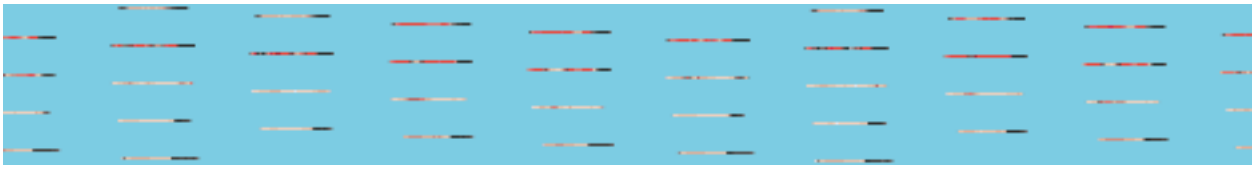
Evolutionary Conservation of Regulatory Elements in Vertebrate *Hox* Gene Clusters

Simona Santini, Jeffrey L. Boore and Axel Meyer

Genome Res. 2003 13: 1111-1122
Access the most recent version at doi:10.1101/gr.700503



Letter

Evolutionary Conservation of Regulatory Elements in Vertebrate *Hox* Gene Clusters

Simona Santini,[1] Jeffrey L. Boore,[2] and Axel Meyer[1,3]

[1]*Department of Biology, University of Konstanz, 78457 Konstanz, Germany;* [2]*Department of Evolutionary Genomics, DOE Joint Genome Institute, Lawrence Berkeley National Laboratory, and University of California, Berkeley, California 94720, USA*

Comparisons of DNA sequences among evolutionarily distantly related genomes permit identification of conserved functional regions in noncoding DNA. *Hox* genes are highly conserved in vertebrates, occur in clusters, and are uninterrupted by other genes. We aligned (PipMaker) the nucleotide sequences of the *Hox*A clusters of tilapia, pufferfish, striped bass, zebrafish, horn shark, human, and mouse, which are separated by approximately 500 million years of evolution. In support of our approach, several identified putative regulatory elements known to regulate the expression of *Hox* genes were recovered. The majority of the newly identified putative regulatory elements contain short fragments that are almost completely conserved and are identical to known binding sites for regulatory proteins (Transfac database). The regulatory intergenic regions located between the genes that are expressed most anteriorly in the embryo are longer and apparently more evolutionarily conserved than those at the other end of *Hox* clusters. Different presumed regulatory sequences are retained in either the Aα or Aβ duplicated *Hox* clusters in the fish lineages. This suggests that the conserved elements are involved in different gene regulatory networks and supports the duplication-deletion-complementation model of functional divergence of duplicated genes.

[Supplemental material is available online at www.genome.org. The sequence data from this study have been submitted to GenBank under accession no. AF538976.]

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 stabilizing 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 heretofore 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. Functional elements tend to evolve at slower rates than nonfunctional regions, because they are subject to selection (Tagle et al. 1988; Hardison et al. 1997; Hardison 2000; Cliften et al. 2001). Due to this slower rate of evolution, comparisons among evolutionarily distantly related genome sequences provide a tool to identify functional regions in the sea of noncoding DNA (Tompa 2001, Blanchette and Tompa 2002), an approach that 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 and Kreitman 2001) have been successfully used to identify regulatory regions, and comparisons between humans and mice (evolutionary distance of approximately 80 million years; Pough et al. 1999) revealed many of the functionally relevant binding sites (Onyango et al. 2000). This is because of their high degree of conservation (on average 93.2%; Wassermann et al. 2000).

Comparisons among closely related species revealed that many nonfunctional noncoding sequences also show a high degree of nucleotide identity, rendering the identification of DNA regions involved in gene regulation more difficult. However, in the alignment of long stretches of DNA sequences from evolutionarily distantly related species, conserved putative regulatory elements will stand out from the background of highly variable nonfunctional regions. This beneficial signal-to-noise ratio among more distantly related species permits the identification of putative regulatory elements.

The search for regulatory elements through comparative genomic approaches in *Hox* gene clusters promises to be particularly successful because their nucleotide sequence and function are extremely conserved in all vertebrates in which they have been studied. *Hox* genes code for transcription factors that are responsible for establishing the animal body plan early in embryonic development. They specify the 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 family, *Hox* genes belong to a subfamily whose members are arranged in genomic clusters. Interestingly, their expression in terms of time of activation and boundary of expression along the anterior–posterior axis is "colinear" with the their chro-

[3]Corresponding author.
E-MAIL axel.meyer@uni-konstanz.de; FAX: 49 7531 883018.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.700503.

mosomal arrangement (e.g., Krumlauf 1994). *Hox* genes occur in strictly packed clusters, which aids their identification and alignment. It may not be surprising that *Hox* genes are highly conserved during evolution because of their importance in development in all animal phyla. Moreover, the degree of conservation in their coding sequences might suggest that their regulatory elements are largely invariant across even great evolutionary distances. There is some evidence for this expectation. One of the selective forces that keeps the genes of *Hox* clusters uninterrupted by reshuffling and insertion of other genes may stem from the fact that adjacent genes share common *cis*-regulatory elements (Peifer et al. 1987). Therefore, adjacent genes must remain closely linked, because translocations or insertions between them would deprive one of them of its *cis*-regulatory elements and, hence, be lethal mutations.

RESULTS

We compared four teleost species: tilapia (*Oreochromis niloticus*), pufferfish (*Fugu rubripes*), striped bass (*Morone saxatilis*), and zebrafish (*Danio rerio*) with two mammalian species (*Homo sapiens* and *Mus musculus*) and the horn shark (*Heterodontus francisci*) as an outgroup species. The *Hox* gene contents for all these species are compared 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 to discover putative regulatory elements.

Genomic Architecture of *Hox* A Clusters

Comparisons of gene lengths and distances between genes of the *Hox*A clusters are shown in Figure 2. The single *Hox* cluster region of the cephalochordate amphioxus (haploid DNA content: C = 0.59 pg; Ohno and Atkin 1966) spans over 400 kb (Garcia-Fernandez and Holland 1994; Ferrier et al. 2000), but the *Hox*A clusters of vertebrates that have been studied are considerably smaller. In the shark (C = 7.25 pg, Stingo et al. 1989), the *Hox*A region is only approximately 110 kb long (AF224262 and AF479755). In this species, the cluster was previously named *Hox*M, but is the ortholog of *Hox*A (Kim et al. 2000). In humans (C = 3.50 pg; Tiersch et al. 1989), the *Hox* A cluster is 110 kb long (AC004079, AC004080, and AC010990), in the mouse (C = 3.25 pg; Vinogradov 1998; Asif et al. 2002) it is 105 kb (AC021667), in the tilapia (C = 0.99 pg; Hinegardner 1976) the *Hox*Aα cluster is 100 kb (AF533976), in the pufferfish (C = 0.40 pg; Brenner et al. 1993) the *Hox*Aα is 64 kb (JGI public database), in the zebrafish (C = 1.75 pg; Vinogradov 1998) the *Hox*Aα is 62 kb (AC107365) and the *Hox*Aβ is 33 kb (AC107364). The *Hox*A cluster of the mouse shows an even base composition, whereas for all other genomes examined the base composition of the *Hox*A clusters is AT-biased (Table 1).

The available striped bass (C = 0.89 pg, Hinegardner 1976) sequence does not cover the entire cluster, but only the region from *Hox*A10α to *Hox*A4α. The region *Hox*A9α to *Hox*A4α in striped bass is 24 kb long (AF089743); the homologous region in the tilapia *Hox*Aa cluster is 23 kb, in the pufferfish *Hox*Aa cluster it is approximately 20 kb, and in the zebrafish *Hox*Aα cluster it is approximately 19 kb (the zebrafish *Hox*Aβ does not contain genes 4, 5, and 7, so therefore cannot be evaluated). In the shark, human, and mouse clusters the region *Hox*A9 to *Hox*A4 is approximately 36 kb. In agreement with the view that *Hox* clusters are reduced in size in 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 previously shown 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; this holds true both in regard to relative lengths not only of orthologous genes among species, but also of spacing between genes, that is, the length of intergenic regions (Fig. 2).

There is increasing evidence for a fish-specific genome duplication that was shared by all (or most) ray-finned fishes (e.g., Amores et al. 1998; Wittbrodt et al. 1998; Taylor et al. 2001). This genome duplication also caused an initial doubling (and some secondary lineage-specific losses) of the number of *Hox* clusters from four to eight. So that, for example, two copies of the initial *Hox*A cluster resulted in the *Hox*Aα and the *Hox*Aβ clusters, which are now expected to be found in all (or most) ray-finned fishes. Independent gene losses in *Hox* clusters have happened in different species of fishes (Fig. 2).

The pufferfish *Hox*Aα cluster was initially thought to lack *Hox*A7α (Aparicio et al. 1997), and it was hypothesized that this loss, together with the loss of other members of the entire paralogy group 7 genes (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; Meyer 1998; Prince et al. 1998; Meyer and Malaga-Trillo 1999). Our comparisons show conservation of *Hox*A7α 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*A7α gene is a pseudogene.

The zebrafish Aα cluster lacks *Hox*A7α and contains only a fragment of exon 2 of *Hox*A10α. It also lacks *Hox*A2α (Amores et al. 1998), but the cluster region corresponding to both *Hox*A2α exons, 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*A1β and *Hox*A3β, *Hox*A4β, *Hox*A5β, and *Hox*A7β genes. In zebrafish, the *Hox*Aβ cluster has been subject to more losses of genes than the *Hox*Aα cluster. Alternatively, the *Hox*5, 4, and 3 genes could have been lost in a single event in *Hox*Aβ cluster. The only genes absent in the *Hox*Aα cluster, but present in the *Hox*Aβ cluster belong to the *Hox*10 and *Hox*2 paralogy groups.

Tilapia has an almost complete *Hox*Aα cluster, in terms of presence of *Hox* genes, and no lineage-specific gene losses relative to other teleost fishes were observed. The Tilapia *Hox*Aα cluster retains the *Hox* 2, 7, and 10 genes, which are absent in the zebrafish *Hox*Aα cluster. Figure 1 summarizes the specific losses of *Hox* genes in different fish lineages. We also have preliminary evidence for the presence of a *Hox*Aβ cluster in tilapia (*Hox*A2β and *Hox*A3β; Malaga-Trillo and Meyer 2001). The increased gene loss of the *Hox*Aβ cluster compared to the *Hox*Aα cluster known from zebrafish may also be repeated in the tilapia genome.

Alignment of Nucleotide Sequences

All *Hox* clusters were screened with RepeatMasker to highlight interspersed repeats. There is a complete absence of any kind

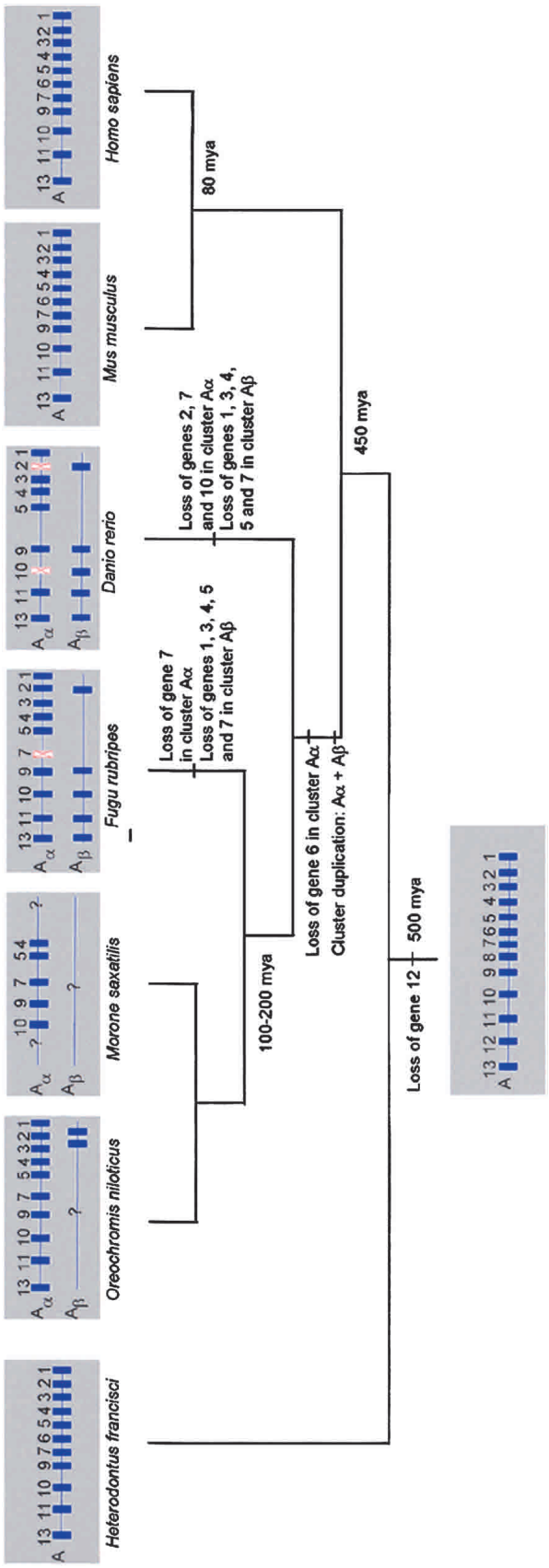


Figure 1 Evolutionary relationships among the species included in this work. The divergence date between the lineage leading to Chondrichthyes (to which *Heterodontus*, the horn shark, belongs) and that leading to the clade of all other taxa on this tree is about 500 millions years. *Actinopterygii* (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 α or β. Pseudogenes are marked with a cross. Question marks represent genomic regions that are not yet characterized.

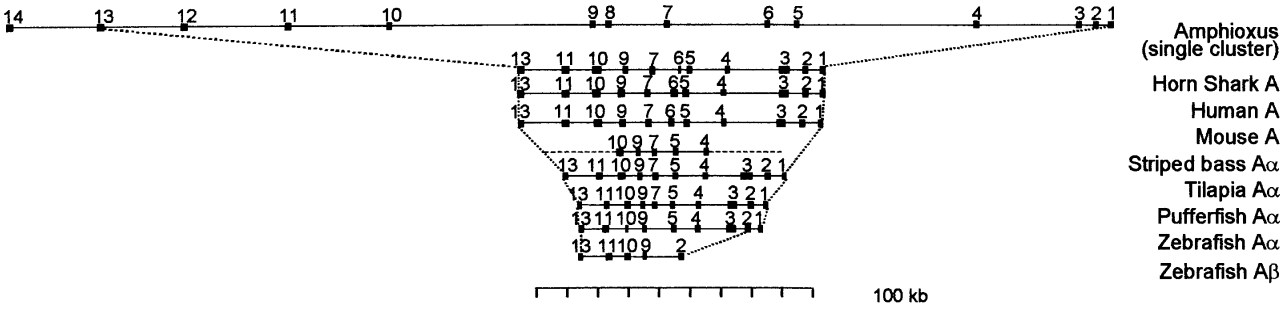


Figure 2 Relative sizes of *Hox*A clusters. Boxes represent individual genes. The duplicated α and β clusters are shown only for zebrafish. The alignable portion of the pseudogenes HoxA7α of pufferfish, HoxA2α and HoxA10α of zebrafish are shown as well.

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 *Hox*A of tilapia, pufferfish, striped bass, shark, human, and mouse clusters, and both *Hox*Aα and *Hox*Aβ clusters from zebrafish. 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 sequence regions in red are conserved regulatory regions that have been previously described in literature. As expected, coding sequences show a particularly high degree of similarity, especially in the second exon (above 75%), which contains the homeobox, while introns are generally less conserved and cannot be aligned for 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 were maintained for a period of about 500 million years of evolution. The fraction of CNSs for each intergenic region for the *Hox*A clusters is shown in Table 3. Interestingly, several 5′ and 3′ untranslated regions adjacent to the *Hox* genes of the clusters are conserved as well, 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 identified CNSs have been tested individually by using BLASTN to exclude their presence in other positions of the genomes. No matches have been found to sequences outside the *Hox* clusters (at the significance threshold of E value < 0.01). Several stretches of sequence involved in the regulation of *Hox* genes have been previously described in the literature (column 11 in Table 2), and these known regulatory sequences were also identified by our method.

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; and the alignment in the Supplementary data files available online at www.genome.org). The total number of conserved nucleotides (over 60% identity) is significantly higher ($P = 0.007$; Fig. 5) in the intergenic regions in the 3′ end of the cluster, and the detected CNSs are longer here.

Table 1. Percent Base Composition of the *Hox*A Clusters

Species	%A	%C	%G	%T
Tilapia	28.356	21.166	20.981	29.496
Pufferfish	28.476	21.398	21.093	29.033
Zebrafish α	31.231	18.816	18.378	31.574
Zebrafish β	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

Description of Some Putative Regulatory Elements

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 even a small number of nucleotides.

The first part of the intron of *Hox*A11α (51 bp) of the tilapia sequence is over 80% identical among tilapia, pufferfish, zebrafish Aα and Aβ, horn shark, humans, and the 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 three homeodomain binding sites (HB1), and it is an evolutionary conserved DNA sequence previously described from 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 paralogy group 4 *Hox* genes in medaka, chicken, the mouse, and humans (Morrison et al. 1995). The HB1 element binds *Drosophila* CAD homeoprotein and CDX-1, its homolog in the 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*Aα cluster of all the species examined. Interestingly, it is also present in the intron of *Hox*A11β of zebrafish.

The region responsible for the *cis*-regulation of the *Hox*A7 gene has previously been described by Knittel et al. (1995) as an enhancer located 1.6 kb upstream of the coding sequence in human and mouse. These authors 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 has a sequence identity of over 80% among the examined species.

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 humans and the mouse were recognized (Fig. 6). These elements are present, with minor variations, among all *Hox* genes of paralogy 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 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 was not previously described as being involved in *Hox*5 and 4 regulation, but the nature and conservation of the long stretches highlighted through this 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 that has a similarity of 85% containing a RARE element (17 bp) that 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 containing 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 containing the AAATAAAA (position 63576–63583) and ATTTAA motifs and a 16-bp stretch that is 94% conserved containing the motif TTTTATTT (position 63882–63889). This is possibly a palindromic sequence for the complementary 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 the mouse *Hox*A cluster. The *Krx20* element and the nearby "box a", described by Nonchev et al. (1996) as being involved in *Hox*A2 *trans*-activation in mouse, and present in tilapia *Hox*A cluster (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 could not identify them.

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 (column 12 in Table 2). The matches most frequently obtained are: nuclear factor *NF1* binding sites (Rossi et al. 1988), abdominal B (*Abd*B) homeobox gene binding sites (Ekker et al. 1994), *Cdx*A 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 these is the *Krx20* binding site, which was found in humans, the mouse, pufferfish, and tilapia clusters (Fig. 7A). *Krx20* binding sites have been described by Nonchev et al. (1996) as being involved in *Hox*A2 regulation as an r3/r5 enhancer that upregulates the expression of those genes in rhombomere3/rhombomere5, where *Krx20* is expressed in humans, chicks, the mouse, and pufferfish. The *Krx20* binding site is 9 bp long and occurs around 2 kb 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).

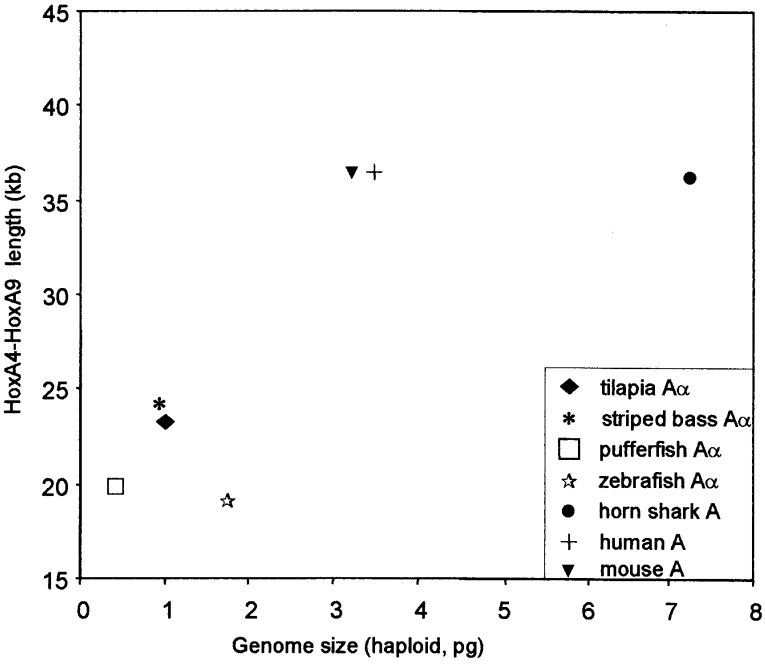
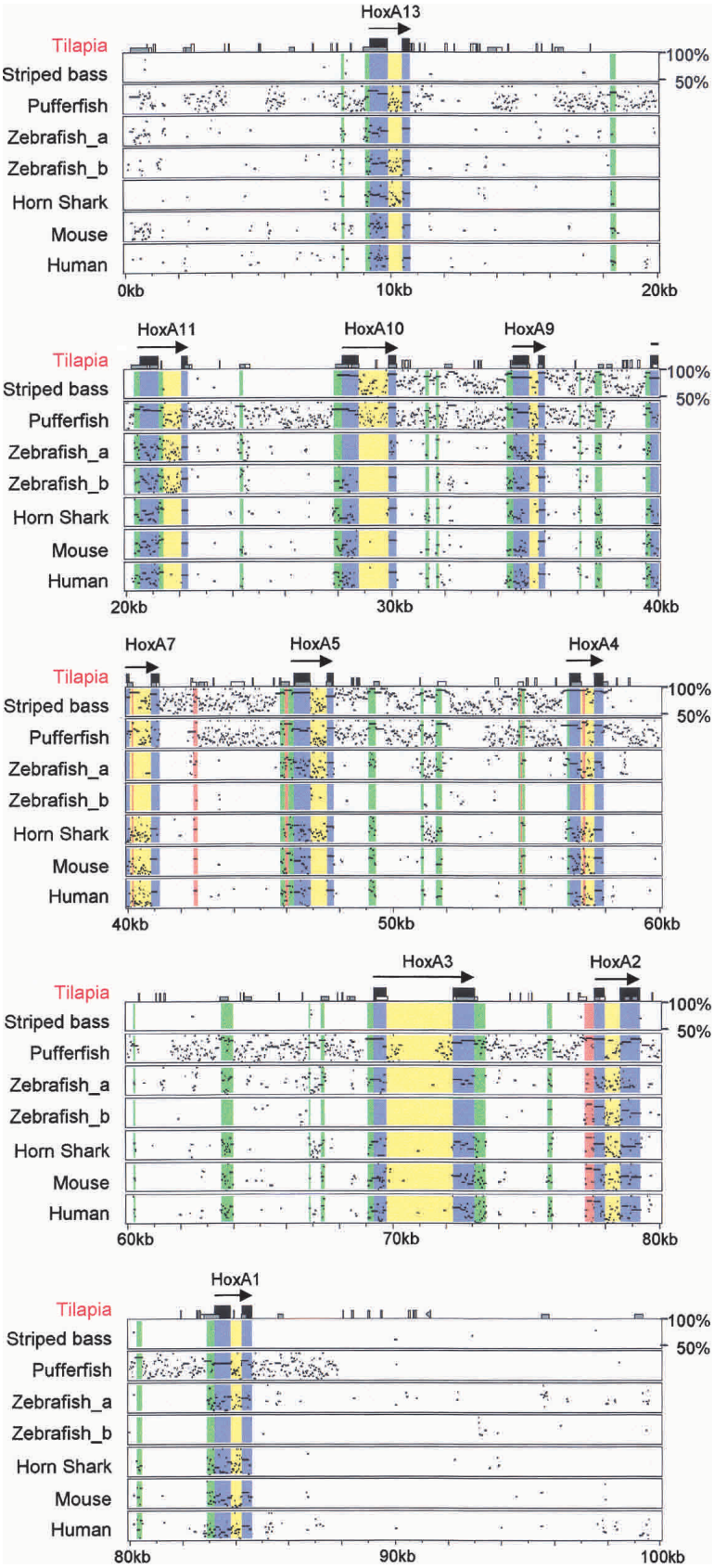


Figure 3 Relationship between genome size and length of the portion *Hox*A4 to *Hox*A9 of *Hox*A clusters. The length of *Hox*A clusters is correlated ($P = 0.06$) with genome size expressed as C value. The *Hox*Aα cluster lengths are shown. To be able to include also striped bass (*Hox*A cluster sequence is available only from *Hox*A4α to *Hox*A10α) and zebrafish (*Hox*Aα cluster lacks *Hox*A10α) into the analysis, only the length of the *Hox*A4 to *Hox*A9 portion of the cluster is shown.

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 the knowledge base about the characteristics, evolutionary conservation, and the position of functional elements with respect to the genes that they control.

We conducted several comparative analyses of the entire *Hox*A clusters for seven species of vertebrates. We compared the position and nucleotide sequence of the genes that constitute the *Hox*Aα cluster from *O. niloticus* with those of the other species in this study. The complete absence of long repetitive elements supports the idea that one of the selective forces keeping



the genes in *Hox* clusters tightly arranged stems from the fact that adjacent genes share common *cis*-regulatory elements. Interestingly, it has been suggested that repetitive elements are frequently involved in chromosomal rearrangement processes, such as inversion, translocation, and excision (Moran et al. 1999; Tomilin 1999). Hence, the absence of repetitive elements might be the result of selections against them, to reduce the risk of events that may interrupt *Hox* cluster compactness.

Degree of Conservation of Intergenic Regions

Teleost fishes, horn shark, and mammals were included in this study, to ensure comparisons of distantly related genomes, because their lineages separated approximately 450–500 millions years ago (e.g., Pough et al. 1999). Our comparative analyses were directed toward identifying conserved blocks of nucleotides among evolutionarily distantly related species that might be *cis*-acting sites for *Hox* gene-regulating factors. Intergenic regions show varying degrees of conservation (Table 3). Intergenic spaces between genes located 3′ in the clusters are significantly more conserved than those in the 5′ portion of the clusters (Fig. 5; Table 3). This pattern might be explained by the different *Hox* genes' expression patterns 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 and Dollé 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 for the correct activation of the subsequent *Hox* expression system. We found a significant

Figure 4 Pip output of the comparison of tilapia *Hox*Aα, striped bass *Hox*Aα, pufferfish *Hox*Aα, zebrafish *Hox*Aα and Aβ, 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 regions, yellow indicates introns, red indicates conserved noncoding sequences (CNSs) previously described in literature, and the green background indicates heretofore 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 gray boxes indicate a CpG/GpC ratio over 0.75. Interspersed repeat elements are shown as triangles (e.g., in position 91 kb).

Table 2. CNSs Identified Through the Comparative Approach

Position	Length (bp)	Striped bass	Pufferfish	Zebrafish a	Zebrafish b	Horn shark	Human	Mouse	Over 95%	Literature	New/similar homeobox binding sites
1 kb upstream 13	63		86	73	82	78	79	80	1 × 19		NF-1 (Rossi et al. 1988)
Imm. upstream 13	188		83	65	63	66	75	71	0		New
13-11	192		89	26		60	68	67	2 × 10		New
Imm. upstream 11	230		89	66	68	63	70	71	5 × 7–28		New
11-10	121		96	84	85		64	63	2 × 6–8		New
Imm. upstream 10	391	92	86		66	68	70	68	2 × 8–25		New
10-9 a	96	95	86	63		66	61	65	2 × 6–7		Abd B (Ekker et al. 1994); RNA pol. II cap signal (Bucher 1990)
10-9 b	95	98	98	89	81	79	73	72	1 × 24		Murine homeotic proteins b.s. (Catron et al. 1993)
Imm. upstream 9	191	94	87	63	56	69	61	63	2 × 5–6		Target sequences chicken CdxA (Margalit et al. 1993)
9-7 a	62	100	98	92	72	78	79	79	2 × 11–15		Abd B (Ekker et al. 1994)
9-7 b	276	96	93	71		71	70	69	3 × 6–11		c-ETS-1 protein b.s. (Woods et al. 1992)
Imm. upstream 7	185	95	88			79	78	78	3 × 9–14		*Hox*A7 enhancer regulatory element, *H. sapiens* (Knittel et al. 1995)
7-5	163	81	78	77		78	81	81	2 × 8–11	H8/7-6 FCS (Kim et al. 2000)	
Imm. upstream 5	529	93	84	69		38	76	76	8 × 6–39	RARE (box c and box d), *H. sapiens, M. musculus* (Odenwald et al. 1989)	
5-4 a	280	99	94	77		82	83	83	7 × 9–33		Pax b.s., (Epstein et al. 1994); Ultrabithorax b.s. (Ekker et al. 1991); target sequences of chicken CdxA homeobox gene (Margalit et al. 1993)
5-4 b	63	97	98	83		83	81	83	2 × 9–19		Dof b.s. (Yanagisawa and Schmidt, 1999)
5-4 c	209	95	93	67		71	69	69	5 × 8–15		NF of C-EBP family (Grange et al. 1991)
5-4 d	239	92	89	81		84	79	78	7 × 7–24	RARE (HoxA4 promoter, *H. sapiens,* Doerksen et al. 1996)	
Imm. upstream 4	83	100	100	89		91	78	76	3 × 6–30	RARE (HoxA4 promoter, *H. sapiens,* Doerksen et al. 1996)	
4-3 a	78		91	69		76	72	66	2 × 6–7		Dof b.s. (Yanagisawa and Schmidt, 1999)
4-3 b	480		87	67		71	66	63	5 × 6–10		New
4-3 c	51		93	72		80	75	73	2 × 6–10		New
4-3 d	136		96	76		76	66	65	4 × 8–12		New
Imm. upstream 3	235		86	90		82	73	79	6 × 7–13		New
3-2 a	476		79			61	66	60	0		New
3-2 b	189		93	72	81	68	69	67	5 × 6–9		New
Imm. upstream 2	382		89	64	78	77	77	77	8 × 8–43	HoxA2 promoter, *M. musculus* (Tan et al. 1992)	
2-1	190		93			61	72	72	3 × 10–11	RARE, *M. musculus* (Frasch et al. 1995)	New
Imm. upstream 1	352		78	60		59	64	61	2 × 6–8	RARE, *M. musculus* (Frasch et al. 1995)	New
Total	6233										

Column 1: position of CNS in the tilapia *Hox*A cluster. Column 2: length in bp of the CNS. Columns 3–9: percentage identity of the corresponding CNSs in the other genomes examined. Column 10: number (x) of occurrences and length in # of bp of highly conserved sequences with over 95% identity among all species. Column 11: reference for previously described CNSs in *Hox* clusters. Column 12: newly suggested CNSs, and reference for known binding sites that show a similar sequence.

Table 3. Base Conservation of the Intergenic Regions of the Tilapia *Hox*A Cluster

Intergenic fragment	% of total noncoding bases	% identified as CNS	% described in literature	% of total CNSs
Evx-13	13	3	0	4
13-11	14	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

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 this particular intergenic fragment.

increase in length of the CNSs between pairs of 3′ genes compared to intergenic regions of genes located 5′ and not involved in hindbrain segmentation (Fig. 5; $P = 0.007$).

In our analyses we also included 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). In addition, also if the region upstream of the *Hox*13 gene contains only regulatory elements for this gene, and the same holds true for the region downstream of the *Hox*1 gene, the trend of increasing 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 study. All the identified CNSs are specific to *Hox* clusters (no matches with any other region of the genome when aligned by using BLASTN).

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 is quite distant (by 1–5 kb; column 1 in Table 2) from both. Thus, these regions are the most interesting, because *cis*-regulatory regions in *Hox* clusters are located in positions that are intermediate between the genes they regulate. An example for this phenomenon is an element named H8/7–6 FCS (Kim et al. 2000) that was shown by Kim et al. (2000) to exist in all four clusters of mammals and shark that they compared. We showed that this element is also present in the *Hox*Aα cluster of fishes (Fig. 4; Table 2). This element is located 1.2 kb downstream of the *Hox*A7α gene and 3.6 kb upstream of the *Hox*A5α 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 the *Krx20* and "box a" in any CNS in our alignment, because the *Krx20* binding site and "box a" are short sequences that are not embedded in a block of at least 50 bp with a conservation of at least 60% in a minimum of four clusters. In this particular case, our criteria defining CNSs were too strict. Furthermore, *Hox*A1 RARE elements described by Langston et al. (1997) could not be identified, because the region downstream of *Hox*A1α was not available for most of the sequences and, hence, 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 downstream of its position, that is, the CNS upstream of *Hox*A10 is present only in *Hox*Aβ cluster, which retains the gene *Hox*A10, and was lost in *Hox*Aα cluster, which does not have the *Hox*10 gene. The same pattern is found in CNSs located upstream of the *Hox*A5, 4, and 3 genes that are present only in the *Hox*Aα 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 humans and the mouse is absent from both zebrafish *Hox* clusters. This is particularly

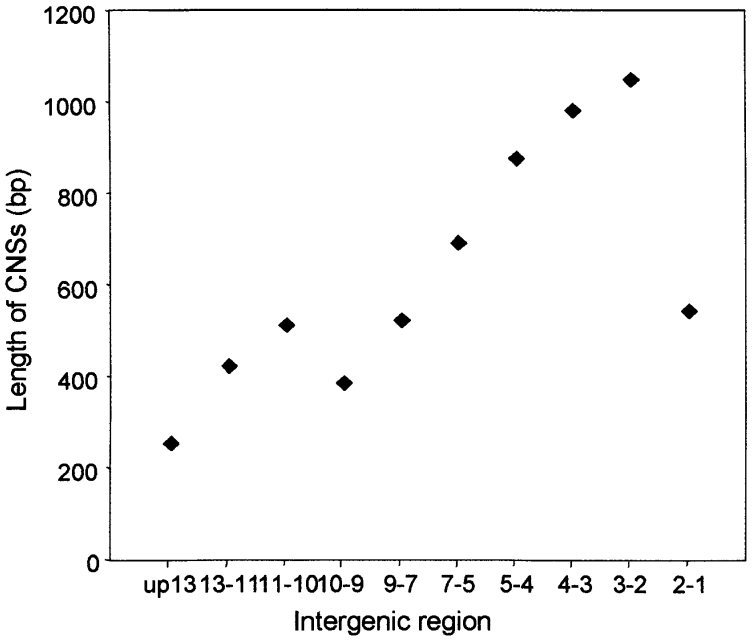


Figure 5 Total lengths of CNSs for each intergenic region. The intergenic regions located 3′ in the cluster are better conserved than those between genes located 5′ in the cluster. The graph shows the number of conserved bases (CNS as defined in text) per intergenic region. There is a significant relationship between the number of conserved bases per intergenic region and the position of the region in the cluster ($P = 0.007$).

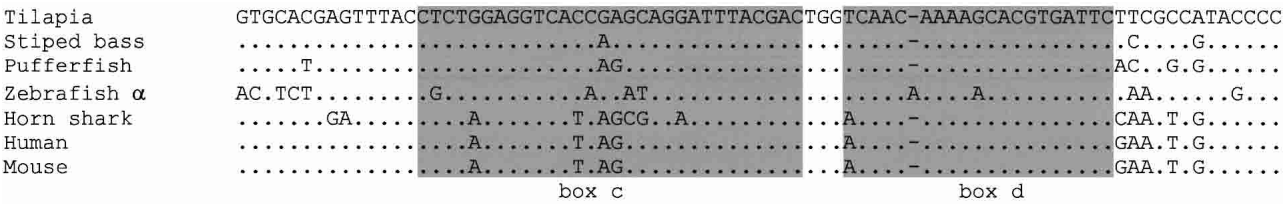


Figure 6 Alignment of known RARE elements. The alignment shows the RARE element described as "box c" and "box d" (Odenwald et al. 1989) immediately upstream of the *Hox*A5 genes. In zebrafish, the RARE element is present only in *Hox*Aα cluster (indicated as zebrafish α).

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 particular CNS has one of the lowest overall conservation levels, with no stretches being over 95% identity. These observations enforce 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 the two duplicated gene copies retain different sets of regulatory elements, and therefore, presumably different function. The set of functions of the initial gene might be divided, "subfunctionalized," by the two duplicated "daughter" copies of the gene. The *Hox*13, 11, and 9 genes are each present in two copies in the zebrafish genome, in the *Hox*Aα and Aβ clusters. The CNSs upstream of these genes are also retained in both 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, and are, hence, an example for the process of subfunctionalization 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. These differences might be due to a different method of identifying CNS sequences. Chiu et al. (2002) described, by comparison of human and horn shark *Hox*A clusters, a great number of Phylogenetic Footprints (PFs). These are defined as short blocks of noncoding DNA, typically 6 bp or more, that are 100% conserved in two taxa that have diverged at least 250 million years ago (Tagle et al. 1988; Blanchette and Tompa 2002). Among PFs, they described as Phylogenetic Footprint Clusters (PFCs) those that were found close to each other (within 200 bp) and located at comparable distances from the gene that is located 3′ of 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 and 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 Methods) (1) identity over 60% in at least four out of eight clusters; (2) presence in at least two species known to have only one *Hox*A cluster (horn shark, human, mouse; see Fig. 1), and (3) a minimum length of 50 base pairs (bp). We therefore identified a smaller number of longer conserved elements, which are shared by a higher number of species/clusters. Moreover, because of the fact that many *trans*-regulatory elements recognize a core sequence that is even shorter than 6 bp and with a certain degree of tolerance, we accepted a 95% lower threshold for the short highly conserved sequences we described (column 10 in Table 2).

Regulatory Elements in Introns

Intronic sequences are typically not conserved among evolutionarily diverged species. A clear exception to this rule are the HB1 elements, believed to be binding sites for several homeoproteins (Haerry and Gehring 1996, 1997). Our analyses show that the HB1 elements, which so far have been 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*Aα and *Hox*Aβ 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 the same *Hox* cluster. Moreover, it would be important to know if other *Hox* clusters also 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%, and is conserved among all the species included in this comparison. Part of this sequence matches a previously 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 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 (column 11 in Table 2). Moreover, many of them contain homeoprotein binding sites that are believed to be responsible for *Hox* gene regulation (column 12 in Table 2). It is reasonable to assume 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). Because *Hox* genes

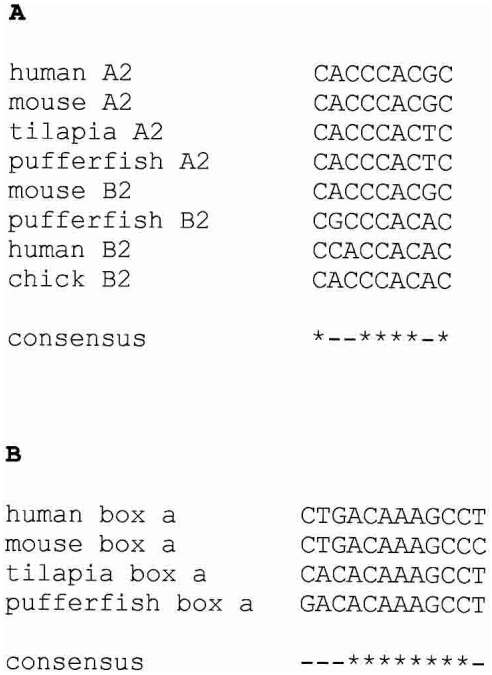


Figure 7 Alignment of known regulatory elements. (*A*) Sequence of *Krx20* binding sites in different species. Krox20 binding sites are involved in *Hox*2 regulation and they are conserved in *Hox*A and B clusters from human, mouse, pufferfish, and *Hox*A from tilapia. Both *Krx20* and the "box a" are widely conserved. The degree of identity is 67% among the species in this comparison. (*B*) Alignment of sequences of the "box a" motif.

have a colinear temporal pattern of differential expression (i.e., *Hox*A1 is expressed before *Hox*A2, and so on), 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.

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. Because functional studies involve a great deal of effort, for example, transgenic animals, it is critical to reduce the number of possible candidates for regulatory function. Sequencing projects of whole genomes (e.g., pufferfish, 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, for example, mammals or amphioxus, highlights the divergence in gene regulation of paralogous genes that evolved subsequent to gene duplication. It is still a subject of discussion whether paralogous genes in ray-finned 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 subfunctionalization 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, for example, mammals. This, in turn, would provide a method by which to elucidate different evolutionarily new functions for duplicated genes.

METHODS

The *Hox* clusters included in this study are: tilapia (*Oreochromis niloticus* AF533976, *Evx*1–*Hox*A1α), pufferfish (*Fugu rubripes*, JGI public database http://www.jgi.doe.gov/programs/fugu/fugu_mainpage.html, *Hox*A13α–*Hox*A1α), striped bass (*Morone saxatilis* AF089743, *Hox*A10α–*Hox*A4α), zebrafish (*Danio rerio* AC107365, *Evx*1–*Hox*A1α and AC107364, *Hox*A13β–*Hox*A2β), 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, *Evx*1–*Hox*A1).

The tilapia *Hox*Aα cluster sequence (Malaga-Trillo and 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 humans and the mouse. Because we used eight clusters from seven species more evolutionarily divergent than only humans and the 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*A cluster (horn shark, humans, mouse; Fig. 1) and minimum length of 50 base pairs (bp). Despite this, when taking into account only the comparison between humans and the mouse, our CNSs also fulfill the definition from Loots et al. (2000). CNSs have been tested in BLASTN (http://www.ncbi.nlm.nih.gov/BLAST/) to confirm that they 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/) to determine if they have been already described as transcription factors binding sites in similar or different biological context.

ACKNOWLEDGMENTS

This work has been supported by a grant of the Deutsche Forschungsgemeinschaft (to A.M.) and by a Marie-Curie fellowship to S.S. The authors thank E. Malaga-Trillo and other members of the Meyer-Lab for library screening, many members of the DOE Joint Genome Institute (JGI) for DNA sequencing, and C. 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, Lawrence Berkeley National Laboratory, under Contract No. DE-AC03-76SF00098.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25:** 3389–3402.

Amores, A., Force, A., Yan, Y.-L., Joly, L., Amemiya, C., Fritz, A., Ho, R.K., Langeland, J., Prince, V., Wang, Y.-L., et al. 1998. Zebrafish *Hox* clusters and vertebrate genome evolution. *Science* **282:** 1711–1714.

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 pufferfish, *Fugu rubripes*. *Proc. Natl. Acad. Sci.* **92:** 1684–1688.

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

Asif, T.C., Cook, L.L., Delehaunty, K.D., Fewell, G.A., Fulton, L.A., Fulton, R.S., Graves, T.A., Hillier, L.W., Mardis, E.R., McPherson, J.D., et al. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420:** 520–562.

Bergman, C.M. 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, W.H., Chou, C.F., Tay, A., Ponniah, S., Wang, Y., and Tan, Y.H. 2002. Conserved regulation of the lymphocyte-specific expression of lck in the Fugu and mammals. *Proc. Natl. Acad. Sci.* **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., Grenier, J., and Weatherbee, S. 2001. *From DNA to diversity—Molecular genetics and the evolution of animal design.* Blackwell Science, Malden, MA.

Catron, K.M., 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, C.H., Amemiya, C., Dewar, K., Kim, C.B., Ruddle, F.H., and Wagner, G.P. 2002. Molecular evolution of the *Hox*A cluster in the three major gnathostome lineages. *Proc. Natl. Acad. Sci.* **99:** 5492–5497.

Chu, D., Kakazu, N., Gorrin-Rivas, M., Lu, H., Kawata, M., Abe, T., Ueda, K., and Adachi, Y. 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, A.G. 2001. The search for meaning in noncoding DNA. *Genome Res.* **11:** 1319–1320.

Cliften, P.F., Hillier, L.W., Fulton, L., Graves, T., Miner, T., Gish, W.R., Waterston, R.H., and Johnston, M. 2001. Surveying *Saccharomyces* genomes to identify functional elements by comparative DNA sequence analysis. *Genome Res.* **11:** 1175–1186.

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

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.

Ekker, S.C., Young, K.E., von Kessler, D.P. and Beachy, P.A. 1991. Optimal DNA sequence recognition by the ultrabithorax homeodomain of *Drosophila*. *EMBO J.* **10:** 1179–1186.

Ekker, S.C., Jackson, D.G., von Kessler, D.P., Sun, B.I., Young, K.E., and Beachy, P.A. 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:** 8355–8361.

Ferrier, D.E., Minguillon, C., Holland, P.W., 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, F.B., Amores, A., Yan, Y.L., and Postlethwait, J. 1999. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151:** 1531–1545.

Frasch, M., Chen, X., 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 Holland, P.W. 1994. Archetypal organization of the amphioxus *Hox* gene cluster. *Nature* **370:** 563–566.

Gehring, W.J. 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. *Nucleic Acids Res.* **19:** 131–139.

Haerry, T.E. and Gehring, W.J. 1996. 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.* **93:** 13884–13889.

———. 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.

Hardison, R.C. 2000. Conserved noncoding sequences are reliable guides to regulatory elements. *Trends Genet.* **16:** 369–372.

Hardison, R.C., 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, H.G., Wolke, U., Lunde, K., Soll, I., Abdelilah-Seyfried, S., Prince, V., and Driever, W. 2002. spiel ohne grenzen/pou2 is required for zebrafish hindbrain segmentation. *Development* **129:** 1645–1655.

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

Holland, P.W. 1997. Vertebrate evolution: Something fishy about *Hox* genes. *Curr. Biol.* **7:** R570–R572.

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.* **97:** 1655–1660.

Knittel, T., Kessel, M., Kim, M.H., 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, A.W., Thompson, J.R., and Gudas, L.J. 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.

Loots, G.G., Locksley, R.M., Blankespoor, C.M., Wang, Z.E., Miller, W., Rubin, E.M., and Frazer, K.A. 2000. Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science* **288:** 136–140.

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 Meyer, A. 2001. Genome duplication and accelerated evolution of *Hox* genes and cluster architecture in teleosts fishes. *Am. Zool.* **41:** 676–686.

Manzanares, M., Wada, H., Itasaki, N., Trainor, P.A., Krumlauf, R., and Holland, P.W.H. 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. *Nucleic 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, J.V., DeBerardinis, R.J., and Kazazian Jr., H.H. 1999. Exon shuffling by L1 retrotransposition. *Science* **283:** 1530–1534.
Morrison, A., Chaudhuri, C., Ariza-McNaughton, L., Muchamore, I., Kuroiwa, A., and Krumlauf, R. 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, M.H., 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, W.F., Garbern, J., Arnheiter, H., Tournier-Lasserve, E., and Lazzarini, R.A. 1989. The *Hox*-1.3 homeo box protein is a sequence-specific DNA-binding phosphoprotein. *Genes & Dev.* **3:** 158–172.
Ohno, S. and Atkin, N.B. 1966. Comparative DNA values and chromosome complements of eight species of fishes. *Chromosoma* **18:** 455–466.
Onyango, P., Miller, W., Lehoczky, J., Leung, C.T., Birren, B., Wheelan, S., Dewark, K., and Feinberg, A.P. 2000. Sequence and comparative analysis of the mouse 1-Megabase region orthologous 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 & Dev.* **1:** 891–898.
Pough, F.H., Janis, C.M., and Heiser, J.B. 1999. *Vertebrate life*. Prentice Hall, Englewood Cliffs, NJ.
Prince, V.E., Joly, L., Ekker, M., and Ho, R.K. 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, P.L., Zelus, D., Hughes, S., and Laudet, V. 2001. Euteleost fish genomes are characterized by expansion of gene families. *Genome Res.* **11:** 781–788.
Rossi, P., Karsenty, G., Roberts, A.B., Roche, N.S., Sporn, M.B., 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-β. *Cell* **52:** 405–414.
Roth, F.P., Hughes, J.D., Estep, P.W., and Church, G.M. 1998. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nat. Biotechnol.* **16:** 939–945.
Schwartz, S., Zhang, Z., Frazer, K.A., 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.
Snell, E.A., Scemama, J.-L., and Stellwag, E.J. 1999. Genomic organization of the *Hoxa4-Hoxa10* region from *Morone saxatilis*: Implications for *Hox* gene evolution among vertebrates. *J. Exp. 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, D.A., Koop, B.F., Goodman, M., Slightom, J.L., Hess, D.L., and Jones, RT. 1988. Embryonic ϵ and γ 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, C.A., Guo, V., and Nirenberg, M. 1992. Murine *Hox*-1.11 homeobox gene structure and expression. *Proc. Natl. Acad. Sci.* **89:** 6280–6284.
Taylor, J.S., 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, T.R., Chandler, R.W., Wachtel, S.S., and Elias, S. 1989. Reference standards for flow cytometry and application in comparative studies of nuclear DNA content. *Cytometry* **10:** 706–710.
Tomilin, N.V. 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, C.P., Alkema, M.J., van Weperen, W.W., Van Leeuwen, H.C., Strating, M.J., and van der Vliet, P.C. 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, A.E. 1998. Genome size and GC-percent in vertebrates as determined by flow cytometry: The triangular relationship. *Cytometry* **31:** 100–109.
Wasserman, W.W., Palumbo, M., Thompson, W., Fickett, J.W., and Lawrence, C.E. 2000. Human–mouse genome comparisons to locate regulatory sites. *Nat. Genet.* **26:** 225–228.
Weiss, K.M. and Fullerton, S.M. 2000. Phenogenetic drift and the evolution of genotype–phenotype relationships. *Theor. Popul. Biol.* **57:** 187–195.
Wittbrodt, J., Meyer, A., and Schartl, M. 1998. More genes in fish? *BioEssays* **20:** 511–515.
Woods, D.B., Ghysdael, J., and Owen, M.J. 1992. Identification of nucleotide preferences in DNA sequences recognized specifically by c-ETS-1 protein. *Nucleic Acids Res.* **20:** 699–704.
Yanagisawa, S. and Schmidt, R.J. 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 SITE 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.

Received August 8, 2002; accepted in revised form March 24, 2003.