Article

Regulatory Roles of Conserved Intergenic Domains in Vertebrate *Dlx* Bigene Clusters

Noël Ghanem,[1,2,7] Olga Jarinova,[1,2,7] Angel Amores,[3] Qiaoming Long,[1,2,5] Gary Hatch,[1] Byung Keon Park,[1,6] John L.R. Rubenstein,[4] and Marc Ekker[1,2,8]

[1]Ottawa Health Research Institute and [2]Department of Cellular and Molecular Medicine, University of Ottawa, Ottawa, Ontario, Canada K1Y 4E9; [3]Institute of Neuroscience, University of Oregon, Eugene, Oregon 97403, USA; [4]Nina Ireland Laboratory of Developmental Neurobiology, Center for Neurobiology and Psychiatry, Department of Psychiatry and Programs in Neuroscience, Developmental Biology and Biomedical Sciences, University of California at San Francisco, California 94143–0984, USA.

Dlx homeobox genes of vertebrates are generally arranged as three bigene clusters on distinct chromosomes. The *Dlx1/Dlx2*, *Dlx5/Dlx6*, and *Dlx3/Dlx7* clusters likely originate from duplications of an ancestral *Dlx* gene pair. Overlaps in expression are often observed between genes from the different clusters. To determine if the overlaps are a result of the conservation of enhancer sequences between paralogous clusters, we compared the *Dlx1/2* and the *Dlx5/Dlx6* intergenic regions from human, mouse, zebrafish, and from two pufferfish, *Spheroides nephelus* and *Takifugu rubripes*. Conservation between all five vertebrates is limited to four sequences, two in *Dlx1/Dlx2* and two in *Dlx5/Dlx6*. These noncoding sequences are >75% identical over a few hundred base pairs, even in distant vertebrates. However, when compared to each other, the four intergenic sequences show a much more limited similarity. Each intergenic sequence acts as an enhancer when tested in transgenic animals. Three of them are active in the forebrain with overlapping patterns despite their limited sequence similarity. The lack of sequence similarity between paralogous intergenic regions and the high degree of sequence conservation of orthologous enhancers suggest a rapid divergence of *Dlx* intergenic regions early in chordate/vertebrate evolution followed by fixation of *cis*-acting regulatory elements.

[Supplemental material is available online at www.genome.org.]

Vertebrates possess anatomical features not seen in their closest living invertebrate relatives, the protochordates such as tunicates and cephalochordates. Genetic changes, such as the evolution of new regulatory pathways, may have permitted the origin of these innovations. Gene duplication followed by functional divergence of paralogs constitutes a major mechanism that permits such changes. An important contribution to the evolutionary divergence of paralogs may be through changes in mechanisms that control gene expression via *cis*-acting regulatory sequences in the noncoding region of genes. However, the identification of *cis*-acting regulatory elements remains challenging, even after the completion of a few vertebrate genome sequences.

The vertebrate *Dlx* genes, which encode a family of homeobox-containing transcription factors related in sequence to the *Drosophila Distal-less* (*Dll*) gene product, constitute one example of functional diversification of paralogs. All vertebrates investigated thus far have at least six *Dlx* genes that are generally arranged as three bigene clusters: *Dlx1/Dlx2*, *Dlx5/Dlx6*, and *Dlx3/Dlx7* (Simeone et al. 1994; McGuinness et al. 1996; Nakamura et al. 1996; Stock et al. 1996; Ellies et al. 1997; Liu et al. 1997). Each bigene cluster is localized on a

distinct chromosome that also contains one of the *Hox* clusters, suggesting that the duplication events that generated the multiple *Dlx* bigene clusters of vertebrates also involved the *Hox* genes (Stock et al. 1996; Amores et al. 1998). The two linked *Dlx* genes are in an inverted configuration and separated by a short intergenic (3.5–16 kb) region. Because only one *Dll*-like gene is found in invertebrates such as *Drosophila* and *Caenorhabditis elegans*, the multiple vertebrate *Dlx* genes are thought to have arisen as a result of tandem gene duplication events from one "hypothetical" common ancestor to nematodes, arthropods, and vertebrates. The presence, in the tunicate *Ciona intestinalis* of pair of *Dll*-like gene with an organization similar to that of the vertebrate *Dlx* (Di Gregorio et al. 1995; Caracciolo et al. 2000) supports the hypothesis that the initial duplication predated the existence of vertebrates.

Gene families such as the *Dlx* family provide attractive models for studying gene regulation and functional divergence between paralogs. The bigene cluster arrangement of *Dlx* genes is conserved amongst distant vertebrates and a direct association is seen between the genomic organization of the genes and their expression pattern in different species (Ellies et al. 1997; Zerucha et al. 2000) suggesting that the mechanisms of regulation might have been conserved, at least in part. Functional conservation among different orthologs, as inferred from comparative expression patterns seems to be applicable to most vertebrate *Dlx* genes (Quint et al. 2000; Zerucha and Ekker 2000). Partial functional redundancy between *Dlx* paralogs is suggested by the overlapping gene expression patterns and phenotypes of mice with targeted *Dlx* mutations (Qiu et al. 1995, 1997; Anderson et al. 1997; Acam-

Present address: [5]Department of Molecular Physiology and Biophysics, Vanderbilt University Medical Center, Nashville, TN 37232, USA; [6]Department of Oral Anatomy, School of Dentistry, Chonbuk National University, Chonju, Republic of Korea.
[7]These authors contributed equally to this work.
[8]Corresponding author.
E-MAIL mekker@ohri.ca; FAX (613) 761-5036.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.716103.

pora et al. 1999; Depew et al. 1999; Robledo et al. 2002). Sharing of *cis*-regulatory elements between members of a *Dlx* bigene cluster may contribute to the overlap in gene expression and to their partial functional redundancy.

Consistent with a model of enhancer-sharing, two highly conserved enhancer elements, I56i and I56ii, were identified in the intergenic region of the *Dlx5/Dlx6* genes of zebrafish, mouse, and human and were able to target expression of reporter transgenes to the forebrain of both mouse and zebrafish in patterns that mimic the endogenous gene expression (Zerucha et al. 2000). Recently, Sumiyama and collaborators conducted a comparative sequence analysis of the mouse and human *Dlx3/Dlx7* (*Dlx3/Dlx4* was suggested as revised nomenclature by Panganiban and Rubenstein 2002) bigene cluster (Sumiyama et al. 2002). Conserved sequences were identified both in the coding and noncoding regions of *Dlx3/Dlx7*. Comparisons of the two mammalian loci with the orthologous *dlx3/dlx7* bigene cluster from zebrafish revealed a much more limited similarity (Sumiyama et al. 2002).

The two genes from the *Dlx1/Dlx2* cluster are expressed in the developing forebrain with patterns that overlap partially with those of *Dlx5* and *Dlx6*. As the *Dlx1/Dlx2* and *Dlx5/Dlx6* bigene clusters probably originate from the duplication of an ancestral cluster, the forebrain expression of *Dlx1* and *Dlx2* could be attributable to enhancer sequences related to I56i and/or I56ii. To address this possibility and to get a comprehensive understanding of *cis*-acting regulatory elements in the *Dlx1/Dlx2* and *Dlx5/Dlx6* intergenic regions, we have performed a homology search (phylogenetic footprinting) between the intergenic regions of the two bigene clusters from five vertebrate species: human, mouse, zebrafish, *Takifugu rubripes* (formerly *Fugu rubripes*) and *Spheroides nephelus*. Sequence conservation between all five species is limited to four distinct sequences of a few hundred base pairs, two in each intergenic region. Each sequence shows enhancer activity in transgenic mice and/or zebrafish. A novel forebrain enhancer, I12b, was identified in the *Dlx1/Dlx2* intergenic region, but surprisingly, it shows almost no sequence similarity to the I56i and I56ii forebrain enhancers, suggesting that highly overlapping patterns of expression can be conferred by highly different *cis*-acting regulatory sequences.

RESULTS

Genomic Organization of *Dlx1/Dlx2* and *Dlx5/Dlx6* Bigene Clusters in Two Species of Pufferfish

The genomic organization of two loci containing *Dlx* genes was examined in *Spheroides nephelus* and *Takifugu rubripes* and was compared to that of zebrafish, mouse, and human. Initial orthology assignment was based on the sequence of the third exon of the genes, which contains part of the homeobox. Orthology was further confirmed by sequence analysis of the intergenic region. As previously described for zebrafish, mouse, and human (Simeone et al. 1994; McGuinness et al. 1996; Ellies et al. 1997; Zerucha et al. 2000), the *dlx1/dlx2* genes and the *dlx5/dlx6* genes of *Spheroides* and *Takifugu* are organized as two pairs of genes, both found in an inverted and convergent configuration (Figs. 1A, 2A).

The size of the *Dlx1/Dlx2* intergenic region in the five species varies between about 4.5–5.0 kb for the two pufferfish to 10.7 kb for human (Fig. 1A). It was difficult to determine with precision the size of the pufferfish intergenic regions because no cDNA sequences are available for the *Dlx1* and *Dlx2* genes from these species and unequivocal polyadenyla-
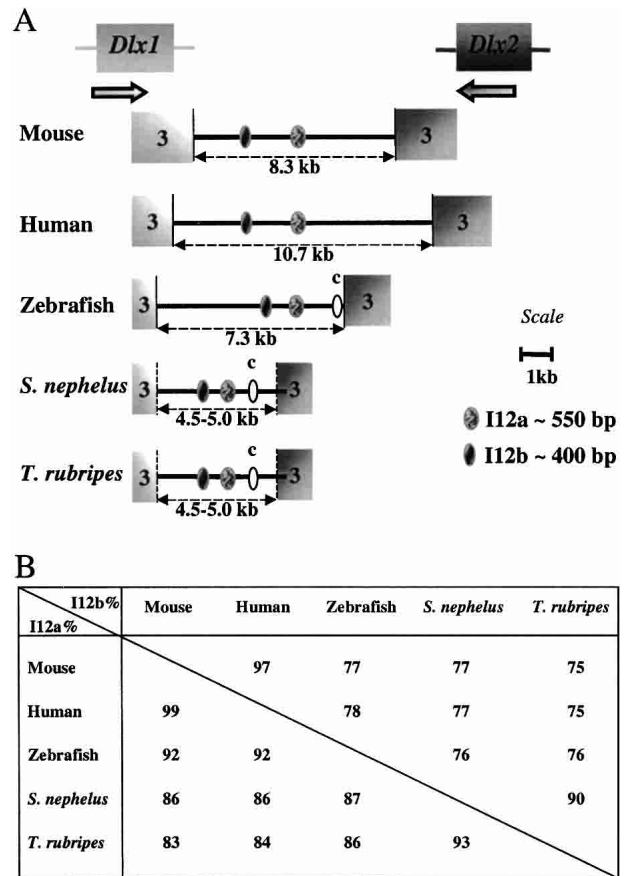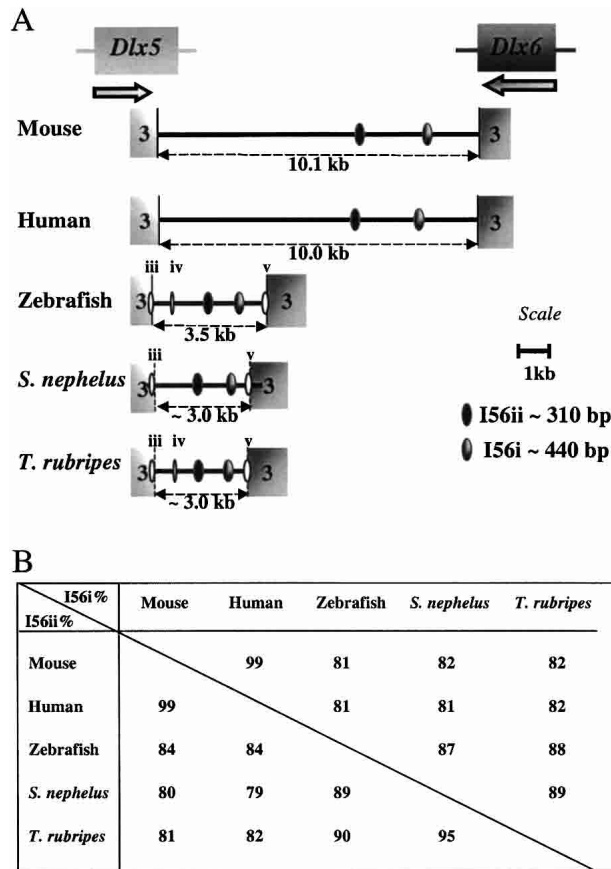

Figure 1 Conserved sequences in the *Dlx1/Dlx2* intergenic region. (*A*) Schematic representation of the *Dlx1/Dlx2* intergenic region of five vertebrate species. The third exons of the *Dlx* genes are indicated. The position of the polyadenylation sequence in the *Dlx* genes of *Spheroides* and *Takifugu* is an estimate. In addition to the I12a and I12b sequences, ovals labeled "c" represent a region of sequence conservation between the three teleost fish species. (*B*) Percentage identity for I12a and I12b in pairwise sequence comparisons.

tion signals were sometimes hard to find in the genomic sequence. The distance that separates the two stop codons is 5.3 kb in both species.

The size of the *Dlx5/Dlx6* intergenic region varied between 10 kb for mouse and human and about 3.0–3.5 kb for the three teleost fish (Fig. 2A). Thus despite the fact that the genome size for *Takifugu rubripes* and *Spheroides nephelus* is ~4 and 8 times smaller than those of the zebrafish and mouse/human, respectively, this is not reflected in proportionally smaller intergenic regions.

Sequence Comparisons and Identification of Highly Conserved Noncoding Sequence Elements in the *Dlx* Intergenic Regions

We examined the *Dlx1/Dlx2* and *Dlx5/Dlx6* intergenic regions of the five vertebrate species for conserved sequences. The mouse and human *Dlx1/Dlx2* intergenic regions were highly similar with 80% overall sequence identity (Fig. 3A). The same applies for the human *Dlx5/Dlx6* intergenic region (78% Fig. 3B) and for the *dlx1/dlx2* and *dlx5/dlx6* intergenic regions of *Takifugu rubripes* and *Spheroides nephelus* with 85% and 87% sequence identity, respectively (data not shown). This reflects

A

B

I56i % I56ii %	Mouse	Human	Zebrafish	*S. nephelus*	*T. rubripes*
Mouse		99	81	82	82
Human	99		81	81	82
Zebrafish	84	84		87	88
S. nephelus	80	79	89		89
T. rubripes	81	82	90	95	

Figure 2 Conserved sequences in the *Dlx5/Dlx6* intergenic region. (*A*) Schematic representation of the *Dlx5/Dlx6* intergenic region of five vertebrate species. The third exon of the *Dlx* genes are indicated. The position of the polyadenylation sequence in the *Dlx* genes of *Spheroides* and *Takifugu* is an estimate. In addition to the I56i and I56ii sequences, ovals labeled iii, iv, and v represent regions of sequence conservation between a subset of the five species. Sequence alignments can be found as supplemental files. (*B*) Percentage identity for I56i and I56ii in pairwise sequence comparisons.

the relatively recent divergence from one common ancestor between mouse and human (~60 million years), on the one hand, and between the two species of pufferfish, on the other hand (between 5–35 million years). Despite the high degree of sequence conservation between orthologous loci, the paralogous intergenic regions, *Dlx1/Dlx2* and *Dlx5/Dlx6*, do not show any striking sequence similarity and no large regions of sequence similarity can be found between the intergenic sequence separating *Dlx3* and *Dlx7* of human, mouse, and zebrafish (Sumiyama et al. 2002).

Two highly conserved sequences that were previously identified in the *Dlx5/Dlx6* intergenic region of zebrafish, mouse, and human (Zerucha et al. 2000), I56i and I56ii, were also found in the *dlx5/dlx6* intergenic regions of *Takifugu* and *Spheroides*. They constitute the only two regions of high sequence similarity between all five species (Fig. 2A, 3B). The sizes of I56i and I56ii are ~440 bp and 310 bp, respectively, and the identity percentages in pairwise comparisons vary between 81 and 99% (Fig. 2B; five-species alignment provided as supplementary Figs. 1 and 2). The relative positions and orientation of the I56i and I56ii sequences with respect to the flanking genes were identical for all five vertebrates. In both

the mouse/human (Fig. 3B) and the *Takifugu/Spheroides* (not shown) alignments, I56i and I56ii reside in a region of overall stronger sequence conservation.

In addition to I56i and I56ii, we found two sequences of 150–200 bp with >80% identity between zebrafish, *Takifugu*, and *Spheroides* (Fig. 2A; alignments provided as supplementary Figs. 3 and 4). The first is found in the 3′UTR sequence of zebrafish *dlx5a* (see note concerning the nomenclature of zebrafish *dlx* genes in the Methods section) and at a corresponding position, with respect to the predicted stop codons of the *Takifugu* and *Spheroides* orthologs (Fig. 2A). The second is found just downstream of the 3′UTR of zebrafish *dlx6a* and at a similar position in the pufferfish orthologs. Finally, a fragment of about 100 bp with 83% sequence identity was found between the end of *dlx5a* and I56ii in zebrafish and *Takifugu* but was not found in *Spheroides* (alignment provided as supplementary Fig. 5). None of the three shorter conserved sequences could be identified in the two mammalian loci.

We identified two highly conserved sequences in the *Dlx1/Dlx2* intergenic regions of the five vertebrates. The first, I12a, is ~550 bp in length and the percentages in sequence identity in pairwise comparisons vary between 83% and 99% (Figs. 1B, 4). The second, I12b, is about 400 bp in length and shows percentages of identity that vary between 75% and 97% (Figs. 1B, 5). The relative positions and orientations of I12a and I12b with respect to the *Dlx1* and *Dlx2* genes were identical in all five species. As for I56i and I56ii, the I12a and I12b sequences reside in a region of overall stronger sequence conservation in mouse/human (Fig. 3A) and in *Takifugu/Spheroides* (not shown) pairwise comparisons.

In addition to I12a and I12b, we identified a sequence of ~320 bp, I12c, that was conserved between *Takifugu*, *Spheroides*, and zebrafish. This sequence is located between the end of *dlx2* and I12a (Fig. 1A; alignment provided as supplementary Fig. 6). Finally, a sequence of ~110 bp was found in or near the 3′UTR of *Dlx1* of mouse and human and the in the zebrafish *dlx1/dlx2* locus, between the 3′end of *dlx1* and I12b (alignment provided as supplementary Fig. 7). This sequence contains a TTA tri-nucleotide repeat but sequence conservation extends beyond this repeat.

The Sequences Conserved Between All Five Vertebrate Species Contain Enhancers

To determine that the conserved *Dlx* intergenic sequences, I56i, I56ii, I12a, and I12b, constitute *cis*-acting regulatory sequences, they were tested in reporter constructs that were injected to produce transgenic mice and zebrafish. As previously reported, I56i and I56ii target expression of *lacZ* reporter constructs to the forebrain of transgenic mice and zebrafish starting at E10 and persisting in adult mice (Zerucha et al. 2000). The mouse I56i sequence can efficiently target expression to the forebrain by itself in 100% of primary transgenic mice expressing the transgene and in three out of four transgenic lines (Fig. 6A; Table 1) (Zerucha et al. 2000). The zebrafish I56i sequence also targeted expression to the forebrain of 12 out of 12 primary transgenic mouse embryos (Zerucha et al. 2000). In both cases, reporter gene expression precisely mimics that of the endogenous *Dlx5* gene and highly overlaps with that of *Dlx6* (Zerucha et al. 2000).

Three primary transgenic mice and two established lines containing a mouse I56ii reporter construct expressed *lacZ* in the forebrain (Fig. 6B), although the intensity of the ß-galactosidase staining was more variable between the telence-
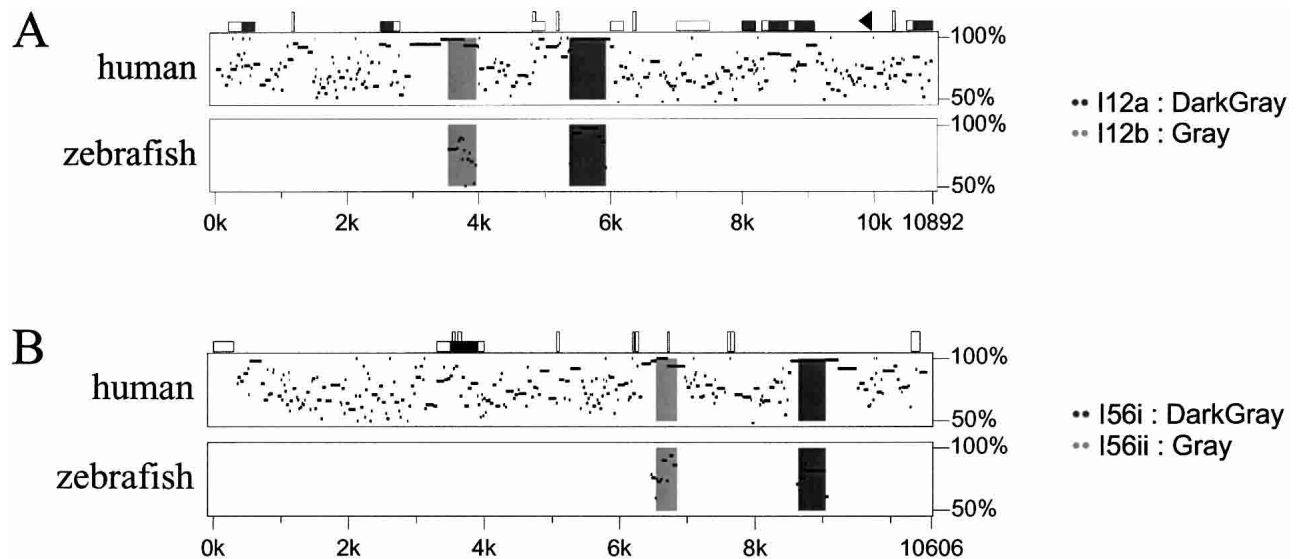
Figure 3 Percentage identity plot (PIP) of the (*A*) *Dlx1/2,* and (*B*) *Dlx5/6* intergenic regions between mouse, human, and zebrafish. The mouse sequence is shown on the horizontal axis and the percentage identity to the human (*top plot*) and zebrafish sequences (*lower plot*) are shown on the vertical axis. Sequences used for comparison include the intergenic regions and the 3′UTRs of both flanking genes. In *A*, *Dlx1* is to the *left* and in *B*, *Dlx5* is to the *left*. Shaded dark and light gray areas indicate the positions of enhancers. Repetitive sequences are shown as follows: black triangles, mammalian interspersed repeats (MIR); vertical rectangles, simple sequence repeats; CpG islands: white horizontal rectangle, CpG ratio >0.60; gray rectangles, CpG ratio >0.75. For further details on PIP analyses, see http://bio.cse.psu.edu/pipmaker.

phalic and diencephalic expression domains, and staining seemed often weaker than that observed with I56i constructs. However, the mouse I56ii (this work) was more efficient at targeting transgene expression to the forebrain than its zebrafish counterpart (Zerucha et al. 2000).

When tested in transgenic zebrafish, a construct containing both zebrafish I56i and I56ii targeted expression of the green fluorescent protein (GFP) reporter transgene to the domains of *dlx* expression in the telencephalon and diencephalon (Fig. 6G,H). In this transgene construct, GFP is placed immediately downstream of a 3.5-kb fragment of the dlx6a 5′-flanking region including the promoter and part of the 5′UTR. This 5′-flanking fragment does not, by itself, target expression of GFP in a specific manner (Fig. 6F; no reproducible pattern in >150 embryos injected). However, in the presence of the zebrafish enhancers, 75–80% of injected embryos (n>400) had forebrain expression starting at 18 h postfertilization (hpf) and lasting until at least 96 hpf. Three transgenic lines could be produced all with comparable expression patterns and intensity. An embryo from one line is shown in Figure 6G and H. In contrast, the same intergenic fragment coupled to the ß-globin minimal promoter, which was used for transgenic mouse constructs, showed forebrain expression in only 8% of injected embryos and only 0.5% of them had more than 10 GFP-positive cells (Zerucha et al. 2000). The difference between efficiency of the human ß-globin minimal promoter fragment between human and zebrafish is, at present, unclear.
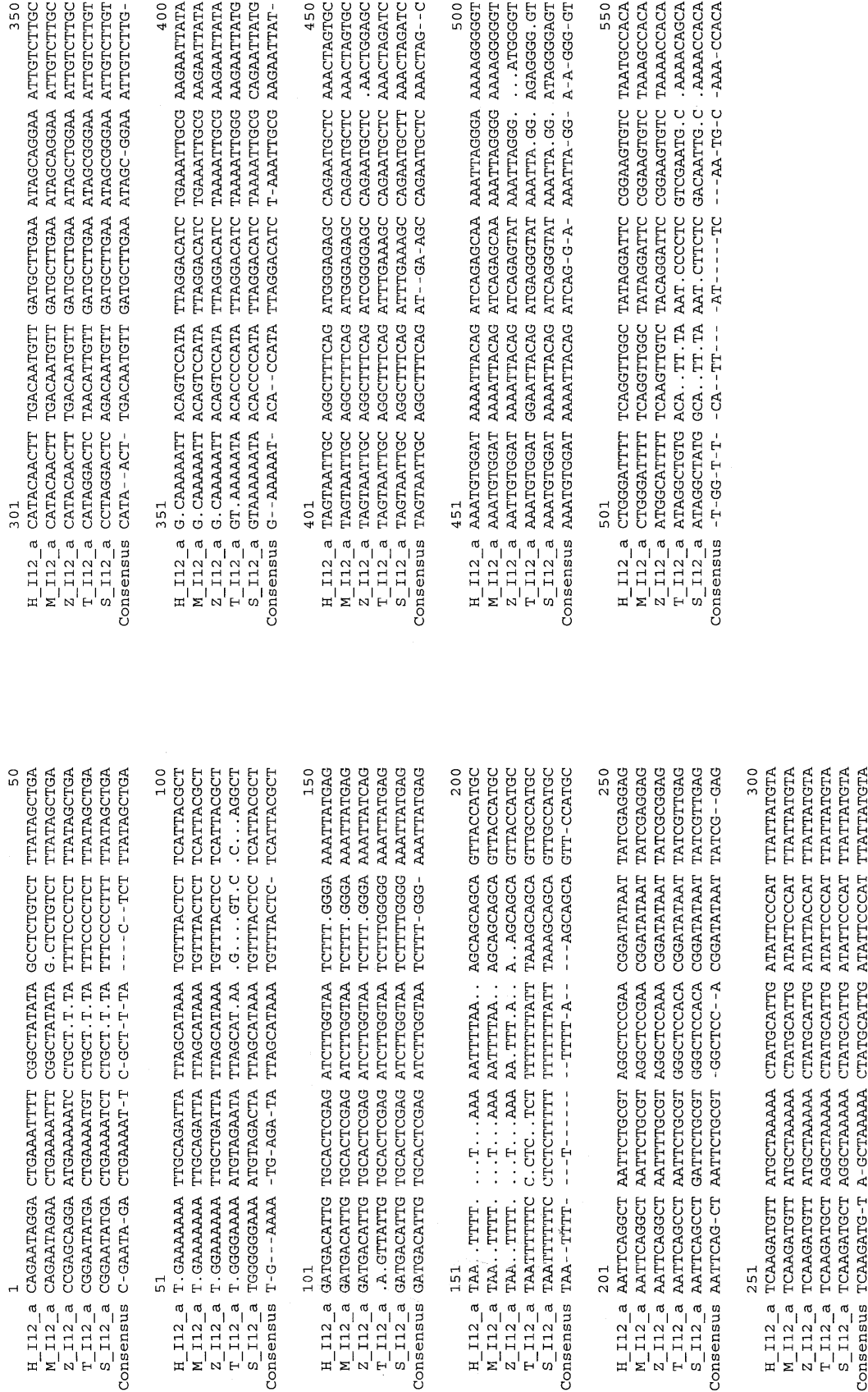
Similar transgene constructs containing the mouse I56i sequence (Fig. 6I) or a combination of I56i and I56ii (Fig. 6J), inserted in the *5′-dlx6a-GFP* plasmid, expressed GFP in the forebrain of transgenic zebrafish although the proportions of transgenic embryos were smaller than those observed with the corresponding construct containing zebrafish sequences. Thus, for both constructs, 35–40% embryos showed forebrain expression (n > 150 for each construct) with most of the GFP-positive cells in the telencephalic domain of *dlx* expression (Fig. 6 I,J).

The mouse I12b conserved sequence targeted reporter transgene expression to the forebrain of transgenic mice, starting at E10 and lasting until E16, the latest time point examined (Fig. 6C; Table 1; 3/3 primary embryos and 5/5 transgenic lines). This construct also produced expression in the apical ectodermal ridge, another site of endogenous *Dlx* expression although expression was more variable in intensity (Table 1) compared to that observed in the forebrain. Preliminary examination of sections of brains from lines of transgenic mice expressing the *I12b-lacZ* construct indicates that the constructs faithfully mimic expression of *Dlx1/Dlx2* in the telencephalon and diencephalon (data not shown). Thus, despite the fact that their sequences are highly divergent (see below), the three intergenic sequences, I56i, I56ii, and I12b, act as *cis*-acting forebrain enhancers with highly overlapping patterns of activity.

A 1.9-kb *Xba*1-*Eco*R1 fragment containing the I12a conserved sequence targeted *lacZ* expression to a subset of *Dlx*-expressing cells in the mesenchyme of the mandibular component of the first branchial arch and in the hyoid arch starting at E9.5 and lasting until at least E16, when expression gradually diminishes (Fig. 6D,E; Table 1; B.K. Park, S. Sperber, B.L. Thomas, G. Hatch, N. Ghanem, P.T. Sharpe, and M. Ekker, unpubl. observations). Reporter transgene expression was observed in six out of seven transgenic lines (Table 1). A 1.6-kb *Xho*1 fragment containing zebrafish I12a targeted expression in one out of two lines of transgenic mice (Table 1).

As the *Dlx1/Dlx2* intergenic regions of mouse and human showed sequence conservation that extended beyond the above two enhancers (Fig. 3A), we produced transgenic mice with reporter constructs containing mouse intergenic fragments outside I12a and I12b. Thus, a construct containing a 1.5-kb DNA fragment located between I12a and I12b, with 80% identity between mouse and human (Figs. 1, 3A), did not show enhancer activity in mouse embryos (zero out of three primary transgenic embryos, as determined by detection of

```
         1                                                          50
H_I12_a   CAGAATAGGA CTGAAAATTT CGGCTATATA GCCTCTGTCT TTATAGCTGA
M_I12_a   CAGAATAGAA CTGAAAATTT CGGCTATATA G.CTCTGTCT TTATAGCTGA
Z_I12_a   CCGAGCAGGA ATGAAAAATC CTGCT.T.TA TTTTCCCTCT TTATAGCTGA
T_I12_a   CGGAATATGA CTGAAAATGT CTGCT.T.TA TTTCCCCTCT TTATAGCTGA
S_I12_a   CGGAATATGA CTGAAAATCT CTGCT.T.TA TTTCCCCTTT TTATAGCTGA
Consensus C-GAATA-GA CTGAAAAT-T C-GCT-T-TA -----C--TCT TTATAGCTGA

         51                                                        100
H_I12_a   T.GAAAAAAA TTGCAGATTA TTAGCATAAA TGTTTACTCT TCATTACGCT
M_I12_a   T.GAAAAAAA TTGCAGATTA TTAGCATAAA TGTTTACTCT TCATTACGCT
Z_I12_a   T.GGAAAAAA TTGCTGATTA TTAGCATAAA TGTTTACTCC TCATTACGCT
T_I12_a   T.GGGGAAAA ATGTAGAATA TTAGCAT.AA .G....GT.C .C...AGGCT
S_I12_a   TGGGGGGAAA ATGTAGACTA TTAGCATAAA TGTTTACTCC TCATTACGCT
Consensus T-G---AAAA -TG-AGA-TA TTAGCATAAA TGTTTACTC- TCATTACGCT

         101                                                       150
H_I12_a   GATGACATTG TGCACTCGAG ATCTTGGTAA TCTTT.GGGA AAATTATGAG
M_I12_a   GATGACATTG TGCACTCGAG ATCTTGGTAA TCTTT.GGGA AAATTATGAG
Z_I12_a   GATGACATTG TGCACTCGAG ATCTTGGTAA TCTTT.GGGA AAATTATCAG
T_I12_a   .A.GTTATTG TGCACTCGAG ATCTTGGTAA TCTTTGGGGG AAATTATGAG
S_I12_a   GATGACATTG TGCACTCGAG ATCTTGGTAA TCTTTTGGGG AAATTATGAG
Consensus GATGACATTG TGCACTCGAG ATCTTGGTAA TCTTT-GGG- AAATTATGAG

         151                                                       200
H_I12_a   TAA..TTTT. ...T...AAA AATTTTAA.. AGCAGCAGCA GTTACCATGC
M_I12_a   TAA..TTTT. ...T...AAA AATTTTAA.. AGCAGCAGCA GTTACCATGC
Z_I12_a   TAA..TTTT. ...T...AAA AA.TTT.A.. A.AGCAGCA GTTACCATGC
T_I12_a   TAATTTTTTC C.CTC..TCT TTTTTTTATT TAAAGCAGCA GTTGCCATGC
S_I12_a   TAATTTTTTC CTCTCTTTTT TTTTTTTATT TAAAGCAGCA GTTGCCATGC
Consensus TAA--TTTT- ---T------ --TTTT-A-- ---AGCAGCA GTT-CCATGC

         201                                                       250
H_I12_a   AATTCAGGCT AATTCTGCGT AGGCTCCGAA CGGATATAAT TATCGAGGAG
M_I12_a   AATTCAGGCT AATTCTGCGT AGGCTCCGAA CGGATATAAT TATCGAGGAG
Z_I12_a   AATTCAGGCT AATTTTGCGT AGGCTCCAAA CGGATATAAT TATCGGGGAG
T_I12_a   AATTCAGGCT AATTCTGCGT GGGCTCCACA CGGATATAAT TATCGTTGAG
S_I12_a   AATTCAGGCT GATTCTGCGT GGGCTCCACA CGGATATAAT TATCGTTGAG
Consensus AATTCAG-CT AATTCTGCGT -GGCTCC--A CGGATATAAT TATCG--GAG

         251                                                       300
H_I12_a   TCAAGATGTT ATGCTAAAAA CTATGCATTG ATATTCCCAT TTATTATGTA
M_I12_a   TCAAGATGTT ATGCTAAAAA CTATGCATTG ATATTCCCAT TTATTATGTA
Z_I12_a   TCAAGATGTT ATGCTAAAAA CTATGCATTG ATATTACCAT TTATTATGTA
T_I12_a   TCAAGATGCT AGGCTAAAAA CTATGCATTG ATATTCCCAT TTATTATGTA
S_I12_a   TCAAGATGCT AGGCTAAAAA CTATGCATTG ATATTCCCAT TTATTATGTA
Consensus TCAAGATG-T A-GCTAAAAA CTATGCATTG ATATTCCCAT TTATTATGTA

         301                                                       350
H_I12_a   CATACAACTT TGACAATGTT GATGCTTGAA ATAGCAGGAA ATTGTCTTGC
M_I12_a   CATACAACTT TGACAATGTT GATGCTTGAA ATAGCAGGAA ATTGTCTTGC
Z_I12_a   CATACAACTT TAACATTGTT GATGCTTGAA ATAGCTGGAA ATTGTCTTGC
T_I12_a   CATAGGACTC CATACATTGT GATGCTTGAA ATAGCGGGAA ATTGTCTTGT
S_I12_a   CCTAGGACTC AGACAATGTT GATGCTTGAA ATAGC-GGAA ATTGTCTTGT
Consensus CATA--ACT- TGACAATGTT GATGCTTGAA ATAGC-GGAA ATTGTCTTG-

         351                                                       400
H_I12_a   G.CAAAAATT ACAGTCCATA TTAGGACATC TGAAATTGCG AAGAATTATA
M_I12_a   G.CAAAAATT ACAGTCCATA TTAGGACATC TGAAATTGCG AAGAATTATA
Z_I12_a   G.CAAAAATT ACAGTCCATA TTAGGACATC TAAAATTGCG AAGAATTATA
T_I12_a   GT.AAAAATA ACACCCCATA TTAGGACATC TAAAATTGGG AAGAATTATG
S_I12_a   GTAAAAAATA ACACCCCATA TTAGGACATC T-AAATTGCG CAGAATTATG
Consensus G--AAAAAT- ACA--CCATA TTAGGACATC T-AAATTGCG AAGAATTAT-

         401                                                       450
H_I12_a   TAGTAATTGC AGGCTTTCAG ATGGGAGAGC CAGAATGCTC AAACTAGTGC
M_I12_a   TAGTAATTGC AGGCTTTCAG ATGGGAGAGC CAGAATGCTC AAACTAGTGC
Z_I12_a   TAGTAATTGC AGGCTTTCAG ATCGGGGAGC CAGAATGCTC .AACTAGGAGC
T_I12_a   TAGTAATTGC AGGCTTTCAG ATTTGAAAGC CAGAATGCTC AAACTAGATC
S_I12_a   TAGTAATTGC AGGCTTTCAG ATTTGAAAGC CAGAATGCTT AAACTAGATC
Consensus TAGTAATTGC AGGCTTTCAG AT--GA-AGC CAGAATGCTC AAACTAG--C

         451                                                       500
H_I12_a   AAATGTGGAT AAAATTACAG ATCAGAGCAA AAATTAGGGA AAAAGGGGGT
M_I12_a   AAATGTGGAT AAAATTACAG ATCAGAGCAA AAATTAGGGG AAAAGGGGGT
Z_I12_a   AATTGTGGAT AAAATTACAG ATCAGAGTAT AAATTAGGG. ...ATGGGGT
T_I12_a   AAATGTGGAT GGAATTACAG ATGAGGGTAT AAATTA.GG. AGAGGGG.GT
S_I12_a   AAATGTGGAT AAAATTACAG ATCAGGGTAT AAATTA.GG. ATAGGGGAGT
Consensus AAATGTGGAT AAAATTACAG ATCAG-G-A- AAATTA-GG- A-A-GGG-GT

         501                                                       550
H_I12_a   CTGGGATTTT TCAGGTTGGC TATAGGATTC CGGAAGTGTC TAATGCCACA
M_I12_a   CTGGGATTTT TCAGGTTGGC TATAGGATTC CGGAAGTGTC TAAAGCCACA
Z_I12_a   ATGGCATTTT TCAGGTTGTC TACAGGATTC CGGAAGTGTC TAAAACCACA
T_I12_a   ATAGGCTGTG ACA..TT.TA AAT.CCCCTC GTCGAATG.C .AAAACAGCA
S_I12_a   ATAGGCTATG GCA..TT.TA AAT.CTTCTC GACAATTG.C .AAAACCACA
Consensus -T-GG-T-T- -CA--TT--- -AT-----TC ---AA-TG-C -AAA-CCACA
```

Figure 4 Multiple sequence alignment of I12a in five vertebrate species; M, mouse; H, human, T, *Takifugu rubripes*; S, *Spheroides nephelus*; and Z, zebrafish. The consensus sequence represents identity in four out of five species.

```
                1                                                          50
T_I12_b   CAGCTGCACA TCCAAGAGGG TCAGTGTTAT TTCGCTGTAT TCTCCTCTTG
S_I12_b   CAGCTGCACC TCCAAGAGGG TCAGTG.TA. TT..C..... TCT.C.CTTG
H_I12_b   CAGCTGCAAA CCCAAGAGGG TCAGCATCAT TTCACTGTAT TCTCTTCTTG
M_I12_b   CAGCTGCAAA CCCAAGAGGG TCAGCATCAT TTCACTGTAT TCTCTTCTTG
Z_I12_b   CAGCTGTGGA GCAGAGAGGG TCAGCGTTAT TTCACTCTAT TCTCCTCTTG
Consensus CAGCTGCA-A -CCAAGAGGG TCAG--T-AT TTC-CT-TAT TCTC-TCTTG

                51                                                        100
T_I12_b   ATTATACAAG CTGAGCCCAT CAAAACCACC ATAATTACAG TCATTTCGCC
S_I12_b   ATTATACAAG CTGAGCCCAT CAAAACCACC ATAATTACAG TCATTTCGCC
H_I12_b   A.T.TACAAG CCGGGCCCAT CAAACACAAC ATAATTACAG TAATTTCAGG
M_I12_b   A.T.TACAAG CCGGGCCCAT CAAACACAAC ATAATTACAG TAATTTCAGG
Z_I12_b   A.T.TACGCG CCGGGCCCAT CAAACCAGAC ATAATTACAG TCATTCCCTC
Consensus A-T-TACAAG C-G-GCCCAT CAAA--CA-C ATAATTACAG T-ATTTC---

                101                                                       150
T_I12_b   CTTATTTATT CTAATGCCGT TT.CCTATCT GTGAGGTAAT TATGAGC.A.
S_I12_b   CTTATTTATT CTAATGCCGT TT.CCTATCT CTGAGGTAAT TATGAAC.AT
H_I12_b   TTTATTTATT CTAATGCAGT TTCCCCATCT CTCTGGTAAT TATGAGCAA.
M_I12_b   TTTATTTATT CTAATGCAGT TTCCCCATCT CTCTGGTAAT TATGAGCAA.
Z_I12_b   CTTACTTATT CTAATGCAGC TT.CCCATCT ACGGGATAAT TATGAGCAAT
Consensus -TTATTTATT CTAATGC-GT TT-CC-ATCT -T--GGTAAT TATGAGC-A-

                151                                                       200
T_I12_b   ..TTTTT.GC AAGGAGAATC TTTCTGTGTT GACAAAGGAG ATA..G.TCA
S_I12_b   TTTTTTT.GC AAGGAGAATC TTTCTGTGTT GACAAAGGAG ATA..G.TCT
H_I12_b   .TTTTTTCGC CCAGGGAATC TTTTTGCATT AACAAAAGAG ATAACGCACT
M_I12_b   .TTTTTTCGC CCAGGGAATC TTTTTGCATT AACAAAAGAG ATAACGCACT
Z_I12_b   TTTTTTTCGC ACGTGGAATC .TTTTG..TT AACAAAATGA ATAGCGCGCT
Consensus -TTTTTT-GC ---G-GAATC TTT-TG--TT -ACAAA-GAG ATA--G--CT

                201                                                       250
T_I12_b   GAAAG..AAA TCTGCTGCCG CAAGACAAAA AAAGAAGAGA AAATAGGTGA
S_I12_b   AAGAG..AAA T.T..T.... ..G.C...A ATAGAA.A.A AAATAGGTGA
H_I12_b   GAAAGCCAAA TTTGCTGTGC .A.TTGAGAA AAGGAAAAAA AAA.AAATCA
M_I12_b   GAAAGCCAAA TTTGCTGCGC .A.TTGAGAA AAGG..GAAA AAA.AAATCA
Z_I12_b   GAAAG.AAAA TTTGCTGTGC GACGGGGGAA AAAGAAAAGG GGG.AAAAAA
Consensus GAAAG--AAA T-TGCTG---  -A-----AA AA-GAA-A-A AAA-A--T-A

                251                                                       300
T_I12_b   GCCTCCCTTT AGTCACTG.T TCCTTTAGAC TTTCGCCTGGA CTATAAGGCT
S_I12_b   GCCTCCTTTT CGTCACTG.T TCCTTTAGAC TTAGGCTGGA CTATGAGGCT
H_I12_b   A.AT.AGGTG CG.AGCTGCC ATCTCTGCAA TTCTCTGGTA CCG.GAGCCG
M_I12_b   A.AT.AGGTG CG.AGCTGCC ATCTCTGCAA TTCTCCGGTA CCG.GAGCCG
Z_I12_b   G.AG.A.GAG CG.AAATAGG TGCGCACCGG TAGT.TATAA GTA.AGGCTG
Consensus CG----CTG- --CT-T--A- TT-----G-A C-----AG-C-

                301                                                       350
T_I12_b   TCTAAATTGC TCGCAGGTGT AGTCTCCATC AC.TGTCACT GCGAAGAGTC
S_I12_b   GCTAAATTGC TCGCAGGTGT AGTCTGCATC AC.TGTCACT GCGAAGAGCC
H_I12_b   GC.AAATTGC TTGCAGGTGT ATGGAGCAAG .CTTGTCAAT GG..CCAGGC
M_I12_b   GC.AAATTGC TTGCAGGTGT ATGGAGCAAG .CTTGTCAAT GG..CCGGGC
Z_I12_b   GC.AAATTGC TTGCAGGTGT ATACCGAGAG ATTTGTCAAT GGAAAGAGGG
Consensus GC-AAATTGC T-GCAGGTGT A----GCA-- -C-TGTCA-T G------AG-C

                351                                           393
T_I12_b   TTCGAAATTA GCAAA.GCAC AATGTAAGAG TGATGACAGA CGC
S_I12_b   TTCAAAATTA GCAAA.GCAC AATATGAGAA CGATCGCAGC CAA
H_I12_b   CTCCAAATTA GCAAATGCAC AGCAGCAAAG TAATGA.AGA CAG
M_I12_b   CTCCAAATTA GCAAATGCAC AGCAGCGAAG TAATGA.AGA CAG
Z_I12_b   ATCAAAATGA GCAAAAGCAC AATGTCCGAG TGATGA.AGG CAG
Consensus -TC-AAATTA GCAAA-GCAC A------AG T-ATGA-AG- CA-
```

Figure 5 Multiple sequence alignment of I12b in five vertebrate species. The consensus sequence represents identity in four out of five species. Sequences similar to the binding site for Dlx protein ([A/C/G]TAATT[G/A][C/G]) (Feledy et al. 1999) are shown in bold. See also Figure 7. Abbreviations as in Figure 4.
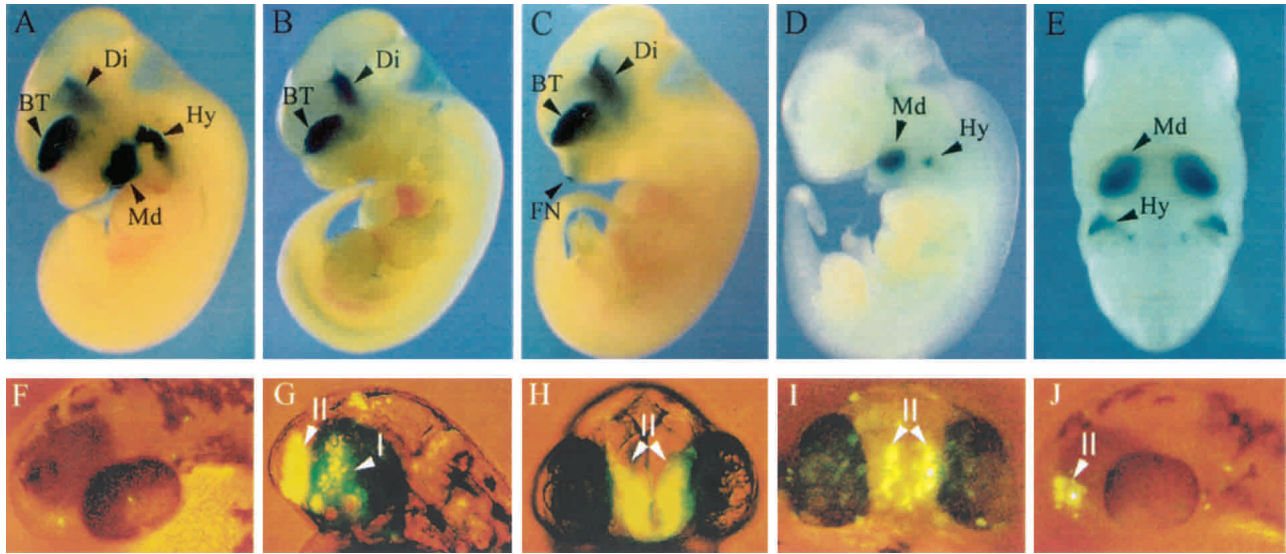
Figure 6 Enhancer activity of conserved *Dlx* intergenic sequences in transgenic mice (*A–E*) and zebrafish (*F–J*). (*A*) Mouse I56i, (*B*) mouse I56ii, and (*C*) mouse I12b each drive reporter gene expression to the telencephalon (BT) and diencephalon (Di) of transgenic mice, as shown here in mouse embryos. (*D, E*) Mouse I12a drives reporter gene expression to a subset of mesenchymal cells in the mandibular (Md) component of the first branchial arch and in the second branchial arch (Hy) of an E11.5 embryo. (*A–D*) are sagittal views and (*E*) is a frontal view of the embryo shown in (*D*). All embryos are at stage E11–12. FN, frontonasal prominence. (*F*) Head of 48 hpf primary transgenic zebrafish embryo, dorso-lateral view, injected with the control *dlx6a-GFP* reporter plasmid. Injection of this construct results in very few GFP-positive cells with no tissue specificity (n > 150). (*G,H*) Lateral and frontal views, respectively, of a 48-hpf zebrafish embryo from a transgenic line produced with a construct made with the *dlx6a-GFP* reporter plasmid that also contained a 1.4-kb *dlx5a/dlx6a* intergenic fragment containing I56i and I56 ii. I and II indicate the diencephalic and telencephalic domains of transgene expression, which also correspond to endogenous *dlx* expression patterns in the zebrafish forebrain. (*I*) Frontal view of a 48-hpf primary transgenic zebrafish embryo injected with a *dlx6a-GFP* that also contained a 4.0-kb mouse *Dlx5/Dlx6* intergenic fragment that comprises I56i. The transgene is expressed predominantly in the telencephalic domain II. (*J*) Lateral view of a 48-hpf primary transgenic zebrafish embryo injected with a *dlx6a-GFP* that also contained a 2.8-kb mouse *Dlx5/Dlx6* intergenic fragment that comprises both I56i and I56ii. GFP-positive cells are seen only in the telencephalic domain, II.

the transgene using PCR). Transgenic analysis of combinations of fragments from the mouse *Dlx1/Dlx2* intergenic region failed to indicate any enhancer activity that could be assessed to sequences outside I12a and I12b. Notably, some of these constructs included I12c (zero out of six PCR-positive embryos) suggesting that this sequence has no enhancer activity by itself, although it cannot be ruled out that it may cooperate with either I12a or I12b in a quantitative manner.

The Three Forebrain Enhancers Show Limited Sequence Similarity

The similar activity of the I12b, I56i, and I56ii enhancers in transgenic mice led us to investigate whether there could be sequence similarities between them. We made pairwise and dot matrix alignments of the three forebrain enhancers in both orientations. We also compared the forebrain enhancers with I12a. We did not find long stretches of sequence similarity among the four enhancers. The best dot matrix alignment was obtained by comparing I12b with I56i (Fig. 7A). A short fragment that extended between 60–80 bp, depending on individual pairwise alignments, was present in all three forebrain enhancers but not in I12a. The two enhancers from the *Dlx5/Dlx6* locus are in opposite orientations in this alignment (shown for the zebrafish sequences in Fig. 7B). The overall similarity over the short region is between 50–60%, thus smaller than the

Table 1. Expression of Reporter Constructs in Primary Transgenic Mouse Embryos and Transgenic Mouse Line

Enhancer element	Primary (P)	Lines (L)	Ventral forebrain	Frontonasal prominence	Apical ectodermal ridge	Visceral arches	Ectopic expression
MI12a	N.D.	7	0L	0L	1L	6L	2L
ZI12a	N.D.	2	0L	0L	0L	1L	0L
MI12b	3	5	3P, 5L	3P, 4L	3P, 5L	1P,[a] 0L	2P, 1L
ZI12b	N.D.	N.D.	N.D.	N.D.	N.D.	N.D.	N.D.
MI56i[c]	13	4	13P, 3L	2P, 1L	1P, 0L	7P, 2L	N.D.
ZI56i[c]	12	N.D.	12P	0P	0P	3P	N.D.
MI56ii	3	2	3P,[b] 2L[b]	1P, 0L	2P, 1L	0P, 0L	1P, 0L
ZI56ii[c]	10	2	1P, 0L	1P, 1L	1P, 0L	0P, 2L	N.D.

All constructs were made by inserting the enhancer fragments into the p1229/1230 plasmids that contain the LacZ reporter gene under the control of a β-globin minimal promoter. M, mouse, Z, zebrafish.

[a]Weak expression in the second arch.
[b]Both primaries and lines showed variable expression patterns in the two domains of the forebrain; thus, 2P showed a weaker expression pattern in the telencephalon (domain II) compared to the diencephalon (domain I), whereas 2L showed expression only in the diencephalon (domain I), and 1P (shown in Fig. 6B) showed an equally strong expression in both domains (I, II).
[c]From Zerucha et al. 2000.

similarity between orthologous enhancer sequences (Figs. 1B, 2B). Interestingly, this region of similarity was also found downstream of the zebrafish *dlx2b* gene, a gene thought to be a duplicate of *dlx2a*, but that is not part of a bigene cluster (A. Amores and M. Ekker, unpubl. observations).

The sequences shown in Figure 7B include a putative *Dlx* binding site, (A/C/G/) TAATT (G/A) (C/G) (Feledy et al. 1999), near both ends of the similarity region. The core binding site for many homeodomain proteins (TAAT/ATTA) was also found between the two putative Dlx binding sites in many of the enhancers (Fig. 7B). The spacing between the Dlx binding sites was similar in all three enhancers. We previously showed that mutagenesis of both Dlx binding sites in I56i abolished almost completely the reporter gene expression in the fore-brain of transgenic mice, suggesting that these sites are essential for activation or maintenance of enhancer activity, possibly through a crossregulatory or autoregulatory mechanism (Zerucha et al. 2000). The Dlx binding sites and surrounding nucleotides are less conserved in I56ii than those in I12b and I56i. The I56ii sequence is not activated by Dlx proteins in transfection assays, contrarily to I56i and I12b (Zerucha et al. 2000; N. Ghanem and M. Ekker, data not shown). This may also explain why it is less efficient than the other two enhancers in targeting a strong and consistent forebrain expression.

We also looked for additional protein-binding sites within the four enhancers (using Genomatix, Matinspector professional software; www.genomatix.de) and could not find any that were consistently found in all of them or in the three forebrain enhancers except for the homeodomain protein-binding sites TAAT/ATTA. Interestingly, the *Dlx* binding site is also a low affinity-binding site (Chen and Schwartz 1995) for members of the *Nkx* family, that are known to be expressed in the forebrain. *Nkx*2.1, for instance, regulates regionalization in a subset of cells in the basal ganglia (Sussel et al. 1999) where the *Dlx* genes are also expressed.

In summary, the similarity between enhancers from paralogous bigene clusters occurs only in a small region of the total enhancer sequence, which, in turn, is highly conserved and over a much longer distance between orthologous, but not paralogous loci.

DISCUSSION

Conserved Organization of the Intergenic Region of Orthologous *Dlx* Bigene Clusters

We have performed a search for homologies in the intergenic region separating the two *Dlx* genes of bigene clusters in five different vertebrate species. Our analysis further illustrates the usefulness of "phylogenetic footprinting" (Muller et al. 2002) to identify *cis*-acting regulatory sequences. Examination of the region that separates the two *Dlx* genes that constitute the *Dlx1/Dlx2* or the *Dlx5/Dlx6* bigene clusters reveals regions of high sequence conservation as well as conserved organization of the intergenic region for orthologous loci of distantly related vertebrates. Each of the two bigene clusters contains two regions of high sequence conservation that extend over a few hundred base pairs as well as a few shorter regions of sequence similarity. For both bi-

Figure 7 Limited similarity between intergenic forebrain enhancer sequences. (*A*) Dot matrix comparison of the zebrafish I12b and I56i. The main two regions of sequence similarity are shown in *B* as multiple sequence alignments between I12b, I56i, and I56ii, and a sequence downstream of the zebrafish *dlx2b*. A three out of four consensus is shown. Putative *Dlx* binding sites, (A/C/G/) TAATT (G/A) (C/G), are indicated in bold, with mismatches highlighted. Additional TAAT/ATTA core homeodomain protein-binding sites are also highlighted.

gene clusters, the relative position and orientation of the conserved intergenic sequences are identical in all five species (Fig. 1, Fig. 2, and deposited sequence data).

The use of compact genomes found in tetraodontid species, such as the two pufferfish *Takifugu rubripes* and *Spheroides nephelus* was initiated to facilitate the search for regulatory elements. This is mainly because large regions of neutral DNA were lost in the course of genome reduction in these species, leaving the noncoding DNA regions enriched for *cis*-acting regulatory elements. We found that the presence of highly conserved sequences in *Dlx* intergenic regions probably contributes to maintain its size even in species with compact genome. Thus, the size of the *Dlx1/Dlx2* and of the *Dlx5/Dlx6* intergenic regions in the two pufferfish, although smaller than their mammalian counterparts, does not follow, proportionally, the smaller size of the genome of the two species.

Orthology assignment for the vertebrate *Dlx* genes was sometimes made difficult by the high degree of sequence similarity in the coding region of *Dlx* genes and by their highly overlapping patterns of expression. Conserved synteny, particularly with the *Hox* clusters, was useful in establishing orthology relationship, as the *Dlx* bigene clusters have been found consistently on the same chromosome as one of the *Hox* clusters (Stock et al. 1996; Amores et al. 1998). Here, we propose that the sequence of the intergenic region is also a reliable predictor of orthology as the paralogous intergenic sequences are quite different while orthologous bigene clusters contain highly conserved sequences.

We examined whether or not the above prediction also applies to a duplicate gene in zebrafish: *dlx2b* (previously, *dlx5*; see comments about nomenclature in Methods). This gene shows high sequence similarity with members of the *Dlx2* and *Dlx5* orthology groups. Mapping of *dlx2b* indicates that it is found in a group of genes with conserved synteny and that are a duplicate of a chromosome region that includes *dlx2* (Amores et al. 1998). We examined about 8 kb of DNA downstream of *dlx2b* and found some sequence similarity with the noncoding sequence elements located in the *Dlx1/Dlx2* intergenic region. Thus, sequences similar to I12a, I12b, and I12c were found (Fig. 7B and supplementary Figs. 6 and 8) although similarity was generally lower than when comparing individual elements between species. No sequence was found that resembled the conserved elements from the *Dlx5/Dlx6* intergenic region except for the short sequence shown in Figure 7B. Thus, in addition to synteny analysis, conservation of noncoding sequence elements can be useful in establishing relationships between duplicate genes.

Highly Conserved *cis*-Acting Regulatory Sequences in the Intergenic Region of *Dlx* Bigene Clusters

The largest conserved sequences found in the *Dlx1/Dlx2* and *Dlx5/Dlx6* intergenic regions are also the only ones conserved in all five species that were examined in the present study. The role of each of these sequences as a *cis*-acting regulatory element is demonstrated by their ability, once coupled to a promoter to drive expression of a reporter transgene in a tissue- and stage-specific manner. Sequence comparisons between mouse and human, or between *Takifugu* and *Spheroides*, reveals an overall high degree of sequence similarity and are therefore of less predictive value in the identification of regulatory elements. This may be because of the small evolutionary distance between the two mammals (~50–60 Mya) as well as the two pufferfish (~5–35 Mya), and to the slow rate of

divergence for neutrally evolving regions among vertebrates in general (0.1% to 0.5% per million years) (Tautz 2000). Intergenic fragments outside the enhancers with 75–80% overall conservation between mouse and human failed to act, by themselves, as enhancers when tested in transgenic mice. Therefore, caution should be exerted when identifying putative *cis*-acting sequences based on comparisons between vertebrates of the same order. Comparisons that include multiple species with some that are distantly related might be a more efficient approach to identify noncoding sequence elements of functional importance, while keeping in mind that absence of sequence conservation does not necessarily indicate absence of functional conservation (Flint et al. 2001).

The relatively high degree of sequence conservation between the mouse and human *Dlx1/Dlx2* intergenic region (80%) or *Dlx5/Dlx6* intergenic region (78%) contrasts with the *Dlx3/Dlx7* intergenic region that is only 69% identical, overall, between the two species (Sumiyama et al. 2002) despite the presence of sequences with higher percentage identity that may have a regulatory function (Sumiyama et al. 2002). However, comparisons of the mammalian *Dlx3/Dlx7* intergenic region with those of zebrafish (Sumiyama et al. 2002), or *Takifugu rubripes* (N. Ghanem and M. Ekker, unpubl. observations) did not show conserved sequences comparable in length or percent identity to the four enhancers that we identified in the *Dlx1/Dlx2* or in the *Dlx5/Dlx6* bigene clusters. Therefore, the *Dlx3/Dlx7* bigene cluster may differ from its two paralogous *Dlx* clusters by a relatively low importance of the intergenic region in the mechanisms that control gene expression or by a higher divergence in regulation mechanisms between the different vertebrate lineages. Consistent with this latter hypothesis is the observation that zebrafish *dlx3/dlx7* have marked differences in their early patterns of expression compared to their mammalian orthologs (Quint et al. 2000).

Function of Intergenic Elements in *Dlx* Regulation and Evolution

The organization of *distal-less*-related genes in bigene clusters may have preceded the evolution of vertebrates as two of the three characterized *Dll* genes of the ascidian *Ciona intestinalis*, *Dll-A*, and *Dll-B* are organized similarly with a short intergenic region (Di Gregorio et al. 1995). Recently, an enhancer located upstream of *Dll-A* was identified and shown to recapitulate most aspects of the endogenous expression pattern (Harafuji et al. 2002). Enhancers have yet to be found in the intergenic region that separates the *Ciona Dll-A* and *Dll-B* genes and preliminary sequence comparisons did not reveal similarities in sequence between this region and the four *cis*-acting regulatory sequence found in vertebrate *Dlx* genes (M. Ekker, unpubl. observations).

Although the three *Dlx* bigene clusters of vertebrates are likely the result of duplication of an ancestral bigene cluster, we did not observe a high degree of conservation between paralogs, regardless of the species. This extends the observation previously made by Sumiyama and collaborators who compared the three human bigene clusters (Sumiyama et al. 2002). This lack of sequence similarity between paralogs is surprising, considering the similarities in expression patterns of genes found in paralogous bigene clusters.

Enhancers with overlapping patterns of activity (Fig. 6) show only a limited conservation in sequence (Fig. 7) that contrasts sharply with the high degree of conservation between orthologous sequences. Furthermore, enhancer se-

quences found in one *Dlx* bigene cluster are not found in the two paralogous clusters. Although one or several *Dlx* intergenic enhancers could originate from a sequence found in the ancestral *Dlx* bigene cluster, they would have diverged following the duplication events that took place early in vertebrate evolution, and that led to the three *Dlx* bigene clusters of modern vertebrates. This divergence happened before the separation of the lineages leading to modern-day teleost and tetrapods. Since then, purifying selection maintained most, if not all, regulatory mechanisms that involve these intergenic sequences, at least for the *Dlx1/Dlx2* and *Dlx5/Dlx6* bigene clusters. The region of limited similarity found between the three forebrain enhancers may suggest that they resulted from a tandem duplication (I56i and I56ii) that also predated the split between the ray-finned fish lineages, and/or represent what subsists from a sequence present in the ancestral *Dlx* bigene cluster.

Although the current study suggests that *cis*-acting regulatory elements of diverse sequence may exert similar enhancer function, the converse may also be true. Thus, I56i from mouse targets expression of a reporter transgene to the forebrain and mesenchymal cells of the branchial arches (Fig. 6A) whereas the orthologous sequence from zebrafish only directs expression to the forebrain, in either transgenic mice or zebrafish (Zerucha et al. 2000) despite the fact that the two sequences are >80% identical (Fig. 2B). Thus, the small differences in sequence between the enhancers from the two species may have a profound effect on enhancer function.

Evidence has been previously presented for cross-regulatory interactions between *Dlx* genes. Thus, the *Dlx1* and *Dlx2* genes are expressed earlier in the forebrain and are involved in either the activation or maintenance of *Dlx5* and *Dlx6* expression through the enhancer(s) found in the *Dlx5/Dlx6* intergenic region (Zerucha et al. 2000). In contrast, there is, at present, no evidence that *Dlx5/6* regulate *Dlx1/2* in the brain. In the branchial arch mesenchyme, *Dlx5/6* regulate *Dlx3*, but not *Dlx1/2* (Depew et al. 2002). Thus, the divergence of the intergenic enhancer sequences may have contributed to the specificity of cross-regulation between *Dlx* genes, allowing for sequential expression of paralogs.

The present study indicates an important role for the intergenic region in the *cis* regulatory mechanisms that are responsible for many aspects of the expression of genes from two *Dlx* bigene clusters. Intergenic regulatory elements are not solely responsible for *Dlx* regulation. Thus, a fragment of the 5′-flanking region of mouse *Dlx2* was shown to recapitulate expression in the epithelial cells of the branchial arches (Thomas et al. 2000). A targeted mutation, that inactivates the function of the mouse *Dlx1* and *Dlx2*, eliminates the entire intergenic region (Anderson et al. 1997). Intriguingly, homozygous mutants expressed truncated *Dlx1* transcripts in the forebrain despite the absence of the I12b sequence (Zerucha et al. 2000). Although our results indicate that I12b is sufficient to confer expression of a reporter transgene to the forebrain (Fig. 6C), distinct sequences located upstream of *Dlx1* also share this property (N. Ghanem and M. Ekker, unpubl. observations), suggesting a cooperative or synergistic effect between multiple and distinct enhancers in forebrain regulation of *Dlx1* and/or *Dlx2*. Distinct mechanisms may take place at the *Dlx5/Dlx6* locus. The *lacZ* reporter gene, introduced in a targeted mutation of *Dlx5/Dlx6* that also removes the intergenic sequence (including I56i and I56ii), is only weakly expressed in the forebrain (Robledo et al. 2002). This suggests that enhancers outside the intergenic region may exist but

that the intergenic enhancers play an essential role in conferring proper levels of gene expression, in as much as detection of transcripts by in situ hybridization can be considered quantitative. Taken together, these observations suggest complex mechanisms of *Dlx* expression control. These mechanisms involve multiple enhancers with overlapping but not necessarily redundant activity and a high degree of conservation in distant vertebrates for at least some of these enhancers.

METHODS

Dlx Gene Nomenclature

To help standardize the nomenclature for vertebrate *Dlx* genes, we found it useful to adopt what was recently suggested by Panganiban and Rubenstein (2002). As the *Dlx* genes are found in regions of conserved synteny that contain the *Hox* clusters, the new nomenclature is aligned with that of the zebrafish *hox* clusters (Amores et al. 1998). Thus, the zebrafish gene we refer to as *dlx5a* in this study is the gene previously named *dlx4* (Akimenko et al. 1994). Similarly, the zebrafish gene previously named *dlx5* is renamed *dlx2b*, as it is a *dlx2* duplicate (see Discussion). The previous *dlx1*, *dlx2*, and *dlx6* genes are renamed *dlx1a*, *dlx2a*, and *dlx6a*, respectively. Finally, the previous *dlx3*, *dlx7*, and *dlx8* genes of zebrafish would be renamed *dlx3b*, *dlx4b*, and *dlx4a*, respectively. We kept the *Dlx3/Dlx7* nomenclature for the mouse genes throughout the current report for the sake of simplicity but indicated the suggested name change.

Isolation and Characterization of *Dlx* Genes From *Spheroides Nephelus*

Clones from a PAC library (Amemiya et al. 2001) were screened using a PCR approach for a conserved region of *Dlx* genes (Stock et al. 1996). The PCR fragments were sequenced to establish a preliminary orthology assignment. Genomic fragments comprising intron B and exon 3 of positive *Dlx* clones plus the intergenic region between *Dlx* genes were obtained by PCR amplification using either specific or degenerate oligonucleotides.

Sequence Analysis

The zebrafish, mouse, and *Spheroides* intergenic sequences were determined from previously isolated genomic clones (McGuinness et al. 1996; Ellies et al. 1997; Depew et al. 1999) or from the *Spheroides* clones described in the above paragraph. They are deposited in GenBank under accession nos. AY168007–AY168012. The sequences from human and *Takifugu rubripes* were obtained from public databases: Human *Dlx1/Dlx2*, GenBank accession no. NT_005332.9; Human *Dlx5/Dlx6*, GenBank accession no. NT_033964.1; *Takifugu dlx1/dlx2*, scaffold 21, position 120318 to 125668, *Takifugu dlx5/dlx6*, scaffold 3932, position 6627–10192. For the Fugu Genome Consortium/JGI (DOE Joint Genome Institute), see http://www.jgi.doe.gov/index.html.

Pairwise sequence alignments are performed with PIPMAKER (available at http://bio.cse.psu.edu/pipmaker/), or with the BestFit, and Mapplot programs of the GCG Wisconsin package. Multiple sequence alignments are performed with the Pileup and Clustal X programs.

Transgenic Animals

For transgenic mice, sequences from the *Dlx* intergenic regions were subcloned into the p1229/p1230 vectors (Yee and Rigby 1993) that contain a human β-globin minimal promoter and the *lacZ* reporter gene. For transgenic zebrafish, intergenic enhancer sequences were inserted into a plasmid containing the GFP reporter gene placed downstream of a 3.5-kb fragment from the immediate 5′-flanking region of zebrafish *dlx6a*, including part of the 5′UTR. This fragment by

itself, does not produce any tissue-specific expression in transgenic zebrafish (Fig. 6F). Subclonings were done using either a PCR-based approach or using convenient restriction sites. Transgenic animals were produced and analyzed as previously described (Zerucha et al. 2000).

ACKNOWLEDGMENTS

REFERENCES

Acampora, D., Merlo, G.R., Paleari, L., Zeraga, B., Postiglione, M.P., Mantero, S., Bober, E., Barbieri, O., Simeone, A., and Levi, G. 1999. Craniofacial, vestibular and bone defects in mice lacking the *Distal-less*-related gene *Dlx5*. *Development* **126:** 3795–3809.

Akimenko, M.-A., Ekker, M., Wegner, J., Lin, W., and Westerfield, M. 1994. Combinatorial expression of three zebrafish genes related to *distal-less*: part of a homeobox gene code for the head. *J. Neurosci.* **14:** 3475–3486.

Amemiya, C.T., Amores, A., Ota, T., Mueller, G., Garrity, D., Postlethwait, J.H., and Litman, G.W. 2001. Generation of a P1 artificial chromosome library of the Southern pufferfish. *Gene* **272:** 283–289.

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.

Anderson, S.A., Qiu, M., Bulfone, A., Eisenstat, D.D., Meneses, J., Pedersen, R., and Rubenstein, J.L.R. 1997. Mutations of the homeobox genes *Dlx-1* and *Dlx-2* disrupt the striatal subventricular zone and differentiation of late-born striatal neurons. *Neuron* **19:** 27–37.

Caracciolo, A., di Gregorio, A., Aniello, F., Di Lauro, R., and Branno, M. 2000. Identification and developmental expression of three *distal-less* homeobox containing genes in the ascidian *Ciona intestinalis*. *Mech. Dev.* **99:** 173–176.

Chen, C.Y. and Schwartz, R.J. 1995. Identification of novel DNA binding targets and regulatory domains of a murine tinman homeodomain factor, nkx-2.5. *J. Biol. Chem.* **270:** 15628–15633.

Depew, M.J., Liu, J.K., Long, J.E., Presley, R., Meneses, J.J., Pedersen, R.A., and Rubenstein, J.L.R. 1999. *Dlx5* regulates regional development of the branchial arches and sensory capsules. *Development* **126:** 3831–3846.

Depew, M.J., Lufkin, T., and Rubenstein, J.L.R. 2002. Specification of jaw subdivisions by *Dlx* genes. *Science* **298:** 381–385.

Di Gregorio, A., Spagnuolo, A., Ristoratore, F., Pischetola, M., Aniello, F., Branno, M., Cariello, L., and Di Lauro, R. 1995. Cloning of ascidian homeobox genes provides evidence for a primordial chordate cluster. *Gene* **156:** 253–257.

Ellies, D.L., Stock, D.W., Hatch, G., Giroux, G., Weiss, K.M., and Ekker, M. 1997. Relationship between the genomic organization and the overlapping embryonic expression patterns of the zebrafish *dlx* genes. *Genomics* **45:** 580–590.

Feledy, J.A., Morasso, M.I., Jang, S.-I., and Sargent, T.D. 1999. Transcriptional activation by the homeodomain protein *distal-less* 3. *Nucl. Acids Res.* **27:** 764–770.

Flint, J., Tufarelli, C., Peden, J., Clark, K., Daniels, R.J., Hardison, R., Miller, W., Philipsen, S., Tan-Un, K.C., McMorrow, T., et al. 2001. Comparative genome analysis delimits a chromosomal domain and identifies key regulatory elements in the α globin cluster. *Hum. Mol. Genet.* **10:** 371–382.

Harafuji, N., Keys, D.N., and Levine, M. 2002. Genome-wide identification of tissue-specific enhancers in the *Ciona* tadpole. *Proc. Natl. Acad. Sci.* **99:** 6802–6805.

Liu, J.K., Ghattas, I., Liu, S., Chen, S., and Rubenstein, J.L.R. 1997. *Dlx* genes encode DNA-binding proteins that are expressed in an overlapping and sequential pattern during basal ganglia differentiation. *Dev. Dyn.* **210:** 498–512.

McGuinness, T., Porteus, M.H., Smiga, S., Bulfone, A., Kingsley, C., Qiu, M., Liu, J.K., Long, J.E., Xu, D., and Rubenstein, J.L.R. 1996. Sequence, organization, and transcription of the *Dlx-1* and *Dlx-2* locus. *Genomics* **35:** 473–485.

Muller, F., Blader, P., and Strahle, U. 2002. Search for enhancers: Teleost models in comparative genomic and transgenic analysis of *cis* regulatory elements. *BioEssays* **24:** 564–572.

Nakamura, S., Stock, D.W., Wynder, K.L., Bollekens, J.A., Takeshita, K., Nagai, B.M., Chiba, S., Kitamura, T., Freeland, T.M., Zhao, Z., et al. 1996. Genomic analysis of a new mammalian distal-less gene: *Dlx7*. *Genomics* **38:** 314–324.

Panganiban, G. and Rubenstein, J.L.R. 2002. Developmental functions of the *Distal-less/Dlx* homeobox genes. *Development* **129:** 4371–4386.

Qiu, M., Bulfone, A., Martinez, S., Meneses, J.J., Shimamura, K., Pedersen, R.A., and Rubenstein, J.L.R. 1995. Null mutations of *Dlx-2* results in abnormal morphogenesis of proximal first and second branchial arch derivatives and abnormal differentiation in the forebrain. *Genes Dev.* **9:** 2523–2538.

Qiu, M., Bulfone, A., Ghattas, I., Meneses, J.J., Christensen, L., Sharpe, P.T., Presley, R., Pedersen, R.A., and Rubenstein, J.L.R. 1997. Role of the *Dlx* homeobox genes in proximodistal patterning of the branchial arches: Mutations of *Dlx-1*, *Dlx-2*, and *Dlx-1* and *-2* alter morphogenesis of proximal skeletal and soft tissue structures derived from the first and second arches. *Dev. Biol.* **185:** 165–184.

Quint, E., Zerucha, T., and Ekker, M. 2000. Differential expression of orthologous *Dlx* genes in zebrafish and mice: Implications for the evolution of the *Dlx* homeobox gene family. *J. Exp. Zool. (Mol. Dev. Evol.)* **288:** 235–241.

Robledo, R.F., Rajan, L., and Lufkin, T. 2002. The *Dlx5* and *Dlx6* homeobox genes are essential for craniofacial, axial, and appendicular skeletal development. *Genes Dev.* **16:** 1089–1101.

Simeone, A., Acampora, D., Pannese, M., Desposito, M., Stornaiuolo, A., Gulisano, M., Mallamaci, A., Kastury, K., Druck, T., Huebner, K., et al. 1994. Cloning and characterization of two members of the vertebrate *Dlx* gene family. *Proc. Natl. Acad. Sci.* **91:** 2250–2254.

Stock, D.W., Ellies, D.L., Zhao, Z., Ekker, M., Ruddle, F.H., and Weiss, K.M. 1996. The evolution of the vertebrate *Dlx* gene family. *Proc. Natl. Acad. Sci.* **93:** 10858–10863.

Sumiyama, K., Irvine, S.Q., Stock, D.W., Weiss, K.M., Kawasaki, K., Shimuzu, N., Shashikant, C.S., Miller, W., and Ruddle, F.H. 2002. Genomic structure and functional control of the *Dlx3–7* bigene cluster. *Proc. Natl. Acad. Sci.* **99:** 780–785.

Sussel, L., Marin, O., Kimura, S., and Rubenstein, J.L.R. 1999. Loss of *Nkx2.1* homeobox gene function results in a ventral to dorsal molecular respecification within the basal telencephalon: Evidence for a transformation of the palladium into the striatum. *Development* **126:** 3359–3370.

Tautz, D. 2000. Evolution of transcriptional regulation. *Curr. Opin. Genet. Dev.* **10:** 575–579.

Thomas, B.L., Liu, J.K., Rubenstein, J.L.R., and Sharpe, P.T. 2000. Independent regulation of *Dlx2* expression in the epithelium and mesenchyme of the first branchial arch. *Development* **127:** 217–224.

Yee, S.-P. and Rigby, P.W.J. 1993. The regulation of myogenin gene expression during the embryonic development of the mouse. *Genes Dev.* **7:** 1277–1289.

Zerucha, T. and Ekker, M. 2000. *Distal-less*-related homeobox genes of vertebrates: Evolution, function, and regulation. *Biochem. Cell Biol.* **78:** 593–601.

Zerucha, T., Stuhmer, T., Hatch, G., Park, B.K., Long, Q., Yu, G., Gambarotta, A., Schultz, J.R., Rubenstein, J.L.R., and Ekker, M. 2000. A highly conserved enhancer in the *Dlx5/Dlx6* intergenic region is the site of cross-regulatory interactions between *Dlx* genes in the embryonic forebrain. *J. Neurosci.* **20:** 709–721.

WEB SITE REFERENCES

http://www.jgi.doe.gov/index.html; Department of Energy Joint Genome Institute. Genomic resources for *Takifugu rubripes*, *Ciona intestinalis*, and other species.

http://bio.cse.psu.edu/pipmaker/; Pipmaker computes alignments of similar regions in two DNA sequences.

www.genomatix.de; software and services including the MatInspector program to search for transcription factor binding sites.