For submission to Genome Research

# Evolutionary conservation of regulatory elements in vertebrate *HOX* gene clusters

Simona Santini[1] Jeffrey L. Boore[2] and Axel Meyer[1]

[1]Department of Biology, University of Konstanz, 78457 Konstanz, Germany

[2] Laboratory of Genomic Diversity, DOE Joint Genome Institute, Lawrence Berkeley National Laboratory, and University of California, Berkeley

Corresponding author: Axel Meyer

Phone: +49 7531 884163

Fax: +49 7531 883018

E-mail: axel.meyer@uni-konstanz.de

Keywords: *Hox*A cluster, comparative genomics, conserved noncoding sequence

ABSTRACT

Due to their high degree of conservation, comparisons of DNA sequences among evolutionarily distantly-related genomes permit to identify functional regions in noncoding DNA. *Hox* genes are optimal candidate sequences for comparative genome analyses, because they are extremely conserved in vertebrates and occur in clusters. We aligned (Pipmaker) the nucleotide sequences of *Hox*A clusters of tilapia, pufferfish, striped bass, zebrafish, horn shark, human and mouse (over 500 million years of evolutionary distance). We identified several highly conserved intergenic sequences, likely to be important in gene regulation. Only a few of these putative regulatory elements have been previously described as being involved in the regulation of *Hox* genes, while several others are new elements that might have regulatory functions. The majority of these newly identified putative regulatory elements contain short fragments that are almost completely conserved and are identical to known binding sites for regulatory proteins (Transfac). The conserved intergenic regions located between the most rostrally expressed genes in the developing embryo are longer and better retained through evolution. We document that presumed regulatory sequences are retained differentially in either Aa or Ab clusters resulting from a genome duplication in the fish lineage. This observation supports both the hypothesis that the conserved elements are involved in gene regulation and the Duplication-Deletion-Complementation model.

# INTRODUCTION

Understanding the mechanisms that underlie gene regulation is one of the major goals of comparative genomics as well as developmental biology. The functions of *cis*-acting regulatory sequences that are located in noncoding regions of DNA are still not well-understood (Clark, 2001). Comparative DNA sequence analyses have become increasingly important since the high degree of conservation of regulatory elements was first recognized (e.g., Aparicio *et al.*, 1995; Manzanares *et al.*, 2000, ). The conservation of protein coding sequences even among evolutionarily distantly related organisms, presumably as a result of purifying selection, has been noted before (e.g., Hardison *et al.*, 1997; Brenner *et al*., 2002). However, only a small portion of organisms' genomes encodes information for proteins. A large portion of the genome (up to 97%, Onyango *et al*., 2000) is noncoding DNA and a hereto forth unknown part of it plays a role in regulating gene expression. The identification of functional elements in noncoding DNA sequences is often complicated by the fact that these elements are typically short (6-15 bp, e.g., Carroll *et al*., 2001) and reside at varying distances from their target gene. Fortunately, among noncoding sequences, functional elements tend to evolve at a slower rate than non-functional regions, because they are subject to selection (Cliften *et al.*, 2001). Due to of this slower rate of evolution, comparisons among evolutionarily distantly-related genome sequences could provide a tool to identify functional regions in noncoding DNA (Tompa 2001, Blanchette and Tompa, 2002). This approach has been termed phylogenetic footprinting (Roth *et al.*, 1998; Venkatesh *et al*., 2000; Cliften *et al.*, 2001). Comparisons among closely related organisms, such as different species of *Saccharomyces* (Cliften *et al.*, 2001) or *Drosophila* (Bergman *et al.*, 2001) have been successfully used to identify regulatory regions, although deeper comparisons such as between human and mouse (Onyango *et al.*, 2000), with an evolutionary distance of approx. 80 million years (Pough, 1999) show many of the functionally

relevant binding sites with a high degree of conservation (on average 93.2%, Wassermann *et al*., 2000) in an otherwise nearly randomized background.

In comparisons among closely related species, many non-functional noncoding sequences will also show a high degree of nucleotide identity, rendering the identification of DNA regions that are involved in gene regulation more difficult. The alignment of long stretches of DNA sequences from evolutionarily distantly related species permits one to search for regulatory elements, which will stand out from the less conserved non-functional regions. This is due to the decrease in "noise" from faster evolving non-functional regions in the alignment that will make the evolutionarily conserved regulatory elements stand out.

*Hox* gene clusters are among the most suitable candidate sequences to perform this kind of comparative genome analyses, because their nucleotide composition and function are extremely conserved in all vertebrates in which they have been studied. *Hox* genes code for transcription factors believed to be responsible for setting the animal body plans early in embryological development. They specify position for developing fields along the anterior-posterior axis, and are characterized by a 183 bp motif, the homeobox, which encodes a conserved DNA binding structure, the homeodomain (reviewed in Gehring, 1993). Within the homeobox gene superfamily, *Hox* genes are a subfamily that are found to be arranged in genomic clusters and to be colinear in chromosomal arrangement with their time of activation and boundary of expression along the anterior-posterior axis (e.g., Krumlauf, 1994). Given their importance in development it may not be surprising that they are highly conserved.

In addition to their coding sequences, it is furthermore expected that their functional sequences are largely invariant across even big evolutionary distances. They occur in strictly packed clusters, which aids their identification and alignment. One of the selective forces keeping the genes of *Hox* clusters together may stem from the fact that adjacent genes share

common *cis*-regulatory elements (Peifer *et al*., 1987) . Therefore, adjacent genes have to remain closely linked, since translocations or insertions between them would deprive one or the other gene of its *cis*-regulatory elements. Moreover, their occurrence in clusters allows better definition of the regions of sequence in which it is expectable to find regulatory elements.

## RESULTS

We compared four teleost species (*Oreochromis niloticus*, *Fugu rubripes*, *Morone saxatilis* and *Danio rerio*) with two mammalian species (*Homo sapiens* and *Mus musculus*) and an outgroup species, the horn shark (*Heterodontus francisci*). Their *Hox* gene contents are shown in Figure 1. Highly conserved homeobox domains in the *Hox* genes permitted "anchoring" of the clusters with each other. Therefore, it was possible to align *Hox*A clusters on the basis of highly conserved regions of exons and thereby align evolutionarily distantly related genomic sequences in order to characterize regulatory elements.

### Genomic architecture of *Hox*A clusters

Comparisons of gene lengths and distances between genes belonging to the *Hox*A cluster are shown in Figure 2. The single *Hox* cluster region of the cephalochordate amphioxus (haploid DNA content: C = 0.59 pg, Atkin & Ohno, 1967) spans over 400 kb (Ferrier *et al*., 2000; Garcia-Fernandez & Holland, 1994), but is smaller for the *Hox*A clusters of vertebrates that have been studied. The region is only approximately 110 kb (AF224262 and AF479755) in shark (C = 7.25 pg, Stingo *et al.*, 1989) *Hox*A (previously named *Hox*M, orthologue of *Hox*A, Kim *et al*., 2000), 110 kb (AC004079, AC004080 and AC010990) in human *Hox*A (C = 3.50 pg, Tiersch *et al.*, 1989), 105 kb (AC021667) in mouse *Hox*A (C = 3.25 pg, Vinogradov, 1998, *Asif et al*., 2002), 100 kb (AF533976) in tilapia *Hox*A (C = 0.99 pg, Hinegardner 1976), 64 kb (JGI public database) in pufferfish *Hox*A (C = 0.40 pg, Brenner *et al.*, 1993), 62 kb

(AC107365) in zebrafish *Hox*A (C = 1.75 pg, Vinogradov, 1998) and 33 kb in zebrafish *Hox*A (AC107364).

The available striped bass (C = 0.89 pg, Hinegardner 1976) sequence covers only the region from *Hox*A10 to *Hox*A4. The region *Hox*A9to *Hox*A4 in striped bass is 24 kb long (AF089743); the homologous region in tilapia is 23 kb, in pufferfish it is approximately 20 kb, and in the zebrafish A is approximately 19 kb (A does not contain genes 4, 5, 7). In the shark it is 35 kb, and in the human and the mouse it is approximately 36 kb. Consistent with the view that *Hox* clusters are reduced in size for vertebrates, this part of the amphioxus cluster is approximately 135 kb long (Fig. 2).

Genome sizes and lengths of the *Hox*A clusters seem to be correlated (Fig. 3). Lengths of *Hox* clusters have been shown previously to be independent of the pattern of gene loss among several fish species (Aparicio *et al*., 1997; Snell *et al*., 1999; Chiu *et al*., 2002). When the same genes are retained, the architecture of *Hox*A clusters is generally conserved among the species under examination, concerning relative lengths not only of orthologous genes among species, but also of spacing between genes (Fig. 2).

Independent gene losses have happened in fishes genomes (Fig. 2). The pufferfish cluster was initially thought to lack *Hox*A7 (Aparicio *et al.*, 1997), and it was hypothesized that this loss, together with the other members of the entire paralogous group 7 (Aparicio *et al.*, 1997), could have been responsible for the absence of ribs and pelvic fins and girdle in this group of fishes (Holland, 1997; Prince *et al.*, 1998; Meyer 1998; Meyer & Malaga-Trillo, 1999). Our comparisons show conservation of *Hox*A7 exons in pufferfish, with the exception of a 84 bp deletion in the homeobox in exon 2. However, the observation that the homeodomain is lacking its central and most conserved part might argue that in pufferfish the *Hox*A7 gene is a pseudogene.

The zebrafish Aβ cluster lacks *Hox*A7 and contains only a fragment of exon 2 of *Hox*A10. It lacks also *Hox*A2 (Amores *et al*., 1998), but the cluster region corresponding to both *Hox*A2 exons and also to the promoter and the intron still shows nucleotide conservation, suggesting that its loss was a relatively recent event in the zebrafish lineage. The zebrafish Aβ cluster lacks the *Hox*A1 and *Hox*A3, *Hox*A5 and *Hox*A7 genes. The *Hox*Aβ cluster in zebrafish has been subject of more losses of genes than the *Hox*Aβ cluster. Tilapia has an almost complete *Hox*Aβ cluster, in terms of presence of *Hox* genes and no lineage-specific gene losses relative to other teleost fishes are observed. Tilapia *Hox*Aβ cluster retains the *Hox* 2, 7 and 10 genes. We have preliminary evidence also for a *Hox*Aβ cluster in tilapia (*Hox*A2β and *Hox*A3β)and it will be interesting to investigate whether the pattern of gene loss resembles that of the zebrafish.

As can be seen in Table 1, the *Hox*A cluster of mouse alone has a nearly identical content of each nucleotide. For all others examined, nucleotide composition of the *Hox*A cluster is significantly biased (chi squared test, data not shown) in favor of bases A and T.

**Comparison of nucleotide sequence**

All clusters were screened with RepeatMasker to highlight interspersed repeats. There is a complete absence of any kind of long repeats between genes of the *Hox*A clusters in all the examined species. We compared the nucleotide sequence of *Hox*A homologous genes from tilapia *Hox*Aβ, pufferfish *Hox*Aβ, striped bass *Hox*Aβ, zebrafish *Hox*Aβ and *Hox*Aβ, shark *Hox*A, human *Hox*A and mouse *Hox*A clusters. In the Pip output (Fig. 4), coding regions are shown with a blue background, introns in yellow, and conserved noncoding sequences (CNSs, Loots *et al*., 2000) not previously described in the literature in green. The red background refers to conserved regions that have been described previously. As expected, coding sequences show

a particularly high degree of similarity, especially in the second exon (above 75%), which contains the homeobox, in all genes of the cluster among all examined species, while introns are generally less conserved and impossible to align over long regions.

**Identification of CNSs**

Several stretches of sequence outside of the recognized coding regions of the *Hox* genes are highly conserved in all species examined (Fig.4 & Table 2). These CNSs have been maintained for a period of over 500 million years of evolution. The fraction of CNSs for each intergenic region is shown in Table 3.Interestingly, several 5' and 3' untranslated regions adjacent to the *Hox* genes of the clusters are conserved, suggesting that they may play an important role in the transcriptional regulation of the genes that they are flanking. A summary of the identified conserved regions is shown in Table 2. All the identified CNSs have been tested by using BLAST to exclude their presence in other positions of the genomes. No significant (E value<1) alignments have been found out of *Hox* clusters.

Several sequences involved in the regulation of *Hox* genes have been previously described in the literature (Table 2). These sequences have been confirmed also by the method we have used to compare the different clusters as being highly conserved.

The intergenic regions between genes located 3' in the clusters are better conserved than those between genes located 5' in the cluster (Fig. 5, Table 3 & alignment in the supplemental data files on Genome Research web site). The number of conserved nucleotides (over 60% identity) is significantly higher ($P = 0.007$) in the intergenic regions in 3' in the cluster and the detected CNSs are longer.

**Description of some hypothetical regulatory element**

Due to the nature of *cis*-regulating elements, which can be as short as 6 bp (Hardison *et al*., 1997), we were interested in finding where such sequences reach the highest degree of conservation for a even small number of nucleotide.

The first part of the intron of *Hox*A11 (51 bp) of the tilapia sequence is over 80% similar among tilapia, fugu, zebrafish A and A, horn shark, human and mouse (data for this region in striped bass are not available). The fragment presents the consensus homeodomain binding sites HB1 located in the intron of the mouse genes *Hox*A4 and 7 (Haerry and Gehring, 1996). The HB1-element consists of a three homeodomain binding sites (HB1) and it is an evolutionary conserved DNA sequence previously described in the intron of *Hox*A7 (Haerry and Gehring, 1996), in the leader (putative autoregulatory element) of its *Drosophila* homolog *Ubx* and in the introns of the paralogous group 4 *Hox* genes in medaka, chicken, mouse and human (Morrison *et al*., 1997). The HB1 element binds *Drosophila* CAD homeoprotein and CDX-1, its homolog in mouse and it therefore is supposed to be a target for various homeodomain proteins in both vertebrates and invertebrates. Our comparative analyses show that the HB1 element is present not only in the introns of *Hox*A4 and 7 as already described in the literature, but also in the intron of *Hox*A11 in the *Hox*A cluster of all the species examined. Interestingly, it is present also in the intron of *Hox*A11 of zebrafish.

The region responsible for *cis*-regulation of the *Hox*A7 gene has been previously described as an enhancer located 1.6 kb upstream of the coding sequence in human and mouse (Knittel *et al.*, 1995). Knittel *et al*. (1995) hypothesized that another proximal regulatory element can cooperate in the expression of *Hox*A7. Immediately upstream of the *Hox*A7 gene we highlighted a 185 bp stretch with more than 84% sequence identity. Our comparison (Fig. 4) shows that there are several completely conserved sequences within this fragment,

characterized by the short motif GTAAA. This long conserved region might be the regulatory element that Knittel *et al.* (1995) hypothesized.

In the intron of the *Hox*A7 the HB1-element shows a sequence identity of over 80% among the species examined. The region immediately upstream of the *Hox*A5 gene (490 bp) is between 70 and 85% similar. The RARE elements described as "box c" and "box d" by Odenwald *et al.* (1989) in human and mouse can be recognized (Fig. 6). These elements are present, with minor variation, among all *Hox* genes of paralog group 5 and are known regulatory binding sites in the mouse *Hox* 1.3 (*Hox*A5) (Odenwald *et al.*, 1989). The conservation percentages within the single boxes are 88% for the "box c" and 96% for the "box d".

Downstream of the *Hox*A5 gene (1.3 kb) a region of 259 bp has an average similarity of 90%, with two 100% identical stretches of 25 and 33 bp in length. The motifs found in this region are ATGAAT (with a repeat following after 13 bp), ATAAA, $(AAGT)_2$ and $(ACATA)_2$. The motifs identified by our comparisons are similar to those described as binding sites of the paired domain of the *Pax* genes (Epstein *et al.*, 1994) and also of the *Ultrabithorax* gene of *Drosophila* (Ekker *et al.*, 1991). This extremely conserved region had not been previously described as being involved in *Hox*5 and 4 regulation, but the nature and conservation of the long stretches pointed out through the comparison suggest that it might be a good candidate region for functional tests.

Upstream of the *Hox*A4 gene we identified a stretch 154 bp long that has a similarity of 85% and it contains a RARE element (17 bp) which is part of the *Hox*A4 promoter, described by Doerksen *et al.* (1996).

In the intron of gene *Hox*A4 a 68 bp long stretch was found and it contains the previously described HB1 element (Haerry and Gehring, 1996).

Downstream of *Hox*A4 (1.7 kb) a 127 bp long sequence is on average 78% conserved with a 26 bp long stretch that is 96% conserved which contains the AAATAAAA (position 63576-63583) and ATTTAA motifs and a 16 bp stretch that is 94% conserved which contains the motif TTTTATTT (position 63882-63889). It is possibly a palindromic sequence for the one in position 63576. Palindromes are frequently associated with regulatory elements (Chu *et al*., 2001).

Immediately upstream of the gene *Hox*A2 we found a 352 bp region that is 85% conserved that constitutes part of the *Hox*A2 promoter described by Tan *et al.* (1992) in mouse.

The *Krx20* element and the nearby box a, described by Nonchev *et al.* (1996) as being involved in *Hox*A2 *trans*-activation in mouse, present in tilapia *Hox*Acluster (Fig. 7a), was not identified by our alignment. To confirm this result we searched specifically for these elements in zebrafish, pufferfish and horn shark clusters, but we could not identify them.

The AT richness of regulatory regions in *Hox* clusters has been previously described by several authors (e.g., Odenwald *et al.*, 1989; Margalit *et al.*, 1993, Shashikant *et al*., 1995) as a common feature of homeodomain binding sites. The most of the DNA regions that our analyses identified as highly conserved are AT-rich (18/30, equal to 60%, Table 2). Although this observation alone clearly cannot be considered as a definitive evidence for the functionality of these sequences, it provides support for this possibility, in addition to the degree of sequence conservation.

**Identification of previously described functional elements**

Extensive searches of the transcription factor database (Transfac) revealed that several of these short 100% conserved sequences match previously described transcription factor binding sites (Table 2). The matches more frequently obtained are: nuclear factor NF1 binding sites (Rossi *et al.*, 1988), abdominal B (AbdB) homeobox gene binding sites (Ekker *et al.*, 1994),

CdxA homeobox gene binding sites (Margalit *et al.*, 1993) and murine homeodomain binding sites (Catron *et al.*, 1993).

Several of the most conserved sequences are highly similar to known transcription factors binding site motifs. One of those is the *Krx20* binding site, that was found in human, mouse, fugu and tilapia clusters (Fig. 7). *Krx20* binding sites have been described by Nonchev *et al.* (1996) as being involved in *Hox*A2 regulation as an r3/r5 enhancer that up-regulates the expression of those genes in rhombomere3/rhombomere5, where *Krx20* is expressed in human, chick, mouse and pufferfish. The *Krx20* binding site is nine bp long and occurs around 2kb upstream of the genes *Hox*A2 and *Hox*B2, with a high degree of conservation (Fig. 7A). It is closely followed by a 12 bp long conserved sequence motif called "box a", which is highly similar to "box1", the corresponding element associated with *Krx20* binding site in cluster B (Fig. 7B). Box 1 is required for r3/r5 enhancer function in transgenic mice (Vesque *et al.*, 1996).

## DISCUSSION

Our analyses confirm the value of comparative evolutionary genomic approaches in the identification and description of regulatory elements in genomes. We expect that this type of analysis will help to increase our knowledge about the characteristics, evolutionary conservation and the position of functional elements with respect to the genes that they control. Consequently, the development of a set of methods which could considerably the characterization of these elements would be desirable.

We conducted several comparative analyses of the entire *Hox*A clusters for seven species of vertebrates. We sequenced the entire *Hox*A cluster from *O. niloticus*, and compared the position and nucleotide sequence of the genes that constitute that cluster with the other species examined. The complete absence of repetitive element agrees with the idea that one of the

selective forces keeping the genes of *Hox* clusters arranged in tight clusters stems from the fact that adjacent genes share common *cis*-regulatory elements. In fact, it has been suggested that repetitive elements are frequently involved in chromosomal rearrangement processes, such as inversion, translocation and excision (Tomilin, 1999; Moran *et al*., 1999). Hence, the absence of repetitive elements could be interpretated as a result of selective pressure against them, to reduce the risk of such events, which may interrupt *Hox* cluster continuity.

We chose to compare teleost fishes, horn shark and mammals to include distantly related genomes, since their lineages separated approximately 450 millions years ago (e.g., Pough *et al*., 1999). Moreover, teleost fish genomes are typically smaller than those of mammals, and conserved sequences between the two groups tend to be restricted to coding sequences and noncoding regions with transcriptional regulatory roles (Aparicio *et al*., 1995).

In zebrafish, *Hox*A cluster seems to be more prone to gene loss than *Hox*A. The only genes present in the *Hox*A cluster but not in the *Hox*A cluster are *Hox*10 and *Hox*2. On the other hand, the *Hox*5, 4, 3 and 1 genes are present only in *Hox*A. One of two daughter clusters preferentially experienced gene losses events. Alternatively, the *Hox*5, 4 and 3 genes could have been lost in a single event in *Hox* cluster.

**Degree of conservation of intergenic regions**

Our comparative analyses were directed toward identifying conserved blocks of nucleotides between evolutionarily distantly related species that might be *cis*-acting sites for *Hox* gene regulating factors. Intergenic regions have varying degrees of conservation (Table 3). Intergenic spaces between genes located 3' in the clusters are significantly more conserved than in the in 5' portion of the clusters (Fig. 6 & Table 3). This pattern might be explained by the different *Hox* genes expression pattern during development. Genes located in 5' position in the cluster are expressed more posteriorly in the embryo and later in its development, while genes

located in position 3' in the cluster are expressed more anteriorly in the embryo and earlier in its development (Duboule & Dolle', 1989). Genes located 3' in the cluster, namely *Hox*1-4, are expressed in the developing hindbrain. Their regulatory elements are evolutionarily highly conserved as was demonstrated through transgenic experiments (e.g., Frasch *et al*., 1995; Manzanares *et al*., 2000). The intergenic regions of *Hox* genes 3' in the clusters are responsible for the activation of the first and more rostral genes to be expressed during development and therefore their extreme conservation might be necessary to guarantee the correct activation of the whole *Hox* system. We found a significant increase in length of CNSs between pairs of 3' genes compared to intergenic regions of genes located 5' and not involved in hindbrain segmentation ($P = 0.007$).

In our analyses we include also the noncoding regions upstream of the *Hox*13 gene and downstream of the *Hox*1 gene. Intergenic regions between two *Hox* genes contain regulatory elements for genes both upstream and downstream (e.g., Peifer *et al*., 1987). Also if the region upstream of the *Hox*13 gene contains only regulatory elements for this gene, and the same for the region downstream of the *Hox*1 gene, the trend of increase in length of CNSs from 5' to 3' within intergenic regions is still significant.

**Search for regulatory sequences**

Several conserved noncoding regions have been identified in this analysis. All the identified CNSs are specific to Hox clusters (no significant BLAST alignment with any other region of the genome, E value<1).

Some of these regions reside immediately 5' and 3'of the genes of the *Hox* clusters and this feature is generally related to functional roles (e.g., reviewed by Maconochie *et al.*, 1996). Promoters are located immediately 5' upstream of genes (e.g., *Hox*A2 promoter, Tan *et al.*, 1992) and RAREs are located 3' of the regulated gene (e.g., Frasch *et al.*, 1995). However, the

largest part of conserved regions we found is located between two genes and quite distant (by 1-5 kb, Table 2) from both. Because of this, these regions are the most interesting, since *cis*-regulatory regions in *Hox* clusters are located in positions that are intermediate between the genes they regulate. An example for this phenomenon is the element named H8/7-6 FCS (Kim *et al.*, 2000) that exists in all four clusters of mammals and shark and in the *Hox*A (at least) cluster of fishes. This element is located 1.2 kb downstream of the *Hox*A7 gene and 3.6 kb upstream of the *Hox*A5 gene in tilapia (Table 2). These *Hox* genes are involved in controlling the development of the branchial region (Krumlauf, 1994). The conservation of the nucleotide sequence and relative position in all clusters examined so far, makes this element an excellent candidate for an evolutionary conserved *cis*-regulatory element. Table 2 lists several other CNSs located between two genes that might contain *cis*-regulatory elements. We could not locate *Krx20* and box a in any CNS through our alignment. The reason is that *Krx20* binding site and box a are short sequences not embedded in a block of at least 50 bp with a conservation of at least 60% in a minimum of 4 clusters. In this particular case out criteria to define CNSs were too strict. Also *Hox*A1 RARE elements described by Langston *et al*. (1997) could not be identified, because the region downstream *Hox*A1 was not available for most of the sequences and then the alignment did not fit the above mentioned criteria for defining CNSs.

All except one of the CNSs identified through our comparisons are present in at least one of the zebrafish *Hox*A clusters and some in both of them (Table 2). A specific CNS is generally conserved in the one of the two zebrafish *Hox*A clusters that still retains the gene located upstream of its position, i.e. the CNS upstream of *Hox*A10 is present only in *Hox*A cluster, which retains the gene *Hox*A10, and was lost in *Hox*A cluster, that does not have the *Hox*10 gene. The same happens with CNSs located upstream of the *Hox*A5, 4 and 3 genes which are present only in the *Hox*A cluster, which still retains those genes. The CNS found immediately

upstream of *Hox*A7 and previously described by Knittel *et al*. (1995) as an enhancer of *Hox*A7 in human and mouse is absent from both the zebrafish cluster. This is particularly interesting, because the *Hox*A7 gene was lost during zebrafish genome evolution. Also the CNS located in the intergenic region between the *Hox*A3 and 2 genes and indicated as 3-2a in Table 2 is absent from both zebrafish clusters. This CNS has one of the lowest overall conservation levels, with none over 95% identity. These observations reinforce the possibility that the CNSs we identified are actually involved in regulatory functions.

The duplication-deletion-complementation model (DDC, Force *et al*., 1999) proposes that duplicated genes retain different sets of regulatory elements. The functions of the initial gene might be divided by the two duplicated "daughter" copies of the gene. The *Hox* 13, 11 and 9 genes are present in two copies in the zebrafish genome, in the *Hox*A and A clusters. The CNSs upstream of these genes are also retained in both the clusters but are different between them. This could indicate that they have been preserved because they are important for the regulation of those genes, but control different patterns of expression, hence accounting for sub-functionalization of the duplicated "daughter" copies of the genes.

Chiu *et al*. (2002) did not observe the same pattern of conservation in zebrafish *Hox*A clusters. That difference might be due to a different method of identification of those sequences. Chiu *et al*. (2002) described, by comparison of human and horn shark *Hox*A clusters, a great number of Phylogenetic Footprints (PFs), which are defined as short blocks of noncoding DNA sequence, typically 6 bp or more, that are 100% conserved in two taxa that have diverged at least 250 million years (Tagle *et al*., 1988, Blanchette and Tompa, 2002). Among those they described as Phylogenetic Footprint Clusters (PFCs) are those that were found close to each other (within 200 bp) and located at comparable distances from the gene that is located 3' to each intergenic region. They found only a small number of PFCs to be

present in at least one of the two zebrafish *Hox*A clusters. They concluded that the essential *Hox* gene functions in zebrafish are performed with different *cis*-regulatory elements (e.g., phenogenetic drift, Weiss & Fullerton, 2000) from those of the ancestral gene, with *cis* elements highly conserved in horn shark and human. We defined a sequence as a CNS using the following criteria (see Materials and Methods): identity over 60% in at least four out of eight clusters; presence in at least two species known to have only one *Hox*A cluster (horn shark, human, mouse, see Fig. 1) and minimum length of 50 base pairs (bp). We identified a smaller number of longer conserved elements, but that are shared by a higher number of species/clusters. Moreover, because of the fact that many *trans*-regulatory elements recognize a core sequence even shorter that 6 bp and with a certain degree of tolerance, we accepted a 95% lower threshold for the short highly conserved sequences we described (Table 2).

**Regulatory elements located in introns**

Intron sequences are typically not conserved among evolutionarily diverged species. A clear exception are the HB1 elements, believed to be binding sites for several homeoproteins (Haerry & Gehring, 1996, 1997). Our analyses show that the HB1 elements, so far described only in the introns of the *Hox*4 and 7 genes, are present also in the intron of the *Hox*11 gene in the *Hox*A cluster (in both *Hox*A and *Hox*A in zebrafish). The *Hox*4, 7 and 11 genes are expressed in different regions of the developing embryos (rhombomeres 6 and 7 in the hindbrain for *Hox*4 paralogous group, thoracic region for *Hox* 7 and caudal region for *Hox* 11) and at different times of development. The spatial regular redundancy of HB1 elements in *Hox* clusters might be related to the different timing of activation of groups of *Hox* genes (anterior, central and caudal) in the developing embryo. It would be of interest to better characterize the function of different HB1 elements within a same *Hox* cluster. Moreover, it would be important

to know if other *Hox* clusters show a similar pattern as the *Hox*A clusters concerning HB1 regulatory elements.

A long (over 600 bp) stretch of intron of gene *Hox*2 is 60-70% conserved among all the species included in this comparison. Part of this sequence matches with a previous described POU protein binding site (Verrijzer *et al*., 1992). The overexpression of homeoprotein POU2 rescues zebrafish *Krx20* and *valentino* mutants (Hauptmann *et al*., 2002), that are caused by disrupted *Hox*2-related patterning of rhombomeres 3/5. It seems likely that *Hox*2 expression and function is related to the conservation of the conservation of the putative regulatory element in its intron.

**Known conserved regions and regulatory elements**

The reliability of our results was confirmed by the observation that some of the highly conserved, possibly functional, noncoding regions that we have identified have been previously described as regulatory elements (Table 2). Moreover, many of them contain homeoprotein binding sites that are believed to be responsible for *Hox* gene regulation (Table 2). It is reasonable that the elements that are evolutionarily conserved are the ones that regulatory proteins bind to and this agrees with the evidence that other classes of homeobox genes are responsible for *Hox* genes regulation. Currently, four groups of transcriptional regulators have been identified that directly regulate *Hox* gene expression in the vertebrate embryo: retinoic acid receptors, *Krx20*, members of the *Pbx/exd* family and the *Hox* genes themselves (reviewed by Lufkin, 1997). Since *Hox* genes have a temporal pattern of differential expression (i.e. *Hox*A1 is expressed before *Hox*A2 and so on), therefore, further studies on homeoprotein binding sites are necessary to define if and how *Hox* genes expressed earlier in embryo development could regulate the expression of *Hox* genes expressed later.

## CONCLUSIONS

It would be particularly interesting to test some of the so far undescribed conserved noncoding regions that we have identified through this comparative genomic approach for a possible functional role in the activation and regulation of *Hox* genes. Since functional studies involve a great deal of effort, e.g., transgenic animals, it is critical to reduce the number of possible candidates for regulatory function. Sequencing projects of whole genomes (e.g., fugu, zebrafish, medaka) offer new possibilities for comparative genomic approaches to study distantly related organisms to uncover putative regulatory elements. Moreover, using distantly related genome comparisons between teleosts and, e.g., mammals or amphioxus highlights the divergence in gene regulation of paralogous genes that evolved subsequent to gene duplication. It is still subject of discussion whether paralogous genes in fishes are due to an early whole genome duplication (Meyer and Schartl, 1999; Taylor *et al.*, 2001), or rather to several independent smaller scale duplication events (Robinson-Rechavi *et al.*, 2001). One of the primary mechanisms by which sub-functionalization of duplicated genes occurs may be through a change in their regulatory elements, whereby mutations or differences in deletions in these elements can lead to differential expression patterns of duplicated genes (Force *et al.*, 1999). The comparison of distantly related genomes may indicate which duplicated genes have divergent regulatory sequences in comparison to organisms for which such a duplication did not occur, as mammals. This in turn would provide a method by which to elucidate different evolutionarily new functions for the duplicated genes.

## MATERIALS AND METHODS

The *Hox* clusters included in this study are: tilapia (*Oreochromis niloticus* AF533976, *Evx*1-*Hox*A1□), pufferfish (*Fugu rubripes*, JGI public database http://www.jgi.doe.gov/programs/fugu/fugu_mainpage.html, *Hox*A13□-*Hox*A1□ ), striped bass

(*Morone saxatilis* AF089743, *Hox*A10α -*Hox*A4α) zebrafish (*Danio rerio* AC107365, Evx1 - *Hox*A1α); zebrafish (*Danio rerio* AC107364, *Hox*A13α -*Hox*A2α); horn shark (*Heterodontus francisci* AF224262 and AF479755 *Hox*M13-*Hox*M1, corresponding to *Hox*A, Kim *et al*., 2000); mouse (*Mus musculus* AC021667, *Hox*A13-*Hox*A1) and *Homo sapiens* (AC004079, AC004080 and AC010990, Evx1-*Hox*A1)

The tilapia *Hox*Aα cluster sequence (Malaga-Trillo & Meyer, 2001) has been used as the template sequence to which the others are compared. It has been filtered for repetitive and other "junk" elements through RepeatMasker, available at University of Washington Genome Center (http://ftp.genome.washington.edu/cgi-bin/RepeatMasker/).

The alignment has been performed using the program MultiPipmaker available at http://bio.cse.psu.edu/pipmaker/. PipMaker (Schwartz *et al.*, 2000) computes alignments of similar regions in two or more DNA sequences. The resulting alignments are summarized with a "percent identity plot", or "pip" for short. All pair wise alignments with the first sequence are computed and then returned as interleaved pips, and it is possible to compute a true multiple alignment of the input sequences to produce a nucleotide-level view of the results. The alignment engine is BlastZ, which is an experimental variant of the Gapped Blast program (Altschul *et al.*, 1997; Zhang *et al.*, 1998).

Loots *et al*. (2000) defined conserved noncoding sequences (CNSs) as conserved noncoding elements with greater or equal to 70% identity over at least 100 bp between human and mouse. Because of the fact we used eight clusters from seven species more evolutionarily divergent that only human and mouse, the following criteria have been used to define CNSs: identity over 60% in at least four out of eight clusters; presence in at least two species known to have only one *Hox*Aα cluster (horn shark, human, mouse, see Fig. 1) and minimum length of 50 base pairs (bp). In spite of this, when taking into account only the comparison between

human and mouse, our CNSs fulfill also the definition from Loots *et al*. (2000). CNSs have been tested in BLAST (http://www.ncbi.nlm.nih.gov/BLAST/) to confirm that are specific to *Hox* clusters.

Within such sequences, stretches between 95 and 100% identity and six nucleotides or more in length, conserved among at least six out of seven examined clusters, have received particular attention. The stretches over 95% identity within CNSs have been used to screen the transfac database (http://transfac.gbf.de/TRANSFAC/) in order to determine if they have been already described as transcription factors binding sites in similar or different biological context.

## ACKNOWLEDGEMENTS

This work has been supported by the grant of Deutsche Forschungsgemeineschaft (DFG) and by the Marie Curie foundation to Simona Santini. The authors wish to thank Dr. Edward Malaga-Trillo for library screening, many members of the DOE Joint Genome Institute (JGI) for DNA sequencing and Dr. Christian Klingenberg for reviewing the manuscript. Part of this work was performed under the auspices of the U.S. Department of Energy, Office of Biological and Environmental Research, by the University of California, Lawrence Berkeley National Laboratory under contract No. DE-AC03-76SF00098.

REFERENCES

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, and Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* **25:** 3389-3402.

Amores A, Force A, Yan Y-L, Joly L, Amemiya C, Fritz A, Ho RK, Langeland J, Prince V, Wang Y-L, Westerfield M, Ekker M, and Postlethwait J. 1998. Zebrafish *Hox* Clusters and Vertebrate Genome Evolution. *Science* **282:** 1711-1714.

Aparicio S, Hawker K, Cottage A, Mikawa Y, Zuo L, Venkatesh B, Chen E, Krumlauf R, and and Brenner S. 1997. Organization of the *Fugu rubripes Hox* clusters: evidence for continuing evolution of vertebrate *Hox* complexes. *Nature Genet.* **16:** 79-83.

Aparicio S, Morrison A, Gould A, Gilthorpe J, Chaudhuri C, Rigby P, Krumlauf R, and Brenner S. 1995. Detecting conserved regulatory elements with the model genome of the Japanese puffer fish, Fugu rubripes. *Proc. Natl. Acad. Sci. USA* **92:** 1684-1688.

Asif T. Chinwalla, Lisa L. Cook, Kimberly D. Delehaunty, Ginger A. Fewell, Lucinda A. Fulton, Robert S. Fulton, Tina A. Graves, LaDeana W. Hillier, Elaine R. Mardis, John D. McPherson, *et al*. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420:** 520-562.

Bergman CM and Kreitman M. 2001. Analysis of conserved noncoding DNA in Drosophila reveals similar constraints in intergenic and intronic sequences. *Genome Res.* **11:** 1335-1345.

Blanchette M, and Tompa M. 2002. Discovery of regulatory elements by a computational method for phylogenetic footprinting. *Genome Res.* **12:** 739-748.

Brenner S, Elgar G, Sandford R, Macrae A, Venkatesh B, and Aparicio S. 1993. Characterization of the pufferfish (Fugu) genome as a compact model vertebrate genome. *Nature* **366:** 265-268.

Brenner S, Venkatesh B, Yap WH, Chou CF, Tay A, Ponniah S, Wang Y, and Tan YH. 2002. Conserved regulation of the lymphocyte-specific expression of lck in the Fugu and mammals. *Proc. Natl. Acad. Sci. USA* **99:** 2936-2941.

Bucher P. 1990. Weight matrix description of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. *J. Mol. Biol.* **212**: 563-578.

Carroll, S., J. Grenier, and S. Weatherbee. 2001. From DNA to diversity-Molecular genetics and the evolution of animal design. *Blackwell Science, Malden Massachusetts***:** 175-186.

Catron KM, Iler N, and Abate C. 1993. Nucleotides flanking a conserved TAAT core dictate the DNA binding specificity of three murine homeodomain proteins. *Mol. Cell. Biol.* **13:** 2354-2365.

Chiu CH, Amemiya C, Dewar K, Kim CB, Ruddle FH, and Wagner GP. 2002. Molecular evolution of the *Hox*A cluster in the three major gnathostome lineages. *Proc. Natl. Acad. Sci. USA* **99:** 5492-5497.

Chu, D., N. Kakazu, M. Gorrin-Rivas, H. Lu, M. Kawata, T. Abe, K. Ueda, and Y. Adachi. 2001. Cloning and characterization of LUN, a novel ring finger protein that is highly expressed in lung and specifically binds to a palindromic sequence. *J. Biol. Chem.* **276:** 14004-14013.

Clark AG. 2001. The search for meaning in noncoding DNA. *Genome Res.* **11:** 1319-1320.

Cliften PF, Hillier LW, Fulton L, Graves T, Miner T, Gish WR, Waterston RH, and Johnston M. 2001. Surveying Saccharomyces genomes to identify functional elements by comparative DNA sequence analysis. *Genome Res.* **11:** 1175-1186.

Doerksen LF, Bhattacharya A, Kannan P, Pratt D, and Tainsky MA. 1996. Functional interaction between a RARE and an AP-2 binding site in the regulation of the human *HOX* A4 gene promoter. *Nucl. Acids Res.* **24:** 2849-2856.

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

Ekker SC, Young KE, von Kessler DP and Beachy PA. 1991. Optimal DNA sequence recognition by the Ultrabithorax homeodomain of Drosophila. *EMBO J*. **10**: 1179-1186

Ekker SC, Jackson DG, von Kessler DP, Sun BI, Young KE, and Beachy PA. 1994. The degree of variation in DNA sequence recognition among four Drosophila homeotic proteins. *EMBO J.* **13:** 3551-3560.

Epstein J, Cai J, Glaser T, Jepeal L and Maas R. 1994. Identification of a Pax paired domain recognition sequence and evidence for DNA-dependent conformational changes. *J Biol Chem*. **269**: 8355-61.

Ferrier DE, Minguillon C, Holland PW, and Garcia-Fernandez J. 2000. The amphioxus *Hox* cluster: deuterostome posterior flexibility and *Hox*14. *Evol Dev* **2:** 284-293.

Force A, Lynch M, Pickett FB, Amores A, Yan YL, and Postlethwait J. 1999. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151:** 1531-1545.

Frasch M, Chen X, and and Lufkin T. 1995. Evolutionary-conserved enhancers direct region-specific expression of the murine *Hox*a-1 and *Hox*a-2 loci in both mice and Drosophila. *Development* **121:** 957-974.

Garcia-Fernandez J and and Holland PW. 1994. Archetipal organization of the amphioxus *Hox* gene cluster. *Nature* **370:** 563-566.

Gehring WJ. 1993. Exploring the homeobox. *Gene* **135:** 215-221.

Grange T, Roux J., Rigaud G and Pictet R. 1991. Cell-type specific activity of two glucocorticoid responsive units of rat tyrosine aminotransferase gene is associated with multiple binding sites for C/EBP and a novel liver-specific nuclear factor. *Nucl. Acids Res*. **19**: 131-139.

Haerry TE and Gehring WJ. 1997. A conserved cluster of homeodomain binding sites in the mouse *Hox*a-4 intron functions in Drosophila embryos as an enhancer that is directly regulated by Ultrabithorax. *Dev. Biol.* **186:** 1-15.

Haerry TE, G. WJ, and . Intron of the mouse *Hox*a-7 gene contains conserved homeodomain binding sites that can function as an enhancer element in Drosophila. *Proc. Natl. Acad. Sci. USA* **93:** 13884-13889.

Hardison RC. 2000. Conserved noncoding sequences are reliable guides to regulatory element. *Trends Genet.* **16:** 369-372.

Hardison RC, Oeltjen J, and Miller W. 1997. Long human-mouse sequence alignments reveal novel regulatory elements: a reason to sequence the mouse genome. *Genome Res.* **7:** 959-966.

Hauptmann G, Belting HG, Wolke U, Lunde K, Soll I, Abdelilah-Seyfried S, Prince V, Driever W. 2002. spiel ohne grenzen/pou2 is required for zebrafish hindbrain segmentation. *Development* **129**:1645-55

Hinegardner R. 1976. The cellular DNA content of sharks, rays and some other fishes. *Comp. Biochem. Physiol. B* **55:** 367-370.

Holland PW. 1997. Vertebrate evolution: something fishy about *Hox* genes. *Curr. Biol.* **7:** R570-572.

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

Knittel T, Kessel M, Kim MH, and Gruss P. 1995. A conserved enhancer of the human and murine *Hox*A-7 gene specifies the anterior boundary of expression during embryonal development. *Development* **121:** 1077-1088.

Krumlauf R. 1994. *Hox* genes in vertebrate development. *Cell* **78:** 191-201.

Langston AW, Thompson JR, and Gudas LJ. 1997. Retinoic acid-responsive enhancers located 3' of the *Hox* A and *Hox* B homeobox gene clusters. Functional analysis. *J Biol. Chem.* **272:** 2167-2175.

Lufkin T. 1997. Transcriptional regulation of vertebrate *Hox* genes during embryogenesis. *Crit. Rev. Eukaryot. Gene Expr.* **7:** 195-213.

Maconochie M, Nonchev S, Morrison A, and Krumlauf R. 1996. Paralogous *Hox* genes: function and regulation. *Annu. Rev. Genet.* **30:** 529-556.

Malaga-Trillo E and and Meyer A. 2001. Genome Duplication and Accelerated Evolution of *Hox* Genes and Cluster Architecture in Teleosts Fishes. *Amer. Zool.* **41:** 676-686.

Manzanares M, Wada H, Itasaki N, Trainor PA, Krumlauf R, and Holland PWH. 2000. Conservation and elaboration of *Hox* gene regulation during evolution of the vertebrate head. *Nature* **408:** 854-857.

Margalit Y, Yarus S, Shapira E, Gruenbaum Y, and Fainsod A. 1993. Isolation and characterization of target sequences of the chicken CdxA homeobox gene. *Nucl. Acids Res.* **21:** 4915-4922.

Meyer A. 1998. *Hox* gene variation and evolution. *Nature* **391:** 225-228.

Meyer A and Malaga-Trillo E. 1999. More fishy tales about *Hox* genes. *Curr. Biol.* **9:** R210-R213.

Meyer A and Schartl M. 1999. Gene and genome duplications in vertebrates: the one-to-four (-to-eight in fish) rule and the evolution of novel gene functions. *Curr. Opin. Cell. Biol.* **11:** 699-704.

Moran JV, DeBerardinis RJ, and Kazazian HH Jr. 1999. Exon shuffling by L1 retrotransposition. *Science* **283:** 1530-1534.

Morrison A, Chaudhuri C, Ariza-McNaughton L, Muchamore I, Kuroiwa A, Krumlauf R, ., and Sep;. 1995. Comparative analysis of chicken *Hox*b-4 regulation in transgenic mice. *Mech. Dev.* **53:** 47-59.

Nonchev S, Vesque C, Maconochie M, Seitanidou T, Ariza-McNaughton L, Frain M, Marshall H, Sham MH, Krumlauf R, and Charnay P. 1996. Segmental expression of *Hox*a-2 in the hindbrain is directly regulated by Krox-20. *Development* **122:** 543-554.

Odenwald WF, Garbern J, Arnheiter H, Tournier-Lasserve E, and Lazzarini RA. 1989. The *Hox*-1.3 homeo box protein is a sequence-specific DNA-binding phosphoprotein. *Genes Devel.* **3:** 158-172.

Ohno S and Atkin NB. 1966. Comparative DNA values and chromosome complements of eight species of fishes. *Chromosoma* **18:** 455-466.

Onyango P, Miller W, Lehoczky J, Leung CT, Birren B, Wheelan S, D. K, and Feinberg AP. 2000. Sequence and comparative analysis of the mouse 1-Megabase region orthologus to the human 11p15 imprinted domain. *Genome Res.* **10:** 1697-1710.

Peifer M, Karch F and Bender W. 1987. The bithorax complex: control of segmental identity.*Genes Devel.* **1**: 891-898.

Pough FH, Janis CM, and Heiser JB. 1999. Vertebrate life. *Prentice Hall Eds.***:** 192 and 267.

Prince VE, Joly L, Ekker M, and Ho RK. 1998. Zebrafish *Hox* genes: genomic organization and modified colinear expression patterns in the trunk. *Development* **125:** 407-420.

Robinson-Rechavi M, Marchand O, Escriva H, Bardet PL, Zelus D, Hughes S, and L. V. 2001. Euteleost fish genomes are characterized by expansion of gene families. *Genome Res.* **11:** 781-788.

Rossi P, Karsenty G, Roberts AB, Roche NS, Sporn MB, and de Crombrugghe B. 1988. A nuclear factor 1 binding site mediates the transcriptional activation of a type I collagen promoter by transforming growth factor-beta. *Cell* **52:** 405-414.

Roth FP, Hughes JD, Estep PW, and Church GM. 1998. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nature Biotech.* **16:** 939-945.

Schwartz S, Zhang Z, Frazer KA, Smit A, Riemer C, Bouck J, Gibbs R, Hardison R, and Miller W. 2000. PipMaker--a web server for aligning two genomic DNA sequences. *Genome Res.* **10:** 577-586.

Shashikant, C., C. Bieberich, H. Belting, J. Wang, M. Borbely, and F. Ruddle. 1995. Regulation of *Hox*c-8 during mouse embryonic development: identification and characterization of critical elements involved in early neural tube expression. *Development* **121:** 4339-4347.

Snell EA, Scemama J-L, and and Stellwag EJ. 1999. Genomic Organization of the *Hoxa4-Hoxa10* Region From Morone saxatilis: Implications for *Hox* Gene Evolution Among Vertebrates. *J. Exper. Zool. (Mol. Dev. Evol.)* **285:** 41-49.

Stingo V, Rocco L, and Improta R. 1989. Chromosome markers and karyology of selachians. *J. Exp. Zool. Suppl.* **2:** 175-185.

Tagle DA, Koop BF, Goodman M, Slightom JL, Hess DL, and Jones RT. 1988. Embryonic epsilon and gamma globin genes of a prosimian primate (Galago crassicaudatus). Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. *J. Mol. Biol.* **203:** 439-455.

Tan D-P, Ferrante J, Nazarali A, Shao X, Kozak CA, G. V, and and NirenbergM. 1992. Murine *Hox*-1.11 homeobox gene structure and expression. *Proc. Natl. Acad. Sci. USA* **89:** 6280-6284.

Taylor JS, Van de Peer Y, Braasch I, and Meyer A. 2001. Comparative genomics provides evidence for an ancient genome duplication event in fish. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **356:** 1661-1679.

Tiersch TR, Chandler RW, Wachtel SS, and Elias S. 1989. Reference standards for flow cytometry and application in comparative studies of nuclear DNA content. *Cytometry* **10:** 706-710.

Tomilin NV. 1999. Control of genes by mammalian retroposons. *Int. Rev. Cytol.* **186:** 1-48.

Tompa M. 2001. Identifying functional elements by comparative DNA sequence analysis. *Genome Res.* **11:** 1143-1144.

Venkatesh B, Gilligan P, and Brenner S. 2000. Fugu: a compact vertebrate reference genome. *FEBS Lett.* **476:** 3-7.

Verrijzer CP, Alkema MJ, van Weperen WW, Van Leeuwen HC, Strating MJ, van der Vliet PC. 1992. The DNA binding specificity of the bipartite POU domain and its subdomains. *EMBO J.* **11**:4993-5003

Vesque C, Maconochie M, Nonchev S, Ariza-McNaughton L, Kuroiwa A, Charnay P, and Krumlauf R. 1996. *Hox*b-2 transcriptional activation in rhombomeres 3 and 5 requires an evolutionarily conserved cis-acting element in addition to the Krox-20 binding site. *EMBO J.* **15:** 5383-5396.

Vinogradov AE. 1998. Genome size and GC-percent in vertebrates as determined by flow cytometry: the triangular relationship. *Cytometry* **31:** 100-109.

Wasserman WW, Palumbo M, Thompson W, Fickett JW, and Lawrence CE. 2000. Human-mouse genome comparisons to locate regulatory sites. *Nature Genet.* **26:** 225-228.

Weiss KM and Fullerton SM. 2000. Phenogenetic drift and the evolution of genotype-phenotype relationships. *Theor. Popul. Biol.* **57**: 187-195.

Woods DB, Ghysdael J and Owen MJ. 1992. Identification of nucleotide preferences in DNA sequences recognized specifically by c-ETS-1 protein. *Nucl. Acids Res*. **20**: 699-704

Yanagisawa S and Schmidt RJ. 1999. Diversity and similarity among recognition sequence of Dof transcription factors. *Plant J*. **17**: 209-214.

Zhang Z, Berman P, and -. Miller W. 1998. Alignments without low-scoring regions. *J Comput. Biol.* **5:** 197-210.

WEB SITES REFERENCES

http://www.jgi.doe.gov/programs/fugu/fugu_mainpage.html, JGI fugu project homepage

http://ftp.genome.washington.edu/cgi-bin/RepeatMasker/, RepeatMasker homepage

http://bio.cse.psu.edu/pipmaker/, PipMaker homepage

http://transfac.gbf.de/TRANSFAC/, Transfac database homepage

http://www.ncbi.nlm.nih.gov/BLAST/, BLAST homepage

Figure Legends

Figure 1: Evolutionary relationships among the species included in this work. The divergence date between the lineage leading to Chondrichthyes (to which *Heterodontus*, the horn shark belongs) and that leading to the clade of all other taxa on this tree is about 500 millions years. *Actinopterygii* (the ray-finned fishes) and *Sarcopterygii* (the tetrapods) diverged about 450 million years ago. Teleosts radiated more than 200 million years ago. The divergence between human and mouse is dated to about 80 millions years (Pough *et al.*, 1999). Horn shark, mouse and human have a single *Hox*A cluster, while all fishes examined so far have two (see text for details). Among fishes, independent gene losses took place in zebrafish and pufferfish relative to tilapia. Solid boxes represent individual genes. Duplicated clusters are designated as □ or □. Pseudogene A2□ and A10□ in zebrafish are marked with a cross. Question marks represent so genomic regions that are not yet characterized.

Figure 2: Relative sizes of *Hox*A clusters. The reduction in size of *Hox*A cluster seems to be independent from the pattern of gene loss. Solid boxes represent individual genes. The duplicated □ and □ clusters are differentiated only for zebrafish. The alignable portion of the pseudogenes HoxA7□ of pufferfish, HoxA2□ and HoxA10□ of zebrafish are shown too. For total length of clusters, refer to the text.

Figure 3: Relationship between genome size and length of the portion HoxA4 to HoxA10 of *Hox*A clusters. The length of *Hox*A clusters is significantly correlated ($P = 0.06$) with the genome size expressed as C value. The *Hox*A□ cluster lengths are shown. To be able to include also striped bass (*Hox*A cluster available only from gene 4 to 10) in the analysis, only the length of the *Hox*A4 to *Hox*A10 portion of the cluster is shown. Both zebrafish *Hox*A clusters are shown.

Figure 4: Pip output of the comparison of tilapia *Hox*A□, striped bass *Hox*A□, pufferfish *Hox*A□, zebrafish *Hox*A□ and A□, horn shark *Hox*A, human *Hox*A and mouse *Hox*A clusters. The tilapia sequence has been used as reference sequence. Kilobase (kb) markings are based on the tilapia sequence. Blue background indicates coding region, yellow indicates intron, red indicates conserved noncoding sequence (CNS) previously described in literature and thr green background indicates heretoforth undescribed CNSs. Horizontal arrows indicate the direction of transcription, tall black boxes show exons, short open boxes indicate a CpG/GpC ratio between 0.6 and 0.75 and short grey boxes indicate a CpG/GpC ratio over 0.75. Interspersed repeat elements are shown as triangles (e.g., in position 91 kb).

Figure 5: Lengths of CNSs in the different intergenic regions. The intergenic regions located 3' in the cluster are better conserved than those between genes located 5' in the cluster. The graph shows the number of conserved bases (>60% identity among at least four of eight clusters, present in at least two species of those known to have only one *Hox*A cluster and minimum length of 50 bp). There is a significant relationship between the degree of conservation and the position in the cluster ($P$ = 0.007).

Figure 6: Alignment of RARE elements described as "box c" and "box d" (Odenwald *et al.*, 1989) immediately upstream of the *Hox*A5 genes.

Figure 7: Alignment of known regulatory elements. (A) Sequence of *Krx20* binding sites in different species. Krox20 binding sites are involved in *Hox*2 regulation and they are conserved in *Hox*A and B clusters from human, mouse, fugu and *Hox*A from tilapia. Both *Krx20* and the Box a are widely

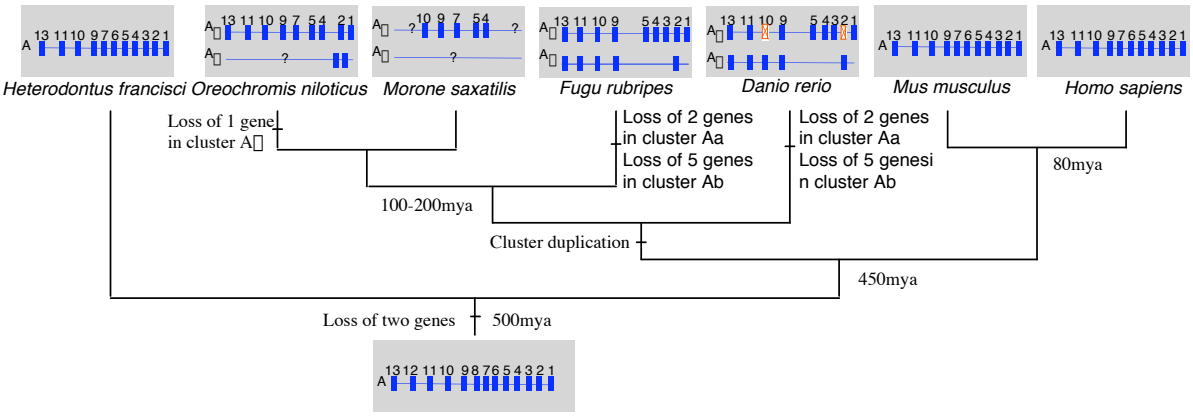
conserved. The degree of identity is 67% among the different species in which they have been found. (B) Alignment of sequences of "box a" motif in different species.

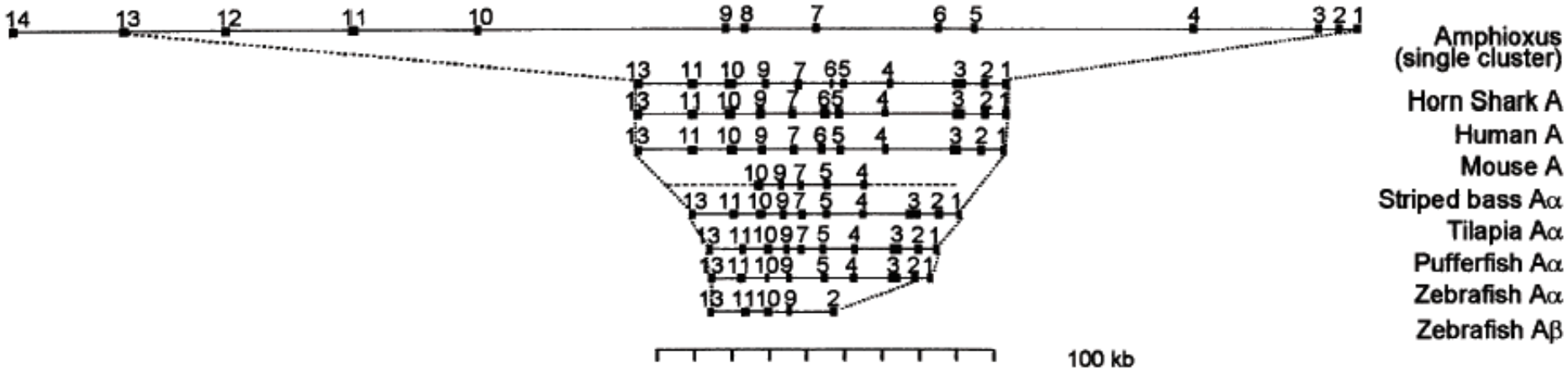
Table 1: Percentual expression of the base composition of the *Hox*A clusters. A bias towards AT richness is present in all examined clusters *Hox*A cluster. In mouse the AT-richness is not significative (chi squared test, data not shown).

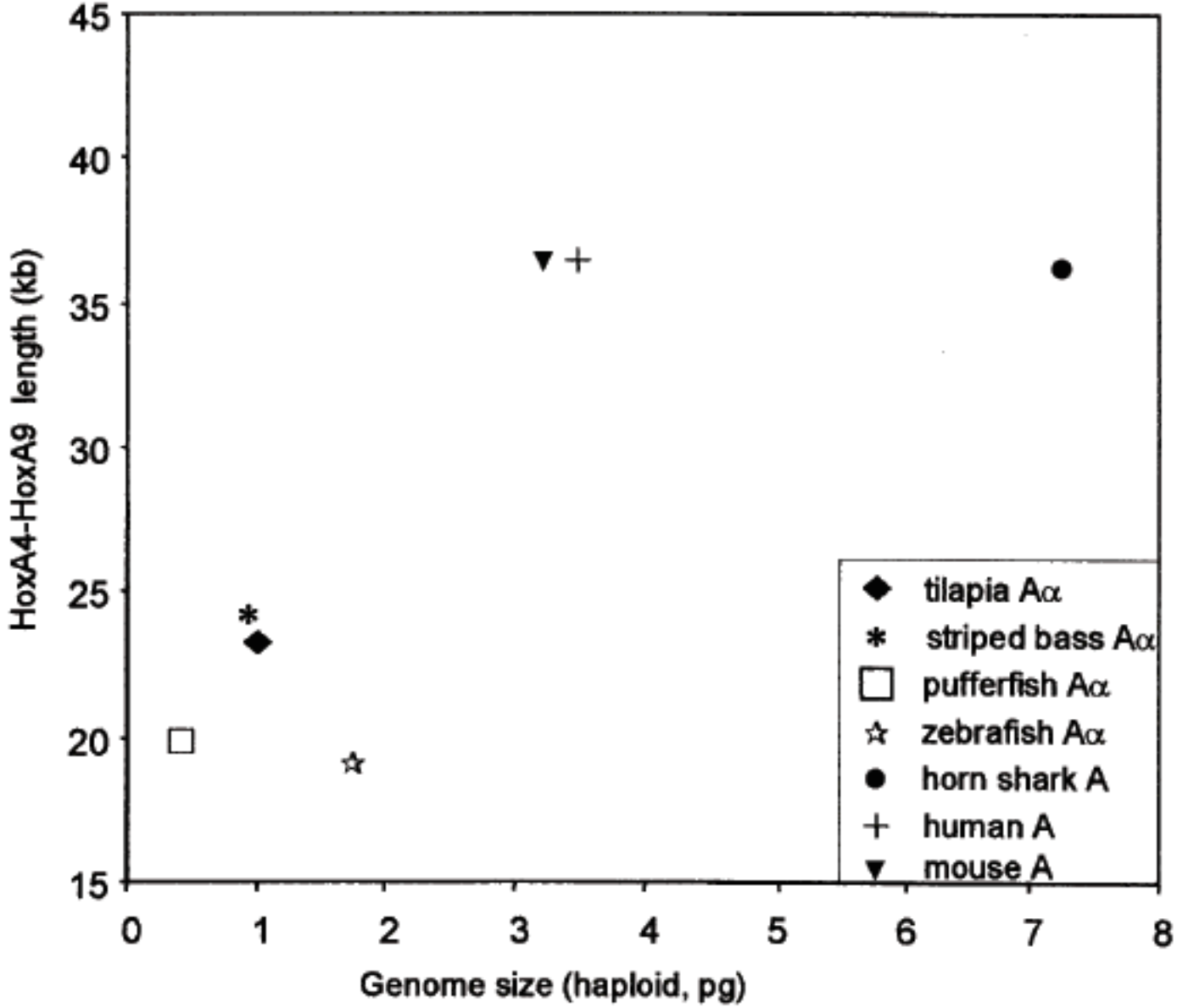
Table 2: CNSs identified through the comparative approach. Column 1: position of CNS in the *Hox*A cluster in tilapia. Column 2: length in bp of the CNS. Column 3-9: percentage of identity of the corresponding region in other genomes. Column 10: number x length of sequence over 95% identity among all species. Column 11: reference for previously described CNSs and for binding sites that show similar sequence.

Table 3: Percentual expression of base conservation per intergenic region of tilapia *Hox*A cluster. Column 1: considered intergenic fragment. Column 2: percentage of total noncoding bases of the tilapia *Hox*A cluster represented by the intergenic region. Column 3: percentage of the intergenic fragment identified as CNS by our analyses. Column 4: percentage of the intergenic fragment previously described in literature as involved in *Hox* genes regulation. Column 5: percentage of total CNSs present in the intergenic fragment.

Figure 1


Heterodontus francisci
Oreochromis niloticus
Morone saxatilis
Fugu rubripes
Danio rerio
Mus musculus
Homo sapiens
Loss of 1 gene in cluster A
Loss of 2 genes in cluster Aa
Loss of 5 genes in cluster Ab
Loss of 2 genes in cluster Aa
Loss of 5 genesi n cluster Ab
80mya
100-200mya
Cluster duplication
450mya
Loss of two genes
500mya


14 13 12 11 10 9 8 7 6 5 4 3 2 1
Amphioxus (single cluster)
Horn Shark A
Human A
Mouse A
Striped bass Aα
Tilapia Aα
Pufferfish Aα
Zebrafish Aα
Zebrafish Aβ
100 kb


45
40
35
30
25
20
15
HoxA4-HoxA9 length (kb)
0
1
2
3
4
5
6
7
8
Genome size (haploid, pg)
tilapia Aα
striped bass Aα
pufferfish Aα
zebrafish Aα
horn shark A
human A
mouse A


HoxA13
Tilapia
Striped bass
Pufferfish
Zebrafish_a
Zebrafish_b
Horn Shark
Mouse
Human
100%
50%
0kb
10kb
20kb
HoxA11
HoxA10
HoxA9
Tilapia
Striped bass
Pufferfish
Zebrafish_a
Zebrafish_b
Horn Shark
Mouse
Human
100%
50%
20kb
30kb
40kb
HoxA7
HoxA5
HoxA4
Tilapia
Striped bass
Pufferfish
Zebrafish_a
Zebrafish_b
Horn Shark
Mouse
Human
100%
50%
40kb
50kb
60kb
HoxA3
HoxA2
Tilapia
Striped bass
Pufferfish
Zebrafish_a
Zebrafish_b
Horn Shark
Mouse
Human
100%
50%
60kb
70kb
80kb
HoxA1
Tilapia
Striped bass
Pufferfish
Zebrafish_a
Zebrafish_b
Horn Shark
Mouse
Human
100%
50%
90kb
100kb

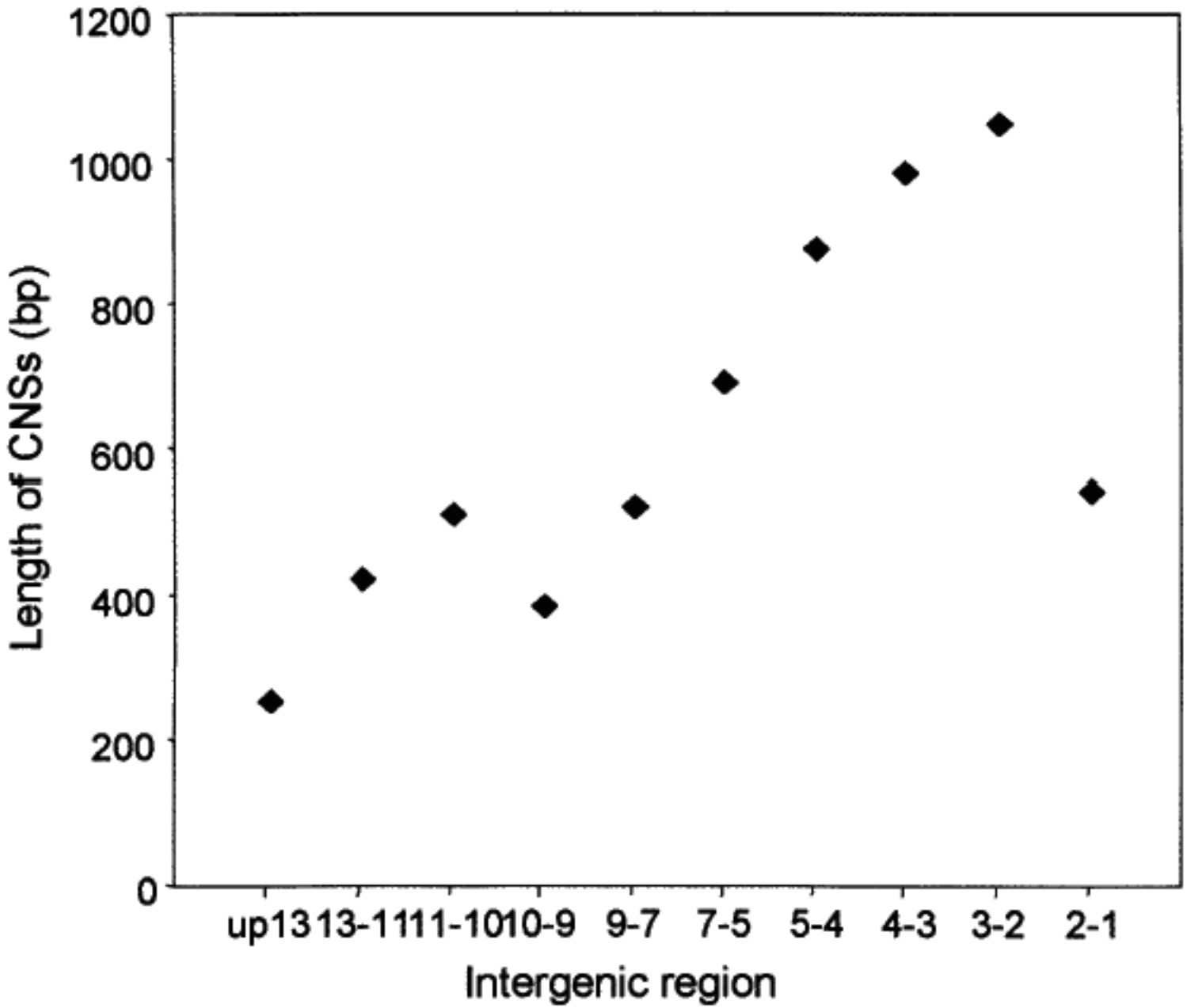

1200
1000
800
600
400
200
0
Length of CNSs (bp)
up13
13-11
11-10
10-9
9-7
7-5
5-4
4-3
3-2
2-1
Intergenic region

Figure 6

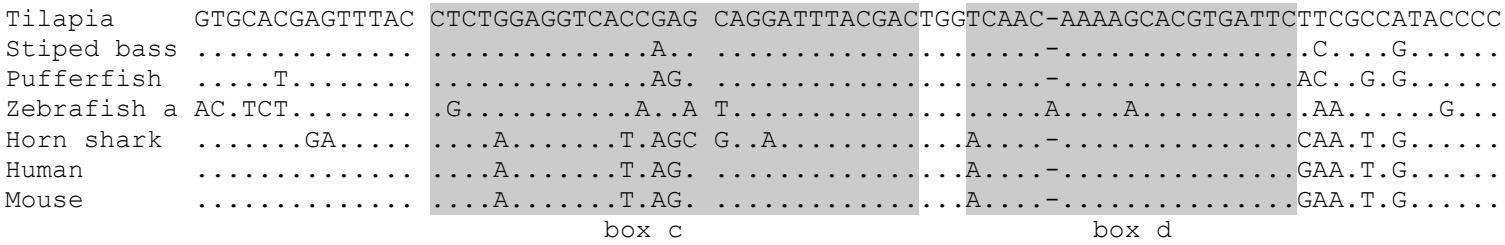

Tilapia GTGCACGAGTTTAC CTCTGGAGGTCACCGAG CAGGATTTACGACTGGTCAAC-AAAAGCACGTGATTCTTCGCCATACCCC
Stiped bass ..... ..... A.. ..... -..... C.... G.....
Pufferfish .....T..... ..... AG. ..... -..... AC..G.G.....
Zebrafish a AC.TCT..... .G..... A..A T..... A....A..... AA.....G...
Horn shark .....GA..... ....A.....T.AGC G..A..... A....-..... CAA.T.G.....
Human ..... ....A.....T.AG. ..... A....-..... GAA.T.G.....
Mouse ..... ....A.....T.AG. ..... A....-..... GAA.T.G.....
box c
box d

Figure 7

```
human A2          CACCCACGC
mouse A2          CACCCACGC
tilapia A2        CACCCACTC
pufferfish A2     CACCCACTC
mouse B2          CACCCACGC
pufferfish B2     CGCCCACAC
human B2          CCACCACAC
chick B2          CACCCACAC

consensus         *--*****-*


human BoxA        CTGACAAAGCCT
mouse BoxA        CTGACAAAGCCC
tilapia BoxA      CACACAAAGCCT
pufferfish BoxA   GACACAAAGCCT

consensus         ---*********-
```

Table 1

| Species | %A | %C | %G | %T |
|---|---|---|---|---|
| Tilapia | 28.356 | 21.166 | 20.981 | 29.496 |
| Pufferfish | 28.476 | 21.398 | 21.093 | 29.033 |
| Zebrafish a | 31.231 | 18.816 | 18.378 | 31.574 |
| Zebrafish b | 32.891 | 18.552 | 16.876 | 31.680 |
| Horn shark | 31.169 | 18.783 | 18.666 | 31.382 |
| Human | 31.169 | 18.783 | 18.666 | 31.382 |
| Mouse | 24.827 | 24.778 | 25.271 | 25.124 |

Table 2

| Position | Length (bp) | Striped bass | Pufferfish | Zebrafish a | Zebrafish b | Horn Shark | Human | Mouse | Over 95% | Literature/Similar binding sites |
|---|---|---|---|---|---|---|---|---|---|---|
| 1kb upstream 13 | 63 | | 86 | 73 | 82 | 78 | 79 | 80 | 1x19nt | NF-1 (Rossi et al., 1988) |
| imm. upstream13 | 188 | | 83 | 65 | 63 | 66 | 75 | 71 | 0 | |
| 13-11 | 192 | | 89 | 26 | | 60 | 68 | 67 | 2x10nt | |
| imm. upstream11 | 230 | | 89 | 66 | 68 | 63 | 70 | 71 | 5x7-28nt | |
| 11-10 | 121 | | 96 | 84 | 85 | | 64 | 63 | 2x6-8 | |
| imm. upstream10 | 391 | 92 | 86 | | 66 | 68 | 70 | 68 | 2x8-25 | |
| 10-9 a | 96 | 95 | 86 | 63 | | 66 | 61 | 65 | 2x6-7 | Abd B (Ekker et al., 1994); RNA pol. II cap signal (Bucher, 1990) |
| 10-9 b | 95 | 98 | 98 | 89 | 81 | 79 | 73 | 72 | 1x24 | Murine homeotic proteins b.s. (Catron et al., 1993) |
| 10-9 c | 21 | 99 | 100 | 81 | 81 | 81 | 95 | 95 | 1x14 | |
| Imm upstream 9 | 191 | 94 | 87 | 63 | 56 | 69 | 61 | 63 | 2x5-6 | target sequences chicken CdxA (Margalit et al., 1993) |
| 9-7 a | 62 | 100 | 98 | 92 | 72 | 78 | 79 | 79 | 2x11-15 | Abd B (Ekker et al., 1994) |
| 9-7 b | 276 | 96 | 93 | 71 | | 71 | 70 | 69 | 3x6-11 | c-ETS-1 protein b.s. (Woods et al., 1992) |
| imm. upstream7 | 185 | 95 | 88 | | | 79 | 78 | 78 | 3x9-14 | enhancer regulatory element, *H. sapiens* (Knittel *et al*., 1995) |
| 7-5 | 163 | 81 | 78 | 77 | | 78 | 81 | 81 | 2x8-11 | H8/7-6 FCS (Kim et al., 2000) |
| imm. upstream5 | 529 | 93 | 84 | 69 | | 38 | 76 | 76 | 8x6-39 | RARE, *H. sapiens, M. musculus* (Odenwald *et al*., 1989) |
| 5.4 a | 280 | 99 | 94 | 77 | | 82 | 83 | 83 | 7x9-33 | Pax b.s., (Epstein et al., 1994); Ultrabithorax b.s. (Ekker et al., 1991); target sequences of chicken CdxA homeobox gene (Margalit.et el., 1993) |
| 5-4 b | 63 | 97 | 98 | 83 | | 83 | 81 | 83 | 2x9-19 | Dof b.s. (Yanagysawa &Schmidt, 1999) |
| 5-4 c | 209 | 95 | 93 | 67 | | 71 | 69 | 69 | 5x8-15 | NF of C-EBP family (Grange et al., 1991) |
| 5-4 d | 239 | 92 | 89 | 81 | | 84 | 79 | 78 | 7x7-24 | RARE (HoxA4 promoter, *H. sapiens*, Doerksen *et al*., 1996) |
| imm. upstream4 | 83 | 100 | 100 | 89 | | 91 | 78 | 76 | 3x6-30 | RARE (HoxA4 promoter, *H. sapiens*, Doerksen *et al*., 1996) |
| 4-3 a | 78 | | 91 | 69 | | 76 | 72 | 66 | 2x6-7 | Dof b.s. (Yanagysawa &Schmidt, 1999) |
| 4-3 b | 480 | | 87 | 67 | | 71 | 66 | 63 | 5x6-10 | No |
| 4-3 c | 51 | | 93 | 72 | | 80 | 75 | 73 | 2x6-10 | No |
| 4-3 d | 136 | | 96 | 76 | | 76 | 66 | 65 | 4x8-12 | No |
| imm. upstream3 | 235 | | 86 | 90 | | 82 | 73 | 79 | 6x7-13 | No |
| 3-2 a | 476 | | 79 | | | 61 | 66 | 60 | 0 | No |
| 3-2 b | 189 | | 93 | 72 | 81 | 68 | 69 | 67 | 5x6-9 | No |
| imm. upstream2 | 382 | | 89 | 64 | 78 | 77 | 77 | 77 | 8x8-43 | HoxA2 promoter, M. musculus (Tan *et al*., 1992) |
| 2-1 | 190 | | 93 | | | 61 | 72 | 72 | 3x10-11 | No |
| imm. upstream1 | 352 | | 78 | 60 | | 59 | 64 | 61 | 2x6-8 | No |

Table 3

| Intergenic fragment | % of total non-coding bases | % identified as CNS | % described in literature | % of total CNSs |
|---|---|---|---|---|
| Evx-13 | 13 | 3 | 0 | 4 |
| 13-11 | 15 | 4 | 0 | 7 |
| 11-10 | 9 | 9 | 0 | 8 |
| 10-9 | 7 | 9 | 0 | 6 |
| 9-7 | 6 | 13 | 12 | 8 |
| 7-5 | 8 | 14 | 4 | 11 |
| 5-4 | 13 | 10 | 4 | 14 |
| 4-3 | 17 | 9 | 0 | 16 |
| 3-2 | 7 | 23 | 10 | 17 |
| 2-1 | 6 | 14 | 14 | 9 |