collected using the same detection
161 threshold for all studied genes. Data were subsequently normalized using the $\Delta\Delta C_t$ method
162 to 18S transcript abundance in samples diluted to 1:2,000.

163

164 **2.5 Clustering analysis**

165 Expression profiles originating from either QPCR and RNA-seq were represented using
166 supervised clustering methods (Eisen et al., 1998). Hierarchical clustering was processed
167 using centroid linkage clustering, that uses the average value of all points in a cluster as a
168 reference to calculate distance of other points, with Pearson's uncentered correlation as
169 similarity metric on data that were normalized and median-centered using the Cluster
170 program (Eisen et al., 1998). Results (colorized matrix) of hierarchical clustering analyses
171 were visualized using the Java TreeView program (Saldanha, 2004).

172

173 **3 Results**

174 **3.1 Shbg gene evolution in vertebrates**

175 In order to decipher phylogenetic relationships among Shbg sequences, a phylogenetic
176 reconstruction of the evolution of Shbg was made based using the alignment of 126
177 vertebrate LamG domain-containing proteins. This phylogeny includes Shbg proteins, growth
178 arrest specific 6 proteins (Gas6) and Vitamin K-dependent protein S (Pros1) and the tree
179 was rooted using as outgroup the zebrafish Laminin subunit alpha 4 (lama4). This analysis
180 (Fig. 1) shows that these vertebrate LamG domains proteins cluster into two major clades
181 containing Shbg proteins on one side and the Gas6 and Pros1 proteins on the other side.
182 These Shbg and Gas6/Pros1 clades are both significantly supported with high bootstrap
183 values (*i.e.* 75% and 100%, respectively). The Shbg clade contains two sub-clades both
184 supported by significant bootstrap values. The Shbga cluster (100% bootstrap support), in
185 red, contains all classical vertebrate Shbg proteins from chondrichthyes to teleosts with the
186 notable exception of birds in which no Shbg proteins have been detected (see also Fig.S2).
187 The Shbgb cluster (93% bootstrap support), in blue, contains not only the salmonid Shbgb
188 proteins, but also other vertebrate sequences outside the salmonid family including
189 sequences from teleosts (European eel and silver arowana), non-teleost bony fishes
190 (reedfish, spotted gar and bowfin), some amphibians and an agnatha, *i.e.*, the sea lamprey
191 (Fig.1 and Fig.S1). The tree topology indicates that Shbgb proteins are not specific to the
192 salmonids lineage and thus suggests a much more ancient origin of *Shbgb* genes in
193 vertebrates than previously hypothesized (Bobe et al., 2008).

194 To strengthen this phylogeny-based analysis of Shbg protein evolutionary history we carried
195 out a synteny analysis in order to better support this hypothesis of an ancient origin *Shbga*
196 and *Shbgb* genes. The synteny analysis first revealed that these two *Shbg* genes are located
197 on two different syntenic chromosome regions in vertebrates (see Fig. 2A for the *Shbga*
198 locus and Fig. 2B for the *Shbgb* locus). In contrast, *Gas6* and *Pros1* genes are located on
199 the same syntenic chromosome region in most studied species, with the exception of

200 primates (Fig. 2C). In vertebrates, *Shbga*, *Shbgb*, and *Gas6/Pros1* are also located in
201 regions containing other syntenic gene families. These neighboring genes are spread over
202 four syntenic regions (Fig. 2A-2D) like for instance for the *Atp*-related genes (*Atp1b2*,
203 *Atp1b3*, *Atp4b*) that are found in all four syntenic chromosome regions. In addition, *Zbtb*-
204 related genes (*Ztb4*, *Zbt38*, *Zbt33*) and *Lamp*-related genes (*Lamp1*, *Lamp2*, *Lamp3*) are
205 present in three of the four different syntenic chromosome regions depending of the gene
206 family. Altogether, these results strongly suggest that the diversity of the gene family present
207 on these four syntenic chromosome regions probably results from early whole genome
208 duplication events (VG1 and/or VG2) that occurred at the root of vertebrate evolution with a
209 subsequent complex pattern of gene retention and gene losses.

210 Using the recently released salmonid genome resources, we also re-investigated the
211 presence of additional copies of *shbg* genes in salmonids and confirmed that *shbga* and
212 *shbgb* were both retained as single copies (Fig. 1) in rainbow trout, Atlantic salmon and coho
213 salmon suggesting that no functional duplicated copies were retained after the salmonid
214 whole genome duplication (SaGD). However, we also found an additional *shbgb* gene
215 conserved in these three salmonid species (Fig.3A), but with many stop codons in its
216 deduced open reading frame (see example for rainbow trout in Fig.3B), suggesting that this
217 gene (*ψ shbgb*) was subsequently pseudogenized after the SaGD.

218

219 **3.2 Expression patterns of *shbga* and *shbgb***

220 In all investigated actinopterygians, *shbga* was found to be mainly expressed in the liver,
221 supporting a conserved role for this blood-secreted Shbg (Fig.4). However, in the silver
222 arowana a low *shbga* expression in the gonads is also detected in addition to the
223 predominant liver expression (Fig.4A and Fig.4B). In contrast to *shbga*, expression of *shbgb*
224 is not predominant in the liver in all investigated actinopterygians (Fig.4A and Fig.4B). In
225 contrast, *shbgb* expression is predominantly detected in the gonads (ovary and/or testis) with
226 the notable exception of the European eel.

227 4 Discussion

228 In this study, we aimed at investigating the diversity of the *Shbg* family in vertebrates and the
229 evolutionary history of *Shbg* genes. To date, despite the recent discovery of a second *shbg*
230 gene (i.e. *shbgb*) in salmonids, the origin and diversity of *Shbg* genes in vertebrates has
231 remained controversial. Because salmonids experienced an additional whole genome
232 duplication (SaGD) approximately 100 Mya (Berthelot et al., 2014; Macqueen and Johnston,
233 2014) compared to other teleost fish and as *shbgb* genes were initially found in salmonids
234 and, never reported at that time in any non-salmonid species, *shbga* and *shbgb* have been
235 first hypothesized to be the result of the SaGD (Bobe et al., 2010, 2008). However pairwise
236 comparison of *Shbga* and *Shbgb* reveals a surprisingly low sequence identity (around 25% at
237 the amino acid level) that was initially interpreted as *Shbga* and *Shbgb* being highly divergent
238 SaGD paralog. However, this paralogy relationship was not supported by the phylogeny
239 reconstruction (Bobe et al., 2008) and this discrepancy was thus explained as being the
240 result of a long-branch attraction artifact resulting from the dramatic divergence of these two
241 sequences (Bobe et al., 2010, 2008). Based on the cloning of another salmonid *shbgb* gene
242 in coho salmon, *Oncorhynchus kisutch* and the low sequence identity between salmonids
243 *Shbga* and *Shbgb*, other authors (Miguel-Queralt et al., 2009), hypothesized that the *shbgb*
244 gene could stem from a much more ancient duplication than the SaGD. In order to
245 decipher the evolutionary history of *Shbg* genes we re-analyzed the phylogenetic
246 relationships of *Shbg* genes, their local synteny context, and the evolution of the
247 phylogenetically and structurally closely related genes i.e., *Gas6* and *Pros1* that also contain
248 LamG domains and are often identified as potential members of the same family. The
249 identification of new *Shbgb* genes in vertebrates, the *Shbg* phylogenetic tree topology and
250 their local synteny relationships strongly suggest that *Shbga* and *Shbgb* genes result from a
251 whole genome duplication event that occurred very early at the root on the vertebrate
252 lineage. The presence of a single *Shbga* gene and a single *Shbgb* gene in amphibians,
253 holosteans, polypteriformes, agnatha and some teleost fishes, suggests that this *Shbg*

254 duplication stems at least from the second round of vertebrate genome duplication (VG2). It
255 is however also possible that *Shbga* and *Shbgb* originate the first round of vertebrate
256 genome duplication (VG1) followed by the loss of one duplicate of each gene before VG2.
257 Following this early duplication, these two *Shgb* paralogs have evolved through many
258 different phylum-specific gene retentions and/or gene losses. Among them the case of birds
259 is interesting as they not only lost their *Shbgb* gene like reported here for many other
260 tetrapods, but also their *Shbga* gene that is found to be conserved in all other vertebrates.
261 This complete absence of *Shbga* in birds has been already reported and it was hypothesized
262 that this specific steroid hormone-binding transport would then be performed by a
263 corticosteroid-binding globulin (Wingfield et al., 1984). Similarly, no *Shbg* homologs were
264 detected by homology searches in Chondrichthyes (data not shown) but their complete
265 absence in this clade requires further in-depth analysis and additional genome information as
266 *Shbg*-like sex steroid binding capacities exist in the serum of the Thorny skate (Freeman and
267 Idler, 1969). In tetrapods, *Shbgb* was only found in Amphibians along with *Shbga*. *Shbgb*
268 was also found in holosteans (spotted gar and bowfin), polypteriformes (reedfish), agnatha
269 (sea lamprey) and in a few teleost fish orders i.e., in Elopomorphs (European eel),
270 Osteoglossiforms (silver arowana) and Salmoniforms even though the protein is frequently
271 misannotated in GenBank (Fig.S1). Interestingly we did not find any retention of additional
272 whole genome [SaGD and the teleost specific duplication (TGD)] paralogs for both *shbga*
273 and *shbgb* gene with the exception of a pseudogenized SaGD *shbgb* paralog (ψ *shbgb*). This
274 indicates that these extra whole genome duplications did no impact the repertoire of *shbg*
275 genes with a maximum of one *shbga* and one *shbgb* functional copies in all investigated
276 teleost clades. This systematic and independent losses of additional *shbga* and *shbgb*
277 duplicated paralogs in teleosts may reflect an evolutionary constraint of maintaining a correct
278 gene and protein dosage as it has been suggested in other organisms (Conant et al., 2014;
279 Gout and Lynch, 2015).

280 In consistency with existing data in mammals, our expression data showed that *shbga* is
281 predominantly expressed in the liver in the different teleost species studied here. This

282 confirms what has previously been reported in various teleost species including zebrafish
283 (*Danio rerio*) (Miguel-Queralt et al., 2004), rainbow trout (*Oncorhynchus mykiss*) (Bobe et al.,
284 2008), Coho salmon (*Oncorhynchus kisutch*) (Miguel-Queralt et al., 2009), pejerrey
285 (*Odontesthes bonariensis*) (González et al., 2017) and sea bass (*Dicentrarchus labrax*)
286 (Miguel-Queralt et al., 2007). In addition, this strong hepatic expression is also observed in
287 spotted gar (*Lepisosteus oculatus*) and bowfin (*Amia calva*) as shown by both RNA-seq and
288 QPCR data.

289 In contrast to *shbga*, data on the tissue distribution of *shbgb* remain scarce. The ovarian
290 predominant expression of *shbgb* was originally reported in rainbow trout, in which the
291 transcript could also be detected at lower levels in muscle and stomach (Bobe et al., 2008).
292 Semi quantitative data in Coho salmon confirmed the expression of *shbgb* in the ovary and
293 stomach and revealed its presence in gills (Miguel-Queralt et al., 2009). Here we show that
294 *shbgb* is also predominant expressed in the ovary in brown trout, silver arowana and
295 grayling. We also report a strong testicular expression of *shbgb* in the two holostean species,
296 spotted gar and bowfin, that appears to be lost in teleosts. In addition, the *shbgb* gene does
297 not exhibit any gonad predominant expression in European eel. Together, our data show that
298 *shbga* and *shbgb* have a very specific expression patterns with a predominant expression in
299 liver and gonads, respectively. This pattern appears to be conserved during evolution without
300 any significant change following whole genome duplications events (TGD and SaGD), with
301 the exception of European eel in which the gonad predominant expression of *shbgb* appears
302 to be lost. Finally, the strong testicular expression of *shbgb* revealed in bowfin and spotted
303 gar is not found in any teleost species suggesting a specific role of *Shbgb* in testicular
304 physiology in holostean species.

305

306 The multiple independent losses of *Shbgb* across vertebrates, while *Shbga*, *Gas6* and *Pros1*
307 have been conserved in almost all vertebrates, could reflect different adaptive and
308 reproductive strategies as *Shbg* have been shown to be important carrier proteins for the
309 blood transport of sex steroids and for their delivery to target reproductive tissues

310 (Hammond, 2011). However, despite this discrepancy among species, the distinct roles of
311 Shgba in hormone transport in the blood and of Shbgb in local hormone action in
312 reproductive organs as well as the associated expression in liver and gonads, respectively,
313 appears to be evolutionary conserved in species that have retained both genes despite a few
314 intriguing species-specific exceptions.

315 **5 Acknowledgement**

316 This work was supported by the French national research Agency (ANR-10-GENM-017–
317 PhyloFish).

318 **6 References**

319 Berthelot, C., Brunet, F., Chalopin, D., Juanchich, A., Bernard, M., Noël, B., Bento, P., Da
320 Silva, C., Labadie, K., Alberti, A., Aury, J.-M., Louis, A., Dehais, P., Bardou, P., Montfort, J.,
321 Klopp, C., Cabau, C., Gaspin, C., Thorgaard, G.H., Boussaha, M., Quillet, E., Guyomard, R.,
322 Galiana, D., Bobe, J., Volff, J.-N., Genêt, C., Wincker, P., Jaillon, O., Roest Crolius, H.,
323 Guiguen, Y., 2014. The rainbow trout genome provides novel insights into evolution after
324 whole-genome duplication in vertebrates. *Nat. Commun.* 5, 3657.
325 <https://doi.org/10.1038/ncomms4657>

326 Bobe, J., Guiguen, Y., Fostier, A., 2010. Diversity and biological significance of sex hormone-
327 binding globulin in fish, an evolutionary perspective. *Mol. Cell. Endocrinol.* 316, 66–78.

328 Bobe, J., Mahé, S., Nguyen, T., Rime, H., Vizziano, D., Fostier, A., Guiguen, Y., 2008. A
329 novel, functional, and highly divergent sex hormone-binding globulin that may participate in
330 the local control of ovarian functions in salmonids. *Endocrinology* 149, 2980–2989.
331 <https://doi.org/10.1210/en.2007-1652>

332 Braasch, I., Gehrke, A.R., Smith, J.J., Kawasaki, K., Manousaki, T., Pasquier, J., Amores, A.,
333 Desvignes, T., Batzel, P., Catchen, J., Berlin, A.M., Campbell, M.S., Barrell, D., Martin, K.J.,
334 Mulley, J.F., Ravi, V., Lee, A.P., Nakamura, T., Chalopin, D., Fan, S., Wcisel, D., Cañestro,

335 C., Sydes, J., Beaudry, F.E.G., Sun, Y., Hertel, J., Beam, M.J., Fasold, M., Ishiyama, M.,
336 Johnson, J., Kehr, S., Lara, M., Letaw, J.H., Litman, G.W., Litman, R.T., Mikami, M., Ota, T.,
337 Saha, N.R., Williams, L., Stadler, P.F., Wang, H., Taylor, J.S., Fontenot, Q., Ferrara, A.,
338 Searle, S.M.J., Aken, B., Yandell, M., Schneider, I., Yoder, J.A., Volff, J.-N., Meyer, A.,
339 Amemiya, C.T., Venkatesh, B., Holland, P.W.H., Guiguen, Y., Bobe, J., Shubin, N.H., Di
340 Palma, F., Alföldi, J., Lindblad-Toh, K., Postlethwait, J.H., 2016. The spotted gar genome
341 illuminates vertebrate evolution and facilitates human-teleost comparisons. *Nat. Genet.* 48,
342 427–437. <https://doi.org/10.1038/ng.3526>
343 Conant, G.C., Birchler, J.A., Pires, J.C., 2014. Dosage, duplication, and diploidization:
344 clarifying the interplay of multiple models for duplicate gene evolution over time. *Curr. Opin.*
345 *Plant Biol.* 19, 91–98. <https://doi.org/10.1016/j.pbi.2014.05.008>
346 Darriba, D., Taboada, G.L., Doallo, R., Posada, D., 2011. ProtTest 3: fast selection of best-fit
347 models of protein evolution. *Bioinforma. Oxf. Engl.* 27, 1164–1165.
348 <https://doi.org/10.1093/bioinformatics/btr088>
349 Dubchak, I., Ryaboy, D.V., 2006. VISTA family of computational tools for comparative
350 analysis of DNA sequences and whole genomes. *Methods Mol. Biol. Clifton NJ* 338, 69–89.
351 <https://doi.org/10.1385/1-59745-097-9:69>
352 Eisen, M.B., Spellman, P.T., Brown, P.O., Botstein, D., 1998. Cluster analysis and display of
353 genome-wide expression patterns. *Proc. Natl. Acad. Sci. U. S. A.* 95, 14863–14868.
354 Fostier, A., Breton, B., 1975. Binding of steroids by plasma of a teleost: the rainbow trout,
355 *Salmo gairdneri*. *J Steroid Biochem.*
356 Freeman, H.C., Idler, D.R., 1969. Sex hormone binding proteins. II. Isolation from serum of
357 an elasmobranch (*Raja radiata*). *Gen. Comp. Endocrinol.* 13, 83–91.
358 [https://doi.org/10.1016/0016-6480\(69\)90224-x](https://doi.org/10.1016/0016-6480(69)90224-x)
359 González, A., Fernandino, J.I., Hammond, G.L., Somoza, G.M., 2017. Sex hormone binding
360 globulin: Expression throughout early development and adult pejerrey fish, *Odontesthes*
361 *bonariensis*. *Gen. Comp. Endocrinol.* 247, 205–214.
362 <https://doi.org/10.1016/j.ygcen.2017.02.004>

- 363 Gout, J.-F., Lynch, M., 2015. Maintenance and Loss of Duplicated Genes by Dosage
364 Subfunctionalization. *Mol. Biol. Evol.* 32, 2141–2148. <https://doi.org/10.1093/molbev/msv095>
- 365 Hammond, G.L., 2011. Diverse roles for sex hormone-binding globulin in reproduction. *Biol.*
366 *Reprod.* 85, 431–441. <https://doi.org/10.1095/biolreprod.111.092593>
- 367 Joseph, D.R., 1997. Sequence and functional relationships between androgen-binding
368 protein/sex hormone-binding globulin and its homologs protein S, Gas6, laminin, and agrin.
369 *Steroids* 62, 578–588. [https://doi.org/10.1016/s0039-128x\(97\)00045-7](https://doi.org/10.1016/s0039-128x(97)00045-7)
- 370 Joseph, D.R., Baker, M.E., 1992. Sex hormone-binding globulin, androgen-binding protein,
371 and vitamin K-dependent protein S are homologous to laminin A, merosin, and *Drosophila*
372 crumbs protein. *FASEB J. Off. Publ. Fed. Am. Soc. Exp. Biol.* 6, 2477–2481.
373 <https://doi.org/10.1096/fasebj.6.7.1532944>
- 374 Langmead, B., Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. *Nat.*
375 *Methods* 9, 357–359. <https://doi.org/10.1038/nmeth.1923>
- 376 Letunic, I., Bork, P., 2016. Interactive tree of life (iTOL) v3: an online tool for the display and
377 annotation of phylogenetic and other trees. *Nucleic Acids Res.* 44, W242-245.
378 <https://doi.org/10.1093/nar/gkw290>
- 379 Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G.,
380 Durbin, R., 1000 Genome Project Data Processing Subgroup, 2009. The Sequence
381 Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078–2079.
382 <https://doi.org/10.1093/bioinformatics/btp352>
- 383 Macqueen, D.J., Johnston, I.A., 2014. A well-constrained estimate for the timing of the
384 salmonid whole genome duplication reveals major decoupling from species diversification.
385 *Proc. Biol. Sci.* 281, 20132881. <https://doi.org/10.1098/rspb.2013.2881>
- 386 Miguel-Queralt, S., Blázquez, M., Piferrer, F., Hammond, G.L., 2007. Sex hormone-binding
387 globulin expression in sea bass (*Dicentrarchus labrax* L.) throughout development and the
388 reproductive season. *Mol. Cell. Endocrinol.* 276, 55–62.
389 <https://doi.org/10.1016/j.mce.2007.06.009>
- 390 Miguel-Queralt, S., Knowlton, M., Avvakumov, G.V., Al-Nouno, R., Kelly, G.M., Hammond,

391 G.L., 2004. Molecular and functional characterization of sex hormone binding globulin in
392 zebrafish. *Endocrinology* 145, 5221–5230. <https://doi.org/10.1210/en.2004-0678>

393 Miguel-Queralt, S., Underhill, C., Devlin, R.H., Hammond, G.L., 2009a. Characterization and
394 measurement of the plasma alpha- and beta-sex hormone-binding globulin paralogs in
395 salmon. *Endocrinology* 150, 366–375. <https://doi.org/10.1210/en.2008-0964>

396 Miguel-Queralt, S., Underhill, C., Devlin, R.H., Hammond, G.L., 2009b. Characterization and
397 Measurement of the Plasma α - and β -Sex Hormone-Binding Globulin Paralogs in Salmon.
398 *Endocrinology* 150, 366–375. <https://doi.org/10.1210/en.2008-0964>

399 Muffato, M., Louis, A., Poisnel, C.-E., Roest Crollius, H., 2010. Genomicus: a database and a
400 browser to study gene synteny in modern and ancestral genomes. *Bioinforma. Oxf. Engl.* 26,
401 1119–1121. <https://doi.org/10.1093/bioinformatics/btq079>

402 Pasquier, J., Cabau, C., Nguyen, T., Jouanno, E., Severac, D., Braasch, I., Journot, L.,
403 Pontarotti, P., Klopp, C., Postlethwait, J.H., Guiguen, Y., Bobe, J., 2016. Gene evolution and
404 gene expression after whole genome duplication in fish: the PhyloFish database. *BMC*
405 *Genomics* 17, 368. <https://doi.org/10.1186/s12864-016-2709-z>

406 Poliakov, A., Foong, J., Brudno, M., Dubchak, I., 2014. GenomeVISTA--an integrated
407 software package for whole-genome alignment and visualization. *Bioinforma. Oxf. Engl.* 30,
408 2654–2655. <https://doi.org/10.1093/bioinformatics/btu355>

409 Rosner, W., Christy, N.P., Kelly, W.G., 1969. Partial purification and preliminary
410 characterization of estrogen-binding globulins from human plasma. *Biochemistry* 8, 3100–
411 3108.

412 Saldanha, A.J., 2004. Java Treeview--extensible visualization of microarray data. *Bioinforma.*
413 *Oxf. Engl.* 20, 3246–3248. <https://doi.org/10.1093/bioinformatics/bth349>

414 Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., Kumar, S., 2011. MEGA5:
415 molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance,
416 and maximum parsimony methods. *Mol. Biol. Evol.* 28, 2731–2739.
417 <https://doi.org/10.1093/molbev/msr121>

418 Thompson, J.D., Higgins, D.G., Gibson, T.J., 1994. CLUSTAL W: improving the sensitivity of

419 progressive multiple sequence alignment through sequence weighting, position-specific gap

420 penalties and weight matrix choice. *Nucleic Acids Res.* 22, 4673–4680.

421 Westphal, U., 1986. Steroid-protein interactions II. *Monogr. Endocrinol.* 27, 1–603.

422 Wingfield, J.C., Matt, K.S., Farner, D.S., 1984. Physiologic properties of steroid hormone-

423 binding proteins in avian blood. *Gen. Comp. Endocrinol.* 53, 281–292.

424 [https://doi.org/10.1016/0016-6480\(84\)90254-5](https://doi.org/10.1016/0016-6480(84)90254-5)

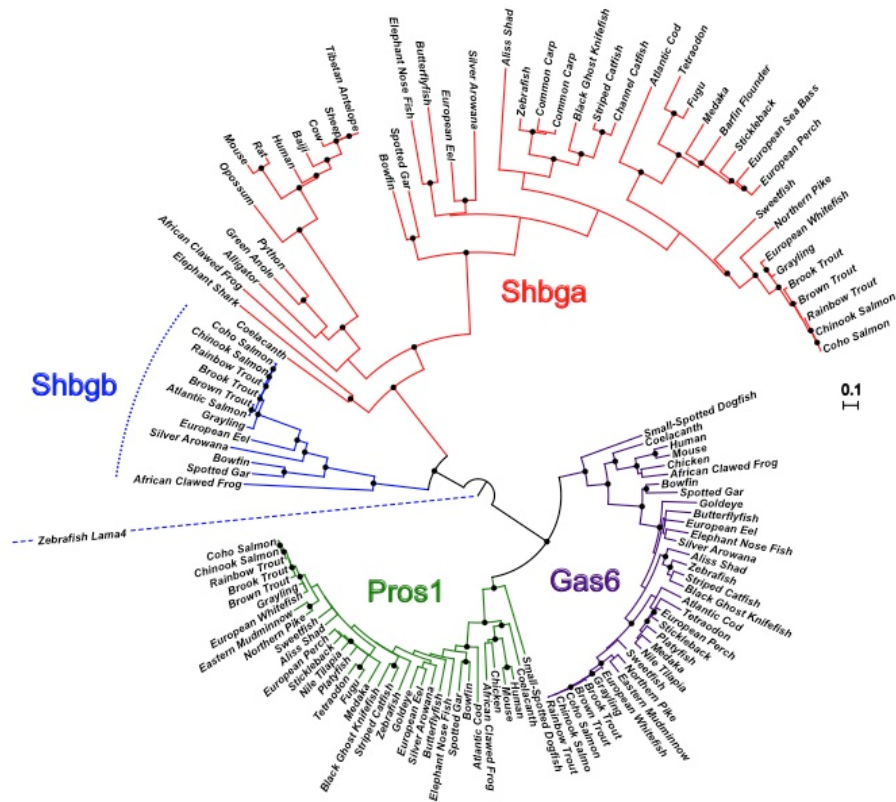
425

426

427

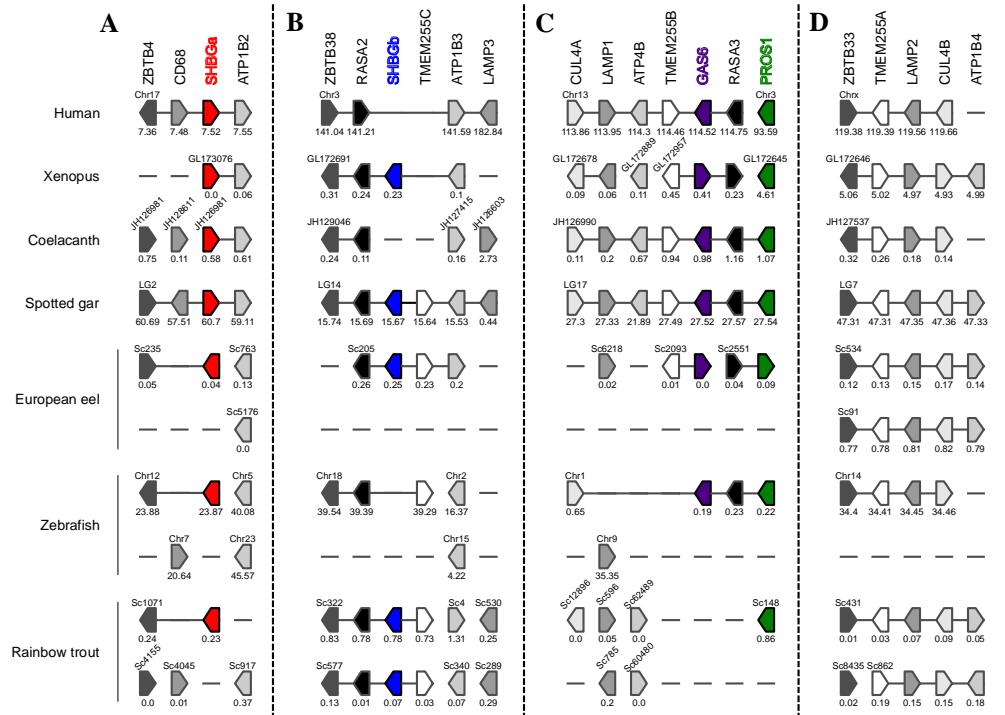
428 **Figures**

429
430



431
432
433
434
435
436
437
438
439

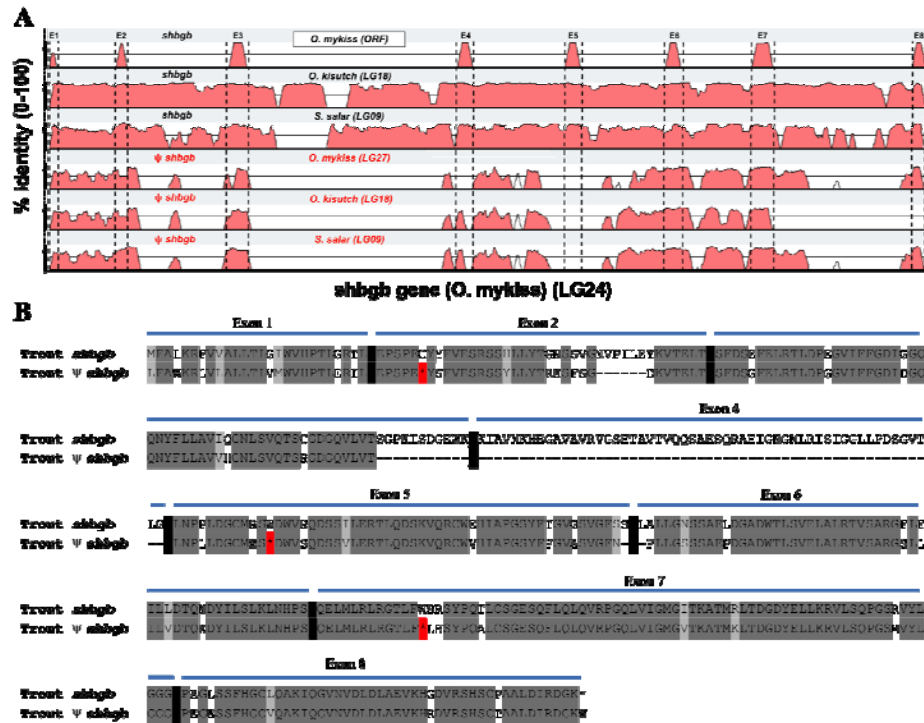
Figure 1: Phylogenetic reconstruction of the evolution of LamG domains proteins (Shbg, Gas6 and Pros1) in vertebrates. Circular NJM phylogenetic tree of LamG domains proteins including Shbga in red, Shbgb in blue, Gas6 in purple and Pros1 in green. The tree is rooted using zebrafish Laminin Subunit Alpha 4 (lama4) and bootstrap values over 0.75 are shown with a black dot on each significant node.



440

441 **Figure 2: Synteny maps of conserved genomic regions around LamG domains**
 442 **proteins (i.e. Shbga, Shbgb, Gas6 and Pros1) in human, Xenopus, coelacanth, spotted**
 443 **gar, European eel, zebrafish and rainbow trout. Synteny maps are given for genomic**
 444 **regions around Shbga (A), Shbgb (B), Gas6 and Pros1 locus (C) and for a fourth region**
 445 **containing homologs of neighbouring Shbga, Shbgb, Gas6 and Pros1 genes (D). Genes**
 446 **are represented by blocks with an arrowed side indicating the gene orientation on chromosomes,**
 447 **linkage group or scaffolds. Gene location on chromosomes (Chr for Human and zebrafish),**
 448 **Linkage group (LG for spotted gar) and scaffolds (Ensembl reference or scaffold number) is**
 449 **given in Mb below each gene block. Genes belonging to the same Chr, LG or scaffolds are**
 450 **linked by a solid line.**
 451

452



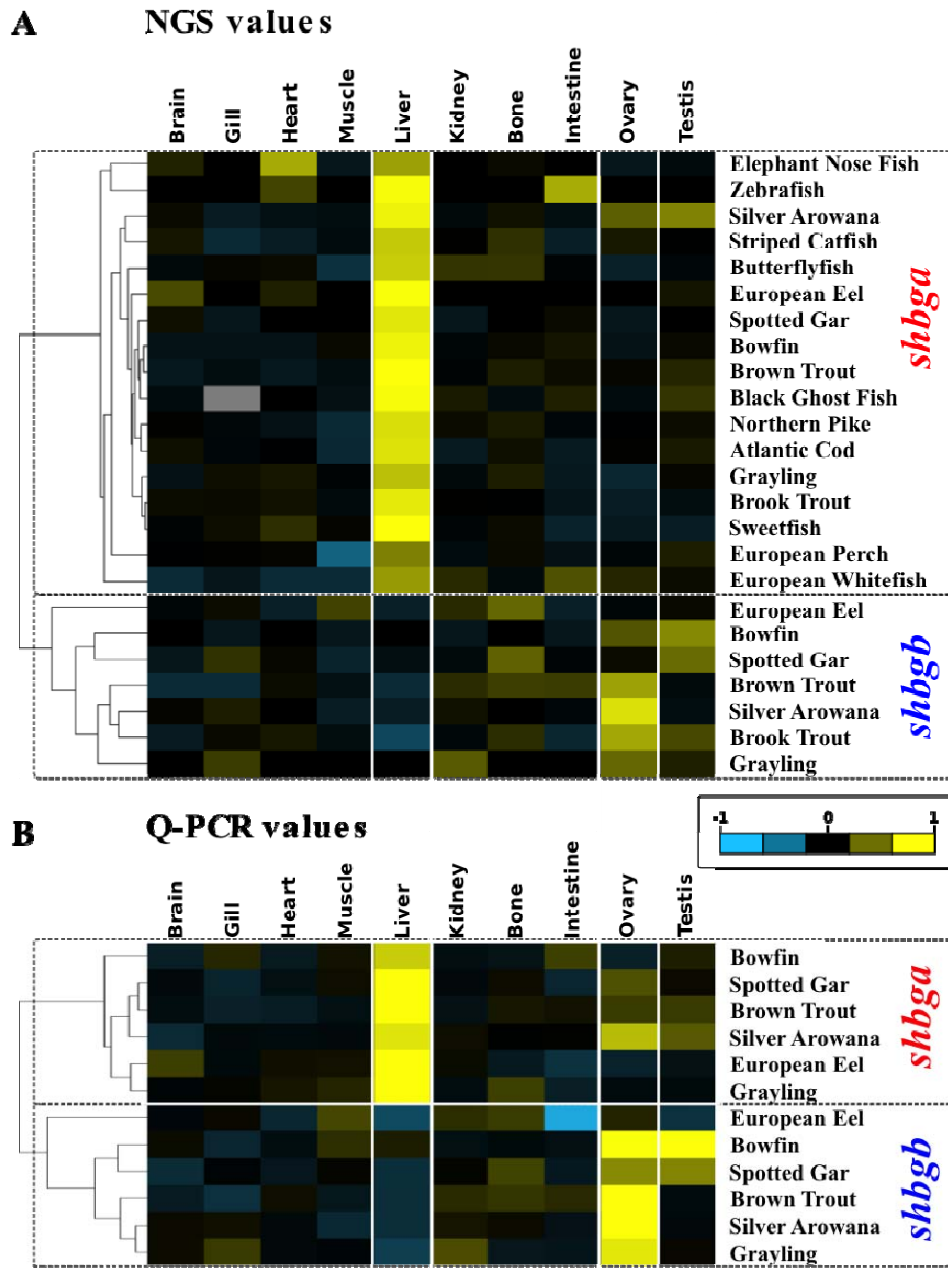
453

454 **Figure 3: Multiple alignments plots of *shbgb* genes in salmonids. (A).** Percentage of
 455 sequence identity of *shbgb* genes in Atlantic salmon, *Salmon salar*, and Coho salmon,
 456 *Oncorhynchus kisutch* compared to the rainbow trout, *Oncorhynchus mykiss*, *shbgb* gene on
 457 linkage group (LG) 24. In addition to the functional *shbgb* genes that were found in all
 458 salmonid species investigated, all these species have an additional *shbgb* homolog
 459 containing multiple stop codons and thus considered as a pseudogene (ψ *shbgb*). (B).
 460 Rainbow trout Shbgb and ψ Shbgb protein alignment showing that the ψ Shbgb contains
 461 multiple stop codons (red asterisks) and a large deletion in exon 4 of the Shbgb protein.
 462

463

464

465

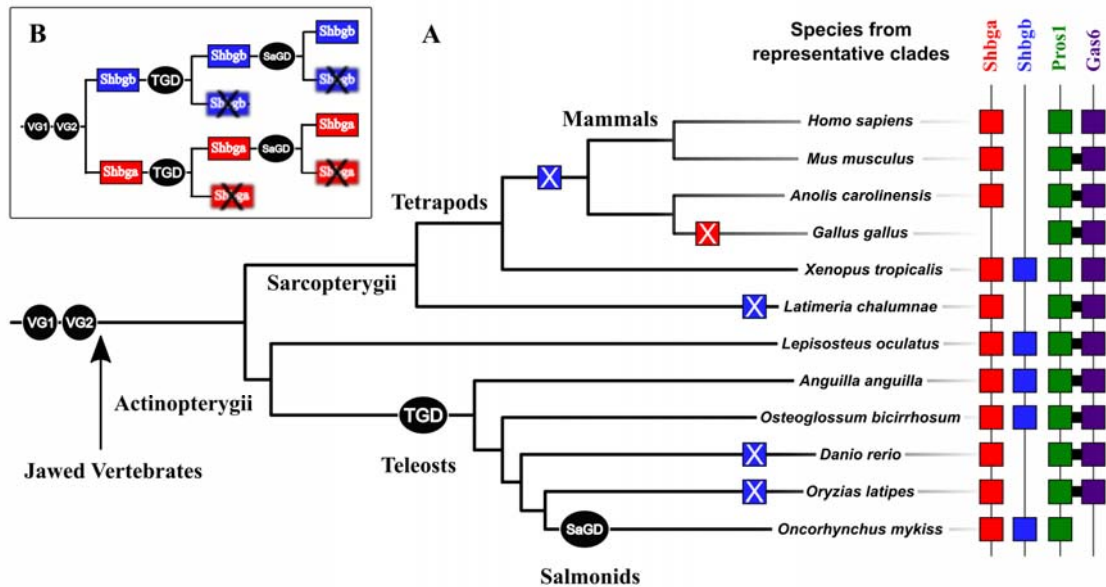


466

467 **Figure 4: Tissue expression profiles of *shbga* and *shbgb* genes.** (A). Heatmap
 468 (colorized matrix) of the hierarchical clustering of tissue RNA-Seq expression profiles (NGS
 469 values) of *shbg* genes in different Holostean (spotted gar and bowfin) and teleost species. A
 470 predominant expression of *shbga* is found in the liver contrasting with the predominant
 471 expression of *shbgb* in gonads (B). Heatmap (colorized matrix) of the hierarchical clustering
 472 of tissue expression profiles of *shbg* genes analyzed by quantitative PCR (QPCR values) in
 473 different Holostean (spotted gar and bowfin) and teleost species. Colorized matrixes highlight
 474 the high expressing tissues in yellow, the low expressing tissues in blue and the median
 475 expression in black (see color scale).
 476

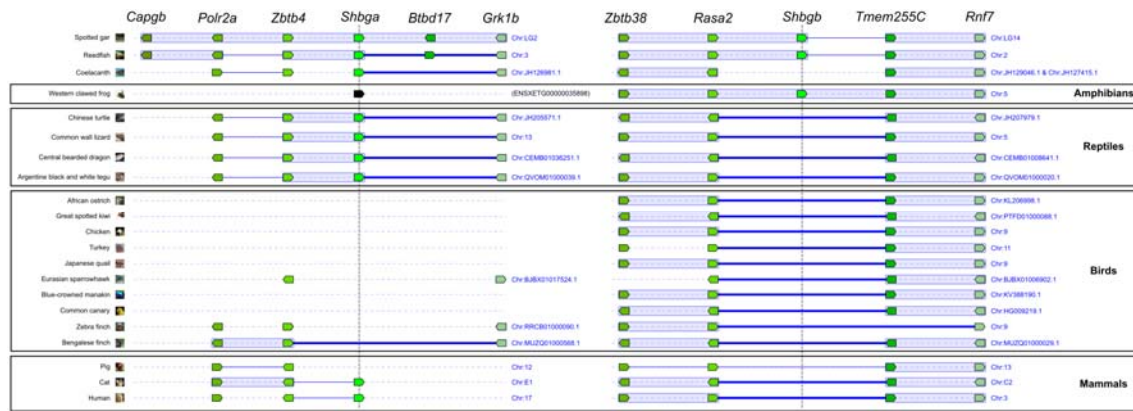
477

478



479

480 **Figure 5: Evolution of *Shbg/shbg*, *Gas6/gas6*, and *Pros1/pros1* genes through**
 481 **successive whole genome duplications and independent gene losses. (A).** Schematic
 482 representation of the evolution of LamG domains proteins including, *Shbga* (red squares),
 483 *Shbgb* (blue squares), *Gas6* (purple squares) and *Pros1* (green squares) in some jawed
 484 vertebrates' representative species. Whole genome duplications (black circles; VG1 and
 485 VG2: vertebrate genome duplications 1 and 2, TGD: teleost genome duplication, SaGD:
 486 salmonid genome duplication) are indicated at each duplication nodes. Gene losses are
 487 represented by square boxes with a cross inside. (B). Simplified representation of the
 488 evolution of *Shbg* proteins after whole genome duplications showing the systematic losses of
 489 one *Shbga* and *Shbgb* paralog after each duplication.
 490
 491



514

515 **Supplementary Figure 2: Genomic context and synteny relationships of some**
 516 **vertebrate Shbga and Shbgb proteins, showing the absence of both Shbga and Shbgb**
 517 **in birds and of Shbgb only in mammals and reptiles.** This analysis was performed using
 518 the Genomicus genome browser (version 99.01,
 519 <https://www.genomicus.biologie.ens.fr/genomicus-99.01/>) using as seed sequences the
 520 accession numbers of Shbga (ENSLOCG00000013875) and Shbgb
 521 (ENSLOCG00000008107) of spotted gar (*Lepisosteus oculatus*). Only a subset of species is
 522 depicted in this figure but the absence of both Shbga and Shbgb in birds and of Shbgb only
 523 in mammals and reptiles was consistently found in all birds (N=35), mammals (N=99) and
 524 reptiles (N=14) genomes available in the Genomicus version 99.01. No Shbgb homolog were
 525 found in the elephant shark, hagfish and lamprey genomes available in this Genomicus
 526 version 99.01.
 527