Unexpectedly large number of conserved non-coding regions within the ancestral chordate Hox cluster

Juan Pascual-Anaya, Salvatore D'Aniello and Jordi Garcia-Fernàndez*.

Departament de Genètica, Facultat de Biologia, Universitat de Barcelona, Barcelona 08028, Spain.

Corresponding author: Jordi Garcia-Fernàndez

e-mail: jordigarcia@ub.edu

phone: +34 934034437

fax: +34 934034420

Total words: 3785

Expected printed pages: 8

Abstract

The single amphioxus Hox cluster contains 15 genes and may well resemble the ancestral chordate Hox cluster. We have sequenced the Hox genomic complement of the European amphioxus *Branchiostoma lanceolatum* and compared it to the American species, *B. floridae*, by phylogenetic footprinting, to gain insights into the evolution of Hox regulation in chordates. We found that Hox intergenic regions are largely conserved between the two amphioxus species, especially in the case of genes located at the 3' of the cluster, a trend previously observed in vertebrates. We further compared the amphioxus Hox cluster with the human HoxA and HoxD clusters finding several conserved non-coding regions, both in intergenic and intronic regions. This suggests that the regulation of Hox genes is highly conserved across chordates, consistent with the similar Hox expression patterns in vertebrates and amphioxus.

Key words: amphioxus, Hox genes, conserved non-coding regions, phylogenetic footprinting.

Introduction

Hox genes are master developmental genes first discovered in *Drosophila* (Lewis, 1978), and so far identified in all studied eumetazoans. Their main function is the anterior-posterior patterning, both of the body axis and of other secondary axes as those of vertebrate limbs (Zakany and Duboule, 2007) and genital system (Podlasek *et al*., 2002; Wagner and Lynch, 2005).

Humans have a total of 39 Hox genes located in four clusters that were originated by the two rounds of whole genome duplications that took place during early vertebrate evolution. They are organized in 13 paralogous groups (PG), whose numbering reflects their position in the cluster. The cephalochordate amphioxus has a single continuous array of 15 Hox genes (Ferrier *et al*., 2000; Garcia-Fernàndez and Holland, 1994; Holland *et al*., 2008), the richest (in terms of gene content) Hox cluster isolated to date, with the same transcriptional orientation for all genes. Each amphioxus Hox gene from the anterior and central groups (Hox1 to Hox8) represents a vertebrate PG, whereas the relationship between posterior genes (from Hox9) is difficult to establish due to their high evolutionary divergence, a phenomenon called ‘deuterostome posterior flexibility’ by Ferrier *et al*. (2000). Amphioxus is the one of the few extant clades among metazoans with a single Hox cluster that has not been broken or is in the process of disintegration, a critical step in the evolution of the small and highly meta-

61 regulated mammalian clusters (Duboule, 2007). Due to these features, the amphioxus
62 Hox cluster is probably the best model to study and understand the origin and evolution
63 of the regulation of vertebrate Hox gene clusters.

64 The most striking characteristic of Hox genes is the spatiotemporal collinear
65 expression, i.e. genes located at 3' of the cluster are expressed in the anterior part of the
66 body and earlier in development than 5' genes (Duboule, 1994; Duboule and Dollé,
67 1989). The expression pattern of Hox genes is conserved between amphioxus and
68 vertebrates, at least at the level of the central nervous system (CNS). Vertebrate Hox
69 genes are expressed in the posterior part of the CNS, with a sharp and nested anterior
70 expression limit in the rhombencephalon, hence each rhombomere presents a different
71 combination of expressed Hox genes (the *Hox* code; Kessel and Gruss, 1991).
72 Amphioxus Hox genes are also expressed in a spatiotemporal collinear manner in the
73 central nervous system (CNS), with expressions restricted to the posterior CNS and
74 excluding the cerebral vesicle (Wada *et al*., 1999; Schubert *et al*., 2006).

75 The similar expression in the CNS of Hox genes of vertebrates and amphioxus
76 suggests that the regulation of these genes may also be conserved between the two
77 lineages. Consistently, previous studies identified several conserved non-coding regions
78 (CNRs) within the different vertebrate Hox clusters (Kim *et al*., 2000; Chiu *et al*., 2002;
79 Santini *et al*., 2003; Richardson *et al*., 2007) and between amphioxus and vertebrates
80 (Amemiya et al., 2008) using phylogenetic footprinting, an approach based on the

principle that functional non-coding sequences are under strong purifying selection, and thus accumulate less mutations than the non-functional non-coding stretches (reviewed in Wasserman and Sandelin, 2004). Nonetheless, phylogenetic footprinting between not closely related clades has been rarely performed because long phylogenetic distances dilute similarities of non coding regions to the extent that identification of conservation is not possible at the primary sequence level, as is generally the case for amphioxus and vertebrates. A way to improve this methodology over these long phylogenetic distances may be the inclusion of another amphioxus species, in order to break the long branch from the chordate ancestor to the extant amphioxus.

Here, we report the sequence of ~192 Kb of the *Hox* genomic sequences of the European amphioxus *Branchiostoma lanceolatum* including the 15 *BlHox* coding regions and surrounding areas. We have compared it to the American *B. floridae* cluster (Amemiya *et al*., 2008; Holland *et al*., 2008) and with the vertebrate clusters. We identified several blocks of conservation with human Hox clusters which may probably be ancestral regulatory regions of the primitive chordate cluster.

Material and methods

Genomic library screening

A Lambda Fix II/Xho I genomic library (Stratagene) of *B. lanceolatum* (Cañestro *et al*., 2000) was intensively screened with $^{[\alpha\text{-}32P]}$dCTP labelled probes by random-hexamer priming. $6x10^5$ recombinant phages were screened at high-stringency conditions (65º C). As a probe for the primary screening an equimolecular mixture corresponding to the sequence from the second and third helixes of the homeobox from *B. floridae Hox1*0 (*BfHox10*) to *BfHox1* genes, amplified by PCR using degenerated primers SO1 (5'-GARYTNGARAARGARTT-3') and SO2 (5'-CKNCKRTTYTGRAACCA-3') (Garcia-Fernandez and Holland, 1994) was used. This strategy allowed the isolation of the corresponding *Hox*-containing phages at once, and then a specific secondary screening for each gene was performed. Missing genes were isolated individually by screening of the genomic library using the correspondent labelled exon 1. Sixteen positive phage clones were isolated and DNA was extracted as described by (Yamamoto *et al*., 1970). The genomic insert was subcloned into *Not*I restriction site of pBlueScript *SK*II$^+$ vector. All the clones were tested by Southern hybridization. Inserts were characterized by restriction mapping and entirely sequenced on both strands by chromosome-walking or by randomly interspersed primer-binding sites technology using a Tn7 transposon-based system (GPS®-1 Genome Priming System, NewEngland BioLabs) and the assembly made by Phred, Phrap and Consed software (Ewing and Green, 1998; Ewing *et al*., 1998; Gordon *et al*., 1998).

Each *B. lanceolatum* Hox (BlHox) gene was annotated by aligning it with the coding region of the correspondent gene from *B. floridae* from Amemiya *et al*. (2008) (accession numbers AC129909, AC129910, AC124817, AC124805 and AC214474) by CLUSTALW. *B.lanceolatum* sequences reported here have been deposited in GenBank under the accession numbers XXXXX to XXXXX.

Phylogenetic footprinting analyses

For interspecies DNA sequence comparison, we used the *B. floridae* Hox cluster sequence from Amemiya *et al*. (2008) and the *B. floridae* genome assembly (v. 1.0) (http://genome.jgi-psf.org/Brafl1/Brafl1.home.html) for the *BfHox15* genomic locus (Holland *et al*., 2008). We retrieved the human HoxA, and HoxD clusters from NCBI database (accession numbers NT_007819, and NT_005403, respectively; (Int. Human Genome Seq. Con., 2004).

Phylogenetic footprinting was performed using the mVISTA (Mayor *et al*., 2000) through the program AVID (Bray *et al*., 2003) when comparing *B. floridae* and *B. lanceolatum* sequences,with parameters values of 100bp for window size, and 70% for identity threshold; and SHUFFLE-LAGAN (Brudno *et al*., 2003) when comparing amphioxus and vertebrates, with lower stringency conditions of window size of 50 bp, and identity threshold of 60%.

Results and discussion

The Hox complement of the European amphioxus

We characterized the Hox genes of the European species *B. lanceolatum* and the surrounding genomic sequences and compare them to the orthologous sequences of *B. floridae* and vertebrates. We isolated 16 clones from a lambda-phage genomic library from *B. lanceolatum* (Cañestro *et al*., 2000) (Fig.1), with an average insert length of 15 Kb. In total, we have sequenced 191825 unique bp of the putative Hox cluster of *B. lanceolatum*, which corresponds to ~40% of the complete cluster length in *B. floridae*. Our study is the largest genomic sequence analyses of the *B. lanceolatum* genome.

Altogether, they comprise 14 complete Hox genes and a part of *BlHox11*, for which we were able to recover a phage with *BlHox12*, the intergenic region between *BlHox11* and *BlHox12*, and the first exon of *BlHox11* (Fig. 1). Moreover, the clone containing *BlHox4* also contains the microRNA *miR-10* which was previously noted in the *B. floridae* genome analyses (Amemiya *et al*., 2008). We also found the ortholog of the new reported Hox15 gene (Holland *et al*., 2008), confirming that *B. floridae* Hox15

160 was not a species-specific oddity. Therefore, *B. lanceolatum* most probably possesses
161 the 15 Hox genes, as *B. floridae*, grouped in a single cluster.
162
163 The coding regions of Hox genes are highly conserved between the two
164 amphioxus species, with more than 96.14% of average similarity, and almost 100%
165 identity in the homeodomain at the amino acid level. Gene length in both species is
166 almost identical, with some indels that comprise no more than 4 amino acids. This
167 extremely high conservation, suggests that a strong selective pressure maintain intact
168 this family in cephalochordates, mainly in the homedomain, the main functional domain
169 of the protein.
170
171 *Conserved non-coding sequences between amphioxus species and vertebrate Hox*
172 *clusters.*
173 Phylogenetic footprinting allows narrowing down putative functional elements
174 by identifying conserved non-coding regions (CNRs) (reviewed in Wasserman and
175 Sandelin, 2004). Using this strategy, we first performed comparative genomic analysis
176 between the two sibling amphioxus species. We aligned both upstream and downstream
177 intergenic regions of *BfHox*'s with our sequences of *BlHox* genes (~15 Kb of genomic
178 sequence for each gene including the open reading frame) using the mVISTA (see
179 Materials and Methods). Strikingly, the intergenic regions between the two amphioxus

180 species are extremely conserved (Fig. 2), despite 100 Myr of independent evolution
181 (Nohara *et al*., 2005; Kon *et al*., 2007). The percentage of identity in the intergenic
182 regions averages 70%, with some stretches reaching more than 80%, as the ~5Kb
183 between *BlHox2* and *BlHox3*, suggesting that some of these regions may be under
184 functional constraints. This analysis comprises the first comparative genomic analysis
185 between two amphioxus species made so far.

186 Interestingly, the conservation of intergenic regions is higher for genes located at
187 3' of the cluster compared to genes located in the 5' portion (Fig. 3). A similar 3'-to-5'
188 gradient has been reported in vertebrate *Hox* clusters (Santini *et al*., 2003). Santini *et al*.
189 (2003) suggested that this different conservation between anterior and posterior parts of
190 the cluster is due to the fact that anterior and central Hox genes are expressed in and are
191 important for the patterning of the rhombencephalon, whereas posterior genes are
192 involved in the development of relatively less constrained structures. Our group
193 previously named this relaxing constraint of the posterior part of the Hox cluster
194 'posterior flexibility' (Ferrier et al. 2000). Similarly, anterior and central amphioxus
195 Hox genes are responsible for neural tube A-P patterning (Wada *et al*., 1999; Schubert
196 *et al*., 2006), and thus mutations in the regulatory elements controlling these genes may
197 be under stronger negative selective pressures.
198

We also compared *B. lanceolatum* sequences with the corresponding regions of the vertebrate counterparts of the human HoxA and HoxD clusters by means of mVISTA. We identified several short regions conserved between *B. lanceolatum* and at least one human Hox cluster (summarized in tables 1 and S1), using unrestrictive conditions (see material and methods). Some of these CNRs are located close to the 5' end of the coding sequences, being proximal promoter or UTR elements. For example, we identified a conserved element in the 5' UTR of *BlHox3*, coincident with a previously described putative functional element, conserved throughout vertebrate evolution (Santini *et al*., 2003).

Another subgroup of identified CNRs was located within introns. Particularly interesting is the case of *Hox4*. Members of PG4 and PG7 in vertebrates and *Drosophila* (*Dfd* and *Ubx* gene, respectively) harbor a highly conserved element, the HB1 (Haerry and Gehring, 1996; 1997; Santini *et al*., 2003) that consists of a cluster of 3 homeobox recognition sites. HB1 is responsible for the autoregulation of Hox4 and cross-regulation by other Hox genes (Haerry and Gehring, 1997; Packer *et al*., 1998).

Interestingly, we have identified non-coding sequences that are conserved between *B. lanceolatum* and both HoxA and HoxD intergenic sequences (highlighted elements in tables 1 and S1), strongly suggesting a functional role for these elements.

Previously, Amemiya *et al*. performed a similar analysis comparing *B. floridae* and vertebrates (Amemiya *et al*. 2008) using *tracker* software to detect CNRs (Prohaska *et*

219 *al*., 2004), finding little conservation of non-coding sequences between amphioxus and
220 gnathostome Hox clusters, being the majority of these footprints conserved between
221 amphioxus and one of the four paralogous vertebrate Hox clusters. The finding of these
222 likely functional elements (e.g. HB1, that has the same function in both *Drosophila* and
223 mouse), using both types of software, shows that these kind of analyses are useful to
224 identify functional CNRs, even among distantly related species.
225
226 An interesting fact is that most of the elements are found to be conserved with
227 just one human cluster. This suggests that after cluster duplication, both genes and
228 regulatory elements were differentially lost in the different clusters, consistent with a
229 DDC model (duplication-degeneration-complementation; Force *et al*., 1999). Therefore,
230 the results of this study and previous works (Amemiya *et al*., 2008) underscore the
231 importance of the amphioxus Hox cluster to unravel the regulatory complement of the
232 ancestral chordate Hox cluster.
233
234 In summary, we have confirmed the ancient origin of the posterior amphioxus
235 Hox genes (i. e. Hox15) since they are present in both cephalochordate species. We
236 have performed the first large-scale genomic comparison in non-vertebrate chordates,
237 providing insights into the different pathways that non coding sequences and gene
238 regulation have followed within the chordate phylum. Amphioxus-human comparison

239 revealed several CNRs despite more than 500 million years of separate evolution,
240 strongly indicating a functional relevance for these elements. These results, together
241 with the similar expression patterns of amphioxus and vertebrates Hox genes, suggest
242 that some of the elements regulating Hox genes and shaping the basic vertebrate
243 bauplan date back to the origin of chordates.
244
245
246 **Acknowlegments**
247
248 We wish to thank Manuel Irimia and Nacho Maeso for critical reading of the
249 manuscript and Senda Jiménez-Delgado for helpful discussions; Ricard Albalat for
250 kindly providing the lambda genomic library of *B. lanceolatum*, and Jon Permanyer for
251 his help in using the GPS®-1 Genome Priming System and Phred/Phrap/Consed
252 software. This research is supported by grant BFU2005-00252 from Ministerio de
253 Educación y Ciencia, Spain. J.P.-A. holds a FI fellowship of the *Generalitat de*
254 *Catalunya* and S.D'A. a 'Juan de la Cierva' postdoctoral contract of the *Ministerio de*
255 *Educación y Ciencia*, Spain.
256
257 **References**
258

Amemiya, C.T., Prohaska, S.J., Hill-Force, A., Cook, A., Wasserscheid, J., Ferrier, D.E.K., Pascual-Anaya, J., Garcia-Fernàndez, J., Dewar, K., and Stadler, P.F. (2008) The amphioxus *Hox* cluster: characterization, comparative genomics, and evolution. J Exp Zoolog (Mol Dev Evol) 310B:n/a.

Bray, N., Dubchak, I., and Pachter, L. (2003) AVID: A Global Alignment Program. Genome Res 13:97-102.

Brudno, M., Do, C.B., Cooper, G.M., Kim, M.F., Davydov, E., Program, N.C.S., Green, E.D., Sidow, A., and Batzoglou, S. (2003) LAGAN and Multi-LAGAN: Efficient Tools for Large-Scale Multiple Alignment of Genomic DNA. Genome Res 13:721-731.

Cañestro, C., Hjelmqvist, L., Albalat, R., Garcia-Fernàndez, J., González-Duarte, R., and Jornvall, H. (2000) Amphioxus alcohol dehydrogenase is a class 3 form of single type and of structural conservation but with unique developmental expression. Eur J Biochem 267:6511-6518.

Chiu, C.-h., Amemiya, C., Dewar, K., Kim, C.-B., Ruddle, F.H., and Wagner, G.P. (2002) Molecular evolution of the HoxA cluster in the three major gnathostome lineages. Proc Natl Acad Sci U S A 99:5492-5497.

Duboule, D. (1994) Temporal colinearity and the phylotypic progression: a basis for the stability fo a vertebrate Bauplan and the evolution of morphologies through heterochrony. Dev Suppl:135-142.

Duboule, D. (2007) The rise and fall of Hox gene clusters. Development 134:2549-2560.

Duboule, D., and Dollé, P. (1989) The structural and functional organization of the murine HOX gene family resembles that of Drosophila homeotic genes. EMBO J 8:1497-1505.

Ewing, B., and Green, P. (1998) Base-Calling of Automated Sequencer Traces Using Phred. II. Error Probabilities. Genome Res 8:186-194.

Ewing, B., Hillier, L., Wendl, M.C., and Green, P. (1998) Base-Calling of Automated Sequencer Traces Using Phred. I. Accuracy Assessment. Genome Res 8:175-185.

Ferrier, D.E.K., Minguillon, C., Holland, P.W.H., and Garcia-Fernandez, J. (2000) The amphioxus Hox cluster: deuterostome posterior flexibility and Hox14. Evol Dev 2:284-293.

Force, A., Lynch, M., Pickett, F.B., Amores, A., Yan, Y.L. and Postlethwait, J. (1999). Preservation of duplicate genes by complementary, degenerative mutations. Genetics 151: 1531-45.

Garcia-Fernandez, J., and Holland, P.W.H. (1994) Archetypal organization of the amphioxus Hox gene cluster. Nature 370:563-566.

Gomez-Skarmeta, J.L., Lenhard, B., and Becker, T.S. (2006) New technologies, new findings, and new concepts in the study of vertebrate cis-regulatory sequences. Dev Dyn 235:870-885.

Gordon, D., Abajian, C., and Green, P. (1998) Consed: A Graphical Tool for Sequence Finishing. Genome Res 8:195-202.

Haerry, T.E., and Gehring, W.J. (1996) Intron of the mouse Hoxa-7 gene contains conserved homeodomain binding sites that can function as an enhancer element in Drosophila. Proc Natl Acad Sci U S A 93:13884-13889.

Haerry, T.E., and Gehring, W.J. (1997) A conserved cluster of homeodomain binding sites in the mouse Hoxa-4 intron functions in Drosophila embryos as an enhancer that is directly regulated by Ultrabithorax. Dev Biol 186:1-15.

Holland, L.Z., Albalat, R., Azumi, K., Benito-Gutiérrez, È., Blow, M.J., Bronner-Fraser, M., Brunet, F., Butts, T., Candiani, S., Dishaw, L.J., Ferrier, D.E.K., Garcia-Fernàndez, J., Gibson-Brown, J.J., Gissi, C., Godzik, A., Hallböök, F., Hirose, D., Hosomichi, K., Ikuta, T., Inoko, H., Kasahara, M., Kasamatsu, J., Kawashima, T., Kimura, A., Kobayashi, M., Kozmik, Z., Kubokawa, K., Laudet, V., Litman, G.W., McHardy, A.C., Meulemans, D., Nonaka, M., Olinski, R.P., Pancer, Z., Pennacchio, L.A., Pestarino, M., Rast, J.P., Rigoutsos, I., Robinson-Rechavi, M., Roch, G., Saiga, H., Sasakura, Y., Satake, M., Satou, Y., Schubert, M., Sherwood, N., Shiina, T., Takatori, N., Tello, J., Vopalensky, P., Wada, S., Xu, A., Ye, Y., Yoshida, K., Yoshizaki, F., Yu, J.-K., Zhang, Q., Zmasek, C.M., Putnam, N.H., Rokhsar, D.S., Satoh, N., and Holland, P.W.H. (2008) The amphioxus genome illuminates vertebrate origins and cephalochordate biology. Genome Res In press.

International Human Genome Sequence Consortium. (2004) Finishing the euchromatic sequence of the human genome. Nature 431:931-945.

Kessel, M., and Gruss, P. (1991) Homeotic transformations of murine vertebrae and concomitant alteration of Hox codes induced by retinoic acid. Cell 67:89-104.

Kim, C.B., Amemiya, C., Bailey, W., Kawasaki, K., Mezey, J., Miller, W., Minoshima, S., Shimizu, N., Wagner, G., and Ruddle, F. (2000) Hox cluster genomics in the horn shark, Heterodontus francisci. Proc Natl Acad Sci U S A 97:1655-1660.

Kon, T., Nohara, M., Yamanoue, Y., Fujiwara, Y., Nishida, M., and Nishikawa, T. (2007) Phylogenetic position of a whale-fall lancelet (Cephalochordata) inferred from whole mitochondrial genome sequences. BMC Evolutionary Biology 7:127.

Lewis, E.B. (1978) A gene complex controlling segmentation in Drosophila. Nature 276:565-570.

Mayor, C., Brudno, M., Schwartz, J.R., Poliakov, A., Rubin, E.M., Frazer, K.A., Pachter, L.S., and Dubchak, I. (2000) VISTA : visualizing global DNA sequence alignments of arbitrary length. Bioinformatics 16:1046-1047.

Nohara, M., Nishida, M., and Nishikawa, T. (2005) New Complete Mitochondrial DNA Sequence of the Lancelet Branchiostoma lanceolatum (Cephalochordata) and the Identity of this Species' Sequences. Zoological Science 22:671-674.

Packer, A.I., Crotty, D.A., Elwell, V.A., and Wolgemuth, D.J. (1998) Expression of the murine *Hoxa4* gene requires both autoregulation and a conserved retinoic acid response element. Development 125:1991-1998.

Podlasek, C., Houston, J., McKenna, K.E., and McVary, K.T. (2002) Posterior Hox gene expression in developing genitalia. Evol Dev 4:142-163.

Prohaska, S.J., Fried, C., Flamm, C., Wagner, G.P., and Stadler, P.F. (2004) Surveying phylogenetic footprints in large gene clusters: applications to Hox cluster duplications. Mol Phylogenet Evol 31:581-604.

Richardson, M.K., Crooijmans, R.P., and Groenen, M.A. (2007) Sequencing and genomic annotation of the chicken (Gallus gallus) Hox clusters, and mapping of evolutionarily conserved regions. Cytogenet Genome Res 117:110-119.

Santini, S., Boore, J.L., and Meyer, A. (2003) Evolutionary conservation of regulatory elements in vertebrate Hox gene clusters. Genome Res 13:1111-1122.

Schubert, M., Holland, N.D., Laudet, V., and Holland, L.Z. (2006) A retinoic acid-Hox hierarchy controls both anterior/posterior patterning and neuronal specification in the developing central nervous system of the cephalochordate amphioxus. Developmental Biology 296:190-202.

Wada, H., Garcia-Fernandez, J., and Holland, P.W.H. (1999) Colinear and segmental expression of amphioxus Hox genes. Dev Biol 213:131-141.

Wagner, G.P., and Lynch, V.J. (2005) Molecular evolution of evolutionary novelties: the vagina and uterus of therian mammals. J Exp Zoolog B Mol Dev Evol 304B:580-592.

Wasserman, W.W., and Sandelin, A. (2004) Applied bioinformatics for the identification of regulatory elements. Nat Rev Genet 5:276-287.

Yamamoto, K.R., Alberts, B.M., Benzinger, R., Lawhorne, L., and Treiber, G. (1970) Rapid bacteriophage sedimentation in the presence of polyethylene glycol and its application to large-scale virus purification. Virology 40:734-744.

Zakany, J., and Duboule, D. (2007) The role of Hox genes during vertebrate limb development. Curr Opin Genet Dev 17:359-366.

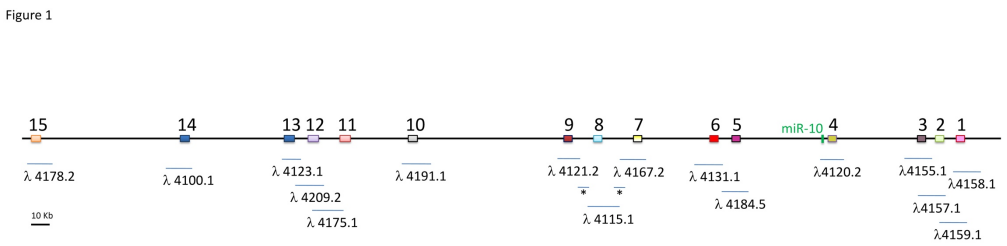


Figure 1. Amphioxus Hox cluster. Position of the 16 lambda phage clones (blue lines) with *B. lanceolatum* Hox genes isolated in this study with respect the *B. floridae* Hox cluster (black line). Asterisk indicates PCR-amplified fragment.

Figure 2

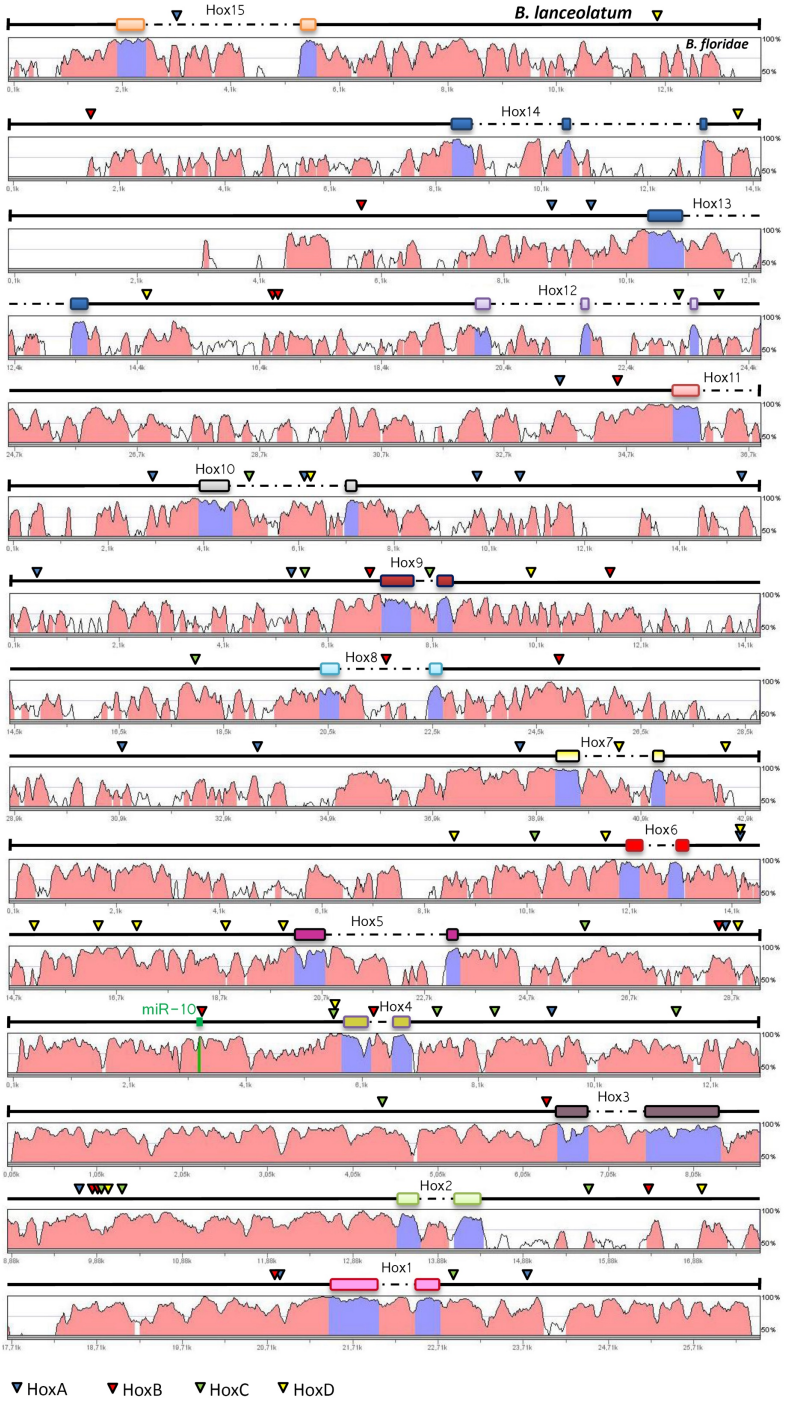


Figure 2. VISTA plot of the comparison between Hox genes of *B. lanceolatum* (base genome) and *B. floridae*. Boxes represent exons in the base genome and dot-lines the introns. Conservation in intergenic regions (pink) is higher in the 3' part (anterior) of the clusters. Exons are in purple in the plot.

Figure 3

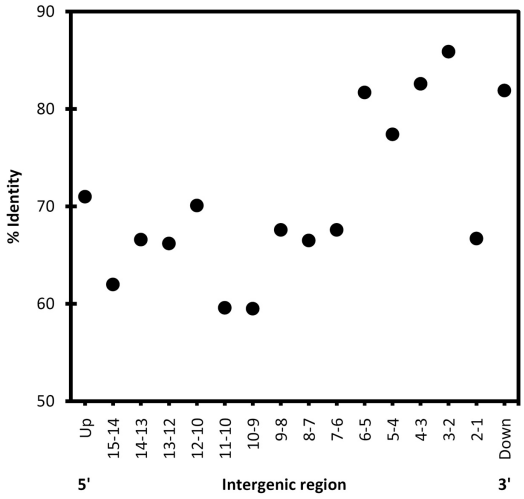
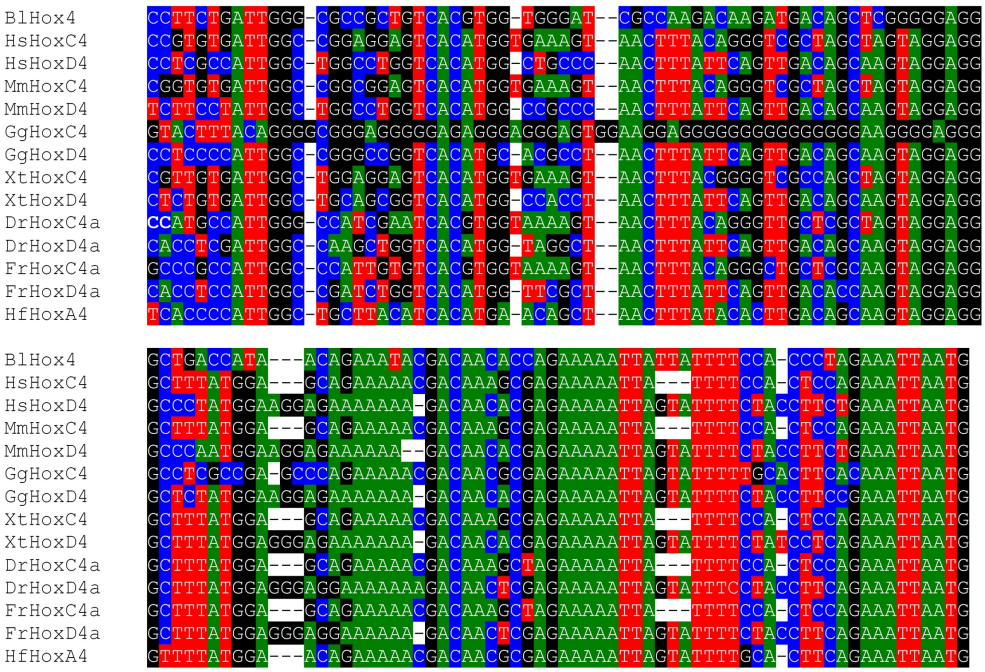


Figure3. Percentage of identity between the intergenic regions of *B. lanceolatum* obtained in this work, and the corresponding ones of *B. floridae*. It is significant the difference of conservation between anterior and posterior intergenic regions, including the outside proximal regions of the cluster. The stretch between Hox1 and Hox2 escape for such conservation, a result also observed when comparing HoxA clusters of vertebrates (Santini *et al*., 2003).

Figure 4



390
391 Figure4. Alignment of Hox4 upstream CNR sequence conserved with amphioxus
392 (*Branchiostoma lanceolatum*), human (*Homo sapiens*), mouse (*Mus musculus*),
393 chicken (*Gallus gallus*), frog (*Xenopus tropicalis*), fish (*Danio rerio* and *Fugu*
394 *rubripes*), and shark (*Heterodontus francisci*)
395

Table 1. This CNEs are obtained using VISTA software with a window size of 50 bp and a identity of 60%. Highlighted regions in amphioxus are conserved also in cluster HoxD.

[a]The positions in amphioxus are calculated for a virtual *B. lanceolatum*'s cluster that resulted of the joining of the different sequences of this work in order.

[b]The positions in Human HoxA cluster correspond to the sequence from NT_007819.

Table S1. This CNEs are obtained using VISTA software with a window size of 50 bp and a identity of 60%. Highlighted regions in amphioxus are conserved also in cluster HoxA.

[a]The positions in amphioxus are calculated for a virtual *B. lanceolatum*'s cluster that resulted of the joining of the different sequences of this work in order.

[b]The positions in Human HoxD cluster correspond to the sequence from NT_005403.

411

412 **Tabla 1.** CNEs between Hox sequences of *B. lanceolatum* and Human HoxA cluster.

Amphioxus[a]		**Human**[b]		**Length (bp)**	**% Identity**	**Position in *BlHox* cluster**
Initial position	Final position	Initial position	Final position			
4547	4596	3177	3226	50	62.0%	intron BlHox15
4828	4893	3903	3973	71	67.6%	intron BlHox15
31255	31311	8659	8713	58	60.3%	intergenic 14-13
31918	31982	9356	9422	67	65.7%	intergenic 14-13
32939	32991	10219	10270	55	61.8%	intergenic 14-13
33015	33064	10275	10317	50	60.0%	intergenic 14-13
33663	33734	10641	10714	74	60.8%	intergenic 14-13
57840	57911	33781	33848	73	63.0%	intergenic 12-11
59113	59185	36655	36727	79	62.0%	intergenic 12-11
70912	70973	50828	50891	64	60.9%	intergenic 9-10
71558	71604	53262	53312	51	60.8%	intergenic 9-10
71840	71889	54825	54871	50	60.0%	intergenic 9-10
72896	72975	57573	57652	80	67.5%	intergenic 9-10
72974	73058	35084	35001	87	65.5%	intergenic 9-10
73676	73725	34402	34354	50	60.0%	intergenic 9-10
75154	75239	33182	33102	86	60.5%	intergenic 9-10
75734	75776	32563	32514	50	62.0%	intergenic 9-10
76880	76927	36736	36790	57	63.2%	intergenic 9-10
77371	77420	37091	37141	51	62.7%	intergenic 9-10
83196	83241	24921	24870	53	62.3%	intergenic 9-10
93418	93479	51316	51377	62	61.3%	intergenic 9-8
94507	94552	55104	55152	51	60.8%	intergenic 9-8
95027	95077	56700	56754	55	60.0%	intergenic 9-8
106935	106986	33948	33996	55	65.5%	intergenic 8-7
107094	107161	34097	34156	69	60.9%	intergenic 8-7
107218	107276	34191	34252	62	62.9%	intergenic 8-7
110364	110423	36948	37010	67	61.2%	intergenic 8-7
110968	111018	37730	37779	51	60.8%	intergenic 8-7
113262	113344	42194	42280	92	63.0%	intergenic 8-7
114534	114584	43200	43251	52	63.5%	intergenic 8-7
133507	133558	59907	59958	53	62.3%	intergenic 6-5
146515	146569	66071	66124	61	62.3%	intergenic 5-miR10
147779	147849	67462	67523	71	63.4%	intergenic 5-miR10
154120	154169	74072	74123	52	78.8%	intergenic miR10-4
154742	154787	74986	75035	50	64.0%	intron *BlHox4*
157789	157841	82989	83035	53	62.3%	intergenic 4-3
171553	171601	97469	97519	52	63.5%	intergenic 3-2
178992	179039	104929	104978	50	64.0%	intergenic 2-1
182229	182280	107490	107543	54	61.1%	intergenic 2-1
182701	182746	108854	108903	52	63.5%	intergenic 2-1
185125	185206	113618	113703	88	62.5%	Downstream

Tabla S1. CNEs between Hox sequences of *B. lanceolatum* and Human HoxD cluster.

Amphioxus[a]		**Human**[b]		**Length (bp)**	**% Identity**	**Position in *BlHox* cluster**
Initial position	Final position	Initial position	Final position			
4843	4911	23739	23671	69bp	68.1%	intron *BlHox15*
4931	4989	23591	23527	65bp	61.5%	intron *BlHox15*
12010	12060	8266	8320	55bp	61.8%	intergenic 15-14
12236	12282	8459	8507	50bp	60.0%	intergenic 15-14
14181	14230	9224	9268	50bp	62.0%	intergenic 15-14
19642	19688	10992	11043	52bp	61.5%	intergenic 15-14
61190	61243	34550	34598	54bp	63.0%	intergenic 11-10
67474	67523	36691	36740	52bp	65.4%	intron *BlHox10*
72934	73037	19324	19434	113bp	61.9%	intergenic 10-9
75385	75447	23668	23730	63bp	71.4%	intergenic 10-9
76837	76891	30387	30442	57bp	63.2%	intergenic 10-9
77438	77485	31857	31907	51bp	60.8%	intergenic 10-9
87560	87614	39044	39095	56bp	60.7%	intergenic 9-8
88796	88845	39340	39391	52bp	61.5%	intergenic 9-8
99832	99899	48211	48277	69bp	65.2%	intergenic 8-7
116515	116566	65555	65599	52bp	67.3%	intron *BlHox7*
118511	118561	52430	52481	52bp	61.5%	intergenic 7-6
118808	118858	66420	66467	51bp	60.8%	intergenic 7-6
119166	119214	66523	66573	51bp	66.7%	intergenic 7-6
126645	126692	59109	59152	50bp	60.0%	intergenic 7-6
127861	127919	69669	69728	60bp	63.3%	intergenic 7-6
128500	128550	60523	60578	57bp	61.4%	intergenic 7-6
130097	130145	62287	62337	54bp	64.8%	intergenic 7-6
132806	132852	56187	56237	51bp	60.8%	intergenic 6-5
133475	133539	57043	57111	70bp	61.4%	intergenic 6-5
134265	134315	68987	69029	51bp	64.7%	intergenic 6-5
135548	135608	70374	70427	61bp	63.9%	intergenic 6-5
136285	136328	74821	74869	50bp	60.0%	intergenic 6-5
139161	139223	63028	63089	64bp	60.9%	intergenic 6-5
145746	145793	56160	56207	50bp	60.0%	intergenic 5-miR10
147159	147203	57798	57847	50bp	60.0%	intergenic 5-miR10
148047	148103	59009	59064	57bp	63.2%	intergenic 5-miR10
154089	154169	64525	64608	86bp	70.9%	intergenic miR10-4
170143	170189	86177	86224	50bp	60.0%	intergenic 3-2
170649	170692	79497	79546	50bp	60.0%	intergenic 3-2
170843	170892	79658	79708	52bp	61.5%	intergenic 3-2
171376	171436	87227	87286	62bp	62.9%	intergenic 3-2
171595	171642	87430	87480	51bp	60.8%	intergenic 3-2
174425	174472	90181	90230	50bp	60.0%	intergenic 3-2
177628	177683	88704	88758	56bp	62.5%	intergenic 2-1
178347	178404	90188	90250	64bp	60.9%	intergenic 2-1