J Mol Evol (2010) 71:427–436
DOI 10.1007/s00239-010-9396-1

Evolution of Conserved Non-Coding Sequences Within the Vertebrate Hox Clusters Through the Two-Round Whole Genome Duplications Revealed by Phylogenetic Footprinting Analysis

Masatoshi Matsunami · Kenta Sumiyama · Naruya Saitou

Received: 27 July 2010 / Accepted: 17 September 2010 / Published online: 28 October 2010

Abstract As a result of two-round whole genome duplications, four or more paralogous Hox clusters exist in vertebrate genomes. The paralogous genes in the Hox clusters show similar expression patterns, implying shared regulatory mechanisms for expression of these genes. Previous studies partly revealed the expression mechanisms of Hox genes. However, *cis*-regulatory elements that control these paralogous gene expression are still poorly understood. Toward solving this problem, the authors searched conserved non-coding sequences (CNSs), which are candidates of *cis*-regulatory elements. When comparing orthologous Hox clusters of 19 vertebrate species, 208 intergenic conserved regions were found. The authors then searched for CNSs that were conserved not only between orthologous clusters but also among the four paralogous Hox clusters. The authors found three regions that are conserved among all the four clusters and eight regions that are conserved between intergenic regions of two paralogous Hox clusters. In total, 28 CNSs were identified in the paralogous Hox clusters, and nine of them were newly found in this study. One of these novel regions bears a RARE motif. These CNSs are candidates for gene expression regulatory regions among paralogous Hox clusters. The authors also compared vertebrate CNSs with amphioxus CNSs within the Hox cluster, and found that two CNSs in the HoxA and HoxB clusters retain homology with amphioxus CNSs through the two-round whole genome duplications.

Keywords Hox cluster · Two-round whole genome duplications · Conserved non-coding sequences · Phylogenetic footprinting · Retinoic acid response element · Amphioxus

Introduction

Vertebrate genomes show evidence of widespread gene duplications compared to invertebrate genomes. Ohno (1970) proposed the existence of two-round whole genome duplications (2R WGD) during the early vertebrate evolution, now known as the 2R hypothesis. Before the amphioxus genome was reported by Putnam et al. (2008), this hypothesis was extensively debated (e.g., Holland et al. 1994; Gibson and Spring 2000; Hughes et al. 2001). Genome duplications generated paralogous genes and complex gene regulatory mechanisms in vertebrate evolution (e.g., Dehal and Boore 2005). These paralogous genes often share the same expression patterns, but some may acquire new expression patterns. The changes of gene expression are mainly because of changes in *cis*-regulatory elements (Carroll 2001).

Identifying the *cis*-regulatory sequences that control spatial and temporal gene expression is a challenging issue. Because gene regulatory elements are expected to be conserved because of their functional importance, searching for evolutionarily conserved non-coding sequences (CNSs) would be an effective strategy for finding candidates of

Electronic supplementary material The online version of this article (doi:10.1007/s00239-010-9396-1) contains supplementary material, which is available to authorized users.

M. Matsunami · K. Sumiyama · N. Saitou
Department of Genetics, School of Life Science, The Graduate University for Advanced Studies (SOKENDAI), Mishima 411-8540, Japan

M. Matsunami · K. Sumiyama · N. Saitou (✉)
Division of Population Genetics, National Institute of Genetics, Yata 1111, Mishima 411-8540, Japan
e-mail: saitounr@lab.nig.ac.jp

functional elements. Let us note that the gene regulatory elements which are not conserved are very rare (Werauch and Hughes 2010). Previous studies have already shown that CNSs are abundant in vertebrate genomes (Bejerano et al. 2004; Woolfe et al. 2005). Genome-wide comparative approaches have also reported the existence of paralogous CNSs (Bejerano et al. 2004; Woolfe et al. 2005; McEwen et al. 2006), and most of them are located in paralogous gene clusters that code for transcriptional factors. These results imply that paralogous CNSs contribute to cluster organization and/or their neighboring gene expression patterns. The authors therefore focused on the vertebrate Hox clusters because they contain abundant CNSs.

The Hox genes orchestrate the development of animal body plans. They consist of more than four physically linked clusters in different chromosomes in vertebrate genomes (Pearson et al. 2005; Lemons and McGinnis 2006). Hox genes of each cluster are expressed along the anterior–posterior body axis in the same order as lining up on the chromosome, a feature called "colinearity" (Garcia-Fernàndez 2005). Paralogous genes of the Hox clusters show the similar expression pattern, which suggests that there might be shared gene expression regulatory mechanisms among paralogous Hox clusters.

The duplication of Hox clusters influences cluster architecture and patterns of non-coding sequence evolution. The duplicated non-coding regions within the Hox clusters are mainly studied for teleost fish (e.g., Chiu et al. 2002; Santini et al. 2003; Prohaska et al. 2004). The third round whole genome duplication occurred after the 2R WGD in the teleost lineage. Chiu et al. (2002) and Prohaska et al. (2004) found massive loss of sequence conservation in teleost HoxA cluster non-coding regions after the 3R WGD. Therefore, teleosts are not suitable for analyzing duplicated Hox cluster non-coding sequences.

In the case of 2R WGD, Kim et al. (2000) described one paralogous CNS within the four Hox clusters. However, analysis of non-coding sequences of the Hox clusters within vertebrates, especially mammalian species, is not sufficient. There are probably two reasons for this. First, the functional paralogous conservation cannot be detected easily. This is because the 2R WGD were very ancient events which occurred approximately half a billion years ago and the non-coding sequences experienced higher evolutionary rates compared to protein coding sequences. This is probably because *cis*-regulatory elements are redundant and may be changed by binding site turnover (Hancock et al. 1999). Second, only a few invertebrate sequences that are more closely related to vertebrates and that still retain cluster structure are available. With the recent abundance in vertebrate genomes sequences, the authors can now analyze the evolution of non-coding sequences within the Hox clusters after 2R WGD. However, identifying CNSs within Hox clusters before 2R WGD remains a challenge.

Recently, Hox cluster sequences of two different amphioxus species, *Branchiostoma floridae* and *B. lanceolatum* were reported by Amemiya et al. (2008) and Pascual-Anaya et al. (2008), respectively. Because amphioxus is the chordate bearing a syntenic Hox cluster which is most closely related to vertebrates, these data would be very informative for inferring the evolution of non-coding regions within Hox clusters before 2R WGD.

Detection of the functional turnover of transcription factor binding site (TFBS) is one interesting problem. In the *Drosophila* genome, the TFBS turnover frequently occurred (Ludwig et al. 2005). Ray et al. (2008) developed a program to find the functional turnover motifs using experimental results as training data. Some *cis*-regulatory regions showed the TFBS turnovers also in vertebrates (Werauch and Hughes 2010). However, these data are difficult to utilize for finding other functional turnover events for various reasons such as insufficient experimental data, short alignment length, and low mutation rate. Therefore, the authors did not examine the functional turnover of the TFBS in this study.

In this study, the authors identified orthologous CNSs within the vertebrate Hox clusters, and found conserved loci among paralogous Hox clusters. The authors compared these CNSs with amphioxus–human CNSs reported by Pascual-Anaya et al. (2008) using phylogenetic footprinting to find CNSs that can be dated back to amphioxus. This study identified and mapped vertebrate CNSs within the four vertebrate Hox clusters using comprehensive genome comparisons.

Materials and Methods

Identification of Vertebrate Hox CNSs

Genomic sequences of Hox clusters were obtained for the following 18 vertebrate species from UCSC Genome Bioinformatics (http://genome.ucsc.edu/): Human (*Homo sapiens*), mouse (*Mus musculus*), chimpanzee (*Pan troglodytes*), orangutan (*Pongo pygmaeus abelii*), rhesus macaque (*Macaca mulatta*), marmoset (*Callithrix jacchus*), rat (*Rattus norvegicus*), guinea pig (*Cavia porcellus*), cat (*Felis catus*), dog (*Canis familiaris*), horse (*Equus caballus*), cow (*Bos taurus*), opossum (*Monodelphis domestica*), platypus (*Ornithorhynchus anatinus*), chicken (*Gallus gallus*), zebra finch (*Taeniopygia guttata*), lizard (*Anolis carolinensis*), and frog (*Xenopus tropicalis*). Partial sequences of the horn shark (*Heterodontus francisci*) that included Hox clusters (DDBJ/EMBL/GenBank accession numbers are AF224262 and

AF224263) were also used for this study. The authors excluded teleost fishes, which have undergone the additional genome duplication in their lineages. Protein coding regions were filtered based on the RefSeq project (http://www.ncbi.nlm.nih.gov/RefSeq/) annotation. Alternative exons were not considered in this analysis. BLAST homology search (Altschul et al. 1997) was performed on this data set with default parameter setting and cutoff scores of >200.

Orthologous CNSs were systematically named based on their genomic locations and BLAST scores. For example, the CNS that is located at the intergenic region between HoxA7 and HoxA6 with the highest BLAST score was named "A76-1."

These CNSs were aligned using CLASTALW (Thompson et al. 1994), and divided into three categories to investigate the depth of conservation: placental mammals, amniotes, and vertebrates. The authors then searched for conserved sequences that were conserved not only between orthologous clusters but also among paralogous four Hox clusters using BLAST search with the cutoff score of less than 30. Annotations of TFBS motifs were mainly based on the TRANSFAC database (http://www.biobase-international.com/pages/index.php?id=transfac).

Analysis of Paralogous CNSs

To investigate the non-coding transcribed regions of the Hox clusters, transcriptional information of mRNAs and ESTs within the human and mouse Hox clusters were obtained from the UCSC Genome Bioinformatics database, and these transcripts were mapped on the region.

Phylogenetic footprinting analysis was carried out for each orthologous CNSs that also have paralogous conservation. Each vertebrate CNS was aligned using CLASTALW. The substitution number of each aligned site was estimated parsimoniously using Fitch's (1971) algorithm. The guide phylogenetic tree (Supplementary Fig. 1) necessary for this analysis was taken from Murphy et al. (2004). In parallel, the likelihood estimation of ancestral sequence of each vertebrate CNS was inferred using PAML 4 (Yang 2007).

Comparison with Amphioxus Hox CNSs

Pascual-Anaya et al. (2008) compared Hox clusters of two different amphioxus species (*Branchistoma floridae* and *B. lanceolatum*) to each human Hox cluster and defined 75 human–amphioxus CNSs (amphiCNS). These amphiCNSs were obtained and were sorted by identity. The authors named amphiCNSs by the order of their identity. The amphiCNSs were compared with vertebrate CNSs to identify significant conserved region among chordates.

Results

Orthologous CNSs Within Vertebrate Hox Clusters

The authors defined 208 CNSs in total: 64, 43, 60, and 41 for HoxA, B, C, and D clusters, respectively. Genomic locations of these CNSs are graphically shown in Fig. 1, and detailed information of all these CNSs is shown in Supplementary Table 1. Many of these orthologous CNSs overlap microRNAs and *cis*-regulatory elements which have previously been described (Mainguy et al. 2003; Yekta et al. 2004). Because sequence information is not complete or homologous sequence is lacking, some CNSs were not found in several species (see Supplementary Table 2). As an example of a *cis*-regulatory element, C98-1 corresponds to the HoxC8 early enhancer which is necessary for proper HoxC8 expression (Juan and Ruddle 2003). Other CNSs might bear similar enhancer functions.

The findings of this study are consistent with previous observations (Prohaska et al. 2004; Chiu et al. 2002), confirming that orthologous CNSs were detected effectively. Moreover, using our criteria, the authors also detected 160 new CNSs (see Fig. 1 and Supplementary Table 1).

The authors detected a larger number of CNSs in Hox5-Hox3 (corresponding to *Drosophila Antp* and *Ubx/abdA*) intergenic sequences than in other intergenic sequences (Fig. 1). This region has abundant alternatively spliced coding RNAs and long non-coding RNAs (Mainguy et al. 2007). This observation suggests that functionally unknown CNSs in this region contribute to these alternative splicing events. In contrast, posterior regions of Hox clusters have poor conservation except for upstream regions of Evx1 and Evx2.

The 208 CNSs were divided into six categories: placental mammals, placental mammals + marsupials, placental mammals + marsupials + monotremes, amniotes, tetrapods, and vertebrates, based on the depth of conservation (Table 1). The level of conservation of orthologous CNSs varies among the four Hox clusters; HoxA has the highest number (64) of CNSs in total, while HoxD has the smallest number (41) of CNSs due to the small numbers of CNSs conserved among amniotes and tetrapods. The HoxC cluster has the highest number (33) of CNSs conserved only among placental mammals, while the HoxB cluster has the highest number (11) of CNSs in placental mammals + marsupials. This result, however, does not mean that the HoxC cluster is the least conserved (see Discussion).

Paralogous CNSs Among Hox Clusters

The authors found 28 paralogous conserved elements in total (8, 6, 6, and 8 for Hox A, B, C, and D clusters,

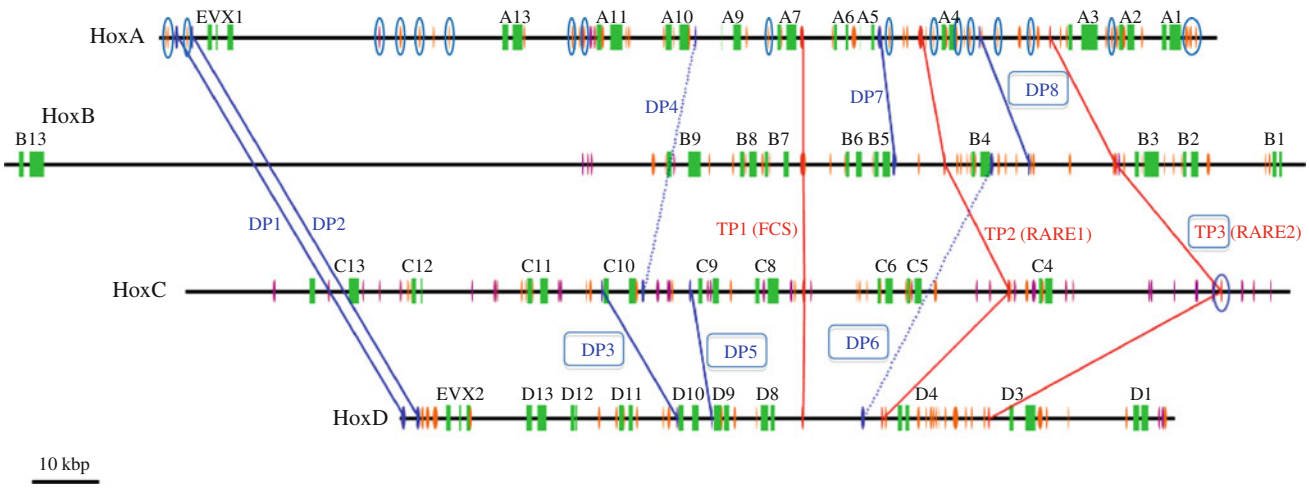

Fig. 1 The schematic diagram of orthologous CNSs and paralogous CNSs among human Hox clusters. Exons of protein coding genes are represented by light *green* boxes. The *orange* ovals are orthologous CNSs. The *blue* and *red* ovals indicate locations of paralogous CNSs conserved also among the two clusters and the four clusters, respectively. The *blue* dotted lines show either microRNA (DP4) or non-syntenic DP (DP6). The paralogous CNSs whose name enclosed by *blue* rectangle is newly detected. Especially the newly detected HoxC CNS of TP3 is highlighted by *blue* circle. The light *blue* circled HoxA CNSs were not identified by Prohaska et al. (2004). *TP* tetra-paralog, *DP* di-paralog (Color figure online)

Table 1 Conservation depth of each CNS

	HoxA	HoxB	HoxC	HoxD
Placental mammals	4	5	33	2
Above + Marsupials	2	11	0	1
Above + Monotremes	8	6	3	6
Amniotes	17	7	8	7
Tetrapods	17	14	16	9
Vertebrates	16	–	–	16
Total	64 (48)	43 (43)	60 (60)	41 (25)

Note: There is no genomic sequence data for horn shark HoxB and HoxC clusters, so "Vertebrates" depth CNSs are not determined, as shown with hyphens. As a result of this, values in parentheses in "Total" are those excluding CNSs shared in all vertebrates

respectively). Three quartets of CNSs are conserved among all the four Hox clusters, and the authors named them TP (tetra-paralogous), as shown in Fig. 1. The authors carried out the phylogenetic footprinting analysis to infer significantly conserved motifs among these three TPs. The authors found the highly conserved region in each CNS, and these overlap with paralogous conserved regions (see Supplementary Fig. 2). Multiple sequence alignments of three TP CNSs are shown in Fig. 2. It should be noted that these sequences are reconstructed ancestral ones. TP2 and TP3 contain retinoic acid response elements (RAREs). Intergenic regions of upstream or downstream of Hox4 genes are abundant with functional RAREs (Mainguy et al. 2003). Despite this, RAREs located downstream of HoxC4 has not been reported before. The authors found a new evolutionarily highly conserved sequence containing RARE in this region. These motifs might maintain gene expression pattern of clusters cooperatively.

The remaining TP1 was discovered by Kim et al. (2000), and they named it four cluster sequence (FCS). Though the authors found conserved motifs in the FCS (Fig. 2), these motifs have no experimental corroboration. Then the authors mapped transcripts within Hox clusters. As a result, 136 CNSs overlap with transcribed regions (see Supplementary Table 1). FCS corresponds to the bidirectional transcript start sites (TSS) which encode alternative spliced RNAs of Hox genes and antisense non-coding RNAs (Fig. 3a). These CNSs might play important roles in the colinear expression pattern of the Hox cluster. Another paralogous CNS between Hox5 and Hox4 overlapped the region of TSS and alternative exons (Fig. 3b), suggesting that CNSs function as *cis* and *trans* regulatory elements.

Eight pairs of CNSs are conserved between two paralogous Hox clusters; and the authors named them DP (di-paralogous), as shown in Fig. 1. Results of phylogenetic footprinting analysis and pairwise sequence alignment are shown in Supplementary Figs. 2 and 3, respectively. The DP6 CNS is not located at syntenic region and the conservation is poor. Other DP CNSs are located at the syntenic region of each cluster and include functional elements (Table 2). DP1 and DP2 which are located at the upstream of Evx1 and Evx2 have *cis*-regulatory functions (Lehoczky et al. 2004). The region called "distal limb enhancer" in the HoxD cluster is essential for the posterior HoxD gene expression of appendicle (Spitz et al. 2001). The DP4 pair corresponds to microRNAs mir-196b and mir-196a-2. They belong to the mir-196 family. This family is composed of

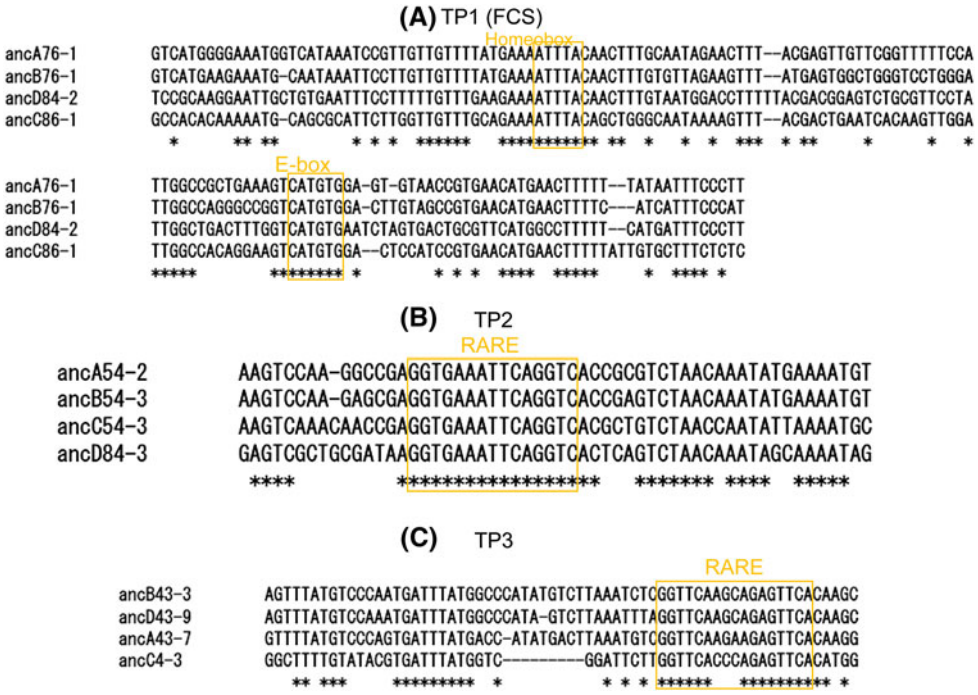

Fig. 2 Multiple alignments of three TP CNS sequences. **a–c** are the results of multiple alignments of paralogous conserved regions derived from each TP CNS. Aligned sequences are ancestral sequences estimated from each CNS using PAML4 program (Yang 2007). Alignments are generated using CLUSTALW (Thomopson et al. 1994). The putative TFBS are highlighted by *orange* (Color figure online)

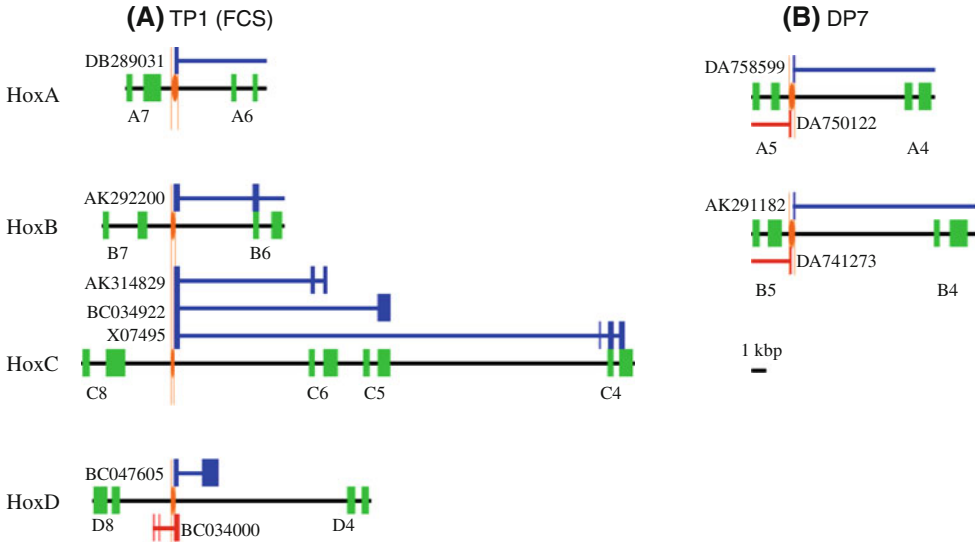

Fig. 3 The scheme of paralogous conserved bidirectional promoters. The authors mapped the paralogous CNSs on bidirectional transcript start sites which code alternative splicing RNAs of Hox genes and antisense RNAs. Paralogous CNSs are **a** TP1 and **b** DP7. The *blue* and *red* lines are sense RNAs and antisense RNAs, respectively. DDBJ/EMBL/GenBank accession numbers of these RNA sequences are also shown (Color figure online)

three members, which are mapped between Hox10 and Hox9 of HoxA, HoxB and HoxC clusters (Yekta et al. 2004). However, another member, mir-196a-1, was difficult to detect because of poor conservation. The authors thus defined only two microRNA members as CNS.

Comparison Between Vertebrate CNSs and Amphioxus CNSs Within Hox Clusters

Phylogenetic footprinting can be used to detect significantly conserved sequences between vertebrates and the amphioxus Hox cluster. Because the conservation of non-coding region between amphioxus and vertebrates is poor, Pascual-Anaya et al. (2008) defined CNS in the case of human–amphioxus comparison as approximately 60% identity and 50-bp length region. They reported 75 amphiCNSs. However, this might include CNSs which are not conserved among all vertebrates, but conserved only between human and amphioxus.

To remove these CNSs and to identify CNSs conserved among all vertebrates, the authors collected multiple orthologous vertebrate sequences and carried out

Table 2 Possible functions of Tetra (TP) and Di (DP) paralogous CNSs

Name	ID	Function	Putative TFBS	References
(A) TP CNSs				
TP1	A76-1	Anterior Hox promoter[a]	Homeobox, E-box	Kim et al. (2000), This study
	B76-1	Anterior Hox promoter[a]	Homeobox, E-box	Kim et al. (2000), This study
	C86-1	Anterior Hox promoter[a]	Homeobox, E-box	Kim et al. (2000), This study
	D84-2	Anterior Hox promoter[a]	Homeobox, E-box	Kim et al. (2000), This study
TP2	A54-2	Hox4 Enhancer	RARE	Mainguy et al. (2003)
	B54-3	Hox4 Enhancer	RARE	Mainguy et al. (2003)
	C54-3	Hox4 Enhancer	RARE	Mainguy et al. (2003)
	D84-3	Hox4 Enhancer	RARE	Mainguy et al. (2003)
TP3	A43-7	Hox3 Enhancer	RARE	Mainguy et al. (2003)
	B43-3	Hox3 Enhancer	RARE	Mainguy et al. (2003)
	C4-3	Hox4 Enhancer[a]	RARE	This study
	D43-9	Hox3 Enhancer	RARE	Mainguy et al. (2003)
(B) DP CNSs				
DP1	E1-1	Hox13 Enhancer (distal limb enhancer)	PPAR-α, GATA-1, POU1F1a, Homeobox	Lehoczky et al. (2004)
	E2-1	Hox13 Enhancer (distal limb enhancer)	PPAR-α, GATA-1, POU1F1a, Homeobox	Spitz et al. (2001)
DP2	E1-2	Hox13 Enhancer (distal limb enhancer)	C-Myb, Homeobox, YY1	Lehoczky et al. (2004)
	E2-3	Hox13 Enhancer (distal limb enhancer)	C-Myb, Homeobox, YY1	Spitz et al. (2001)
DP3	C1110-2	Hox10 enhancer[a]	SF1, CP1, Homeobox	This study
	D1110-3	Hox10 enhancer[a]	SF1, CP1, Homeobox	This study
DP4	A109-2	MicroRNA (mir-196 family)	–	Yekta et al. (2004)
	C109-2	MicroRNA (mir-196 family)	–	Yekta et al. (2004)
DP5	C109-4	Hox9 enhancer[a]	GR, E-box, CAT-box	This study
	D109-1	Hox9 enhancer[a]	GR, E-box, CAT-box	This study
DP6	B43-1	Hox3 enhancer[a]	GR	This study
	D84-1	Hox3 enhancer[a]	GR	This study
DP7	A54-1	Bidirectional promoter[a]	E-box, NF-1, E-box, CAT-box, TATA-box	This study
	B54-1	Bidirectional promoter	E-box, NF-1, E-box, CAT-box, TATA-box	Dinger et al. (2008)
DP8	A43-12	Hox3 enhancer[a]	USF, Homeobox	This study
	B43-5	Hox3 enhancer[a]	USF, Homeobox	This study

[a] Putative function

phylogenetic footprinting analysis. The authors then identified the highly conserved "core region" of each vertebrate CNS. By comparing amphiCNSs with the vertebrate CNSs, the authors found that only 16 out of 75 amphiCNSs overlap with the vertebrate CNSs. Eight of them show deep conservation; they are conserved among all the vertebrates used in this study(see vertebrate CNSs information shown in Supplementary Table 2). Two of eight amphiCNSs were aligned with the "core region" of the vertebrate CNSs; they are conserved among all the chordates used in this study (Fig. 4a). These are located at the HoxA and HoxB anterior regions, and supported a previous observation that the posterior region is more divergent than the anterior region (Ferrier et al. 2000).

The remaining six amphiCNSs did not correspond to the "core region" of the vertebrate CNSs (Fig. 4b and Supplementary Fig. 4). Interestingly, the "core region" of the vertebrate CNSs is often adjacent to the amphioxus–human conserved regions. At last, only 2 out of 75 amphiCNSs are significantly highly conserved among chordates.

Discussion

The authors defined 208 CNSs within the vertebrate Hox clusters. To infer the depth of sequence conservation, the authors investigated the existence of orthologous CNSs from vertebrate species. The depth of conservation is different with each cluster. The HoxC cluster shows the shallowest conservation. Despite this result, the HoxC cluster retains some paralogous CNSs. Shallow conservation of the HoxC

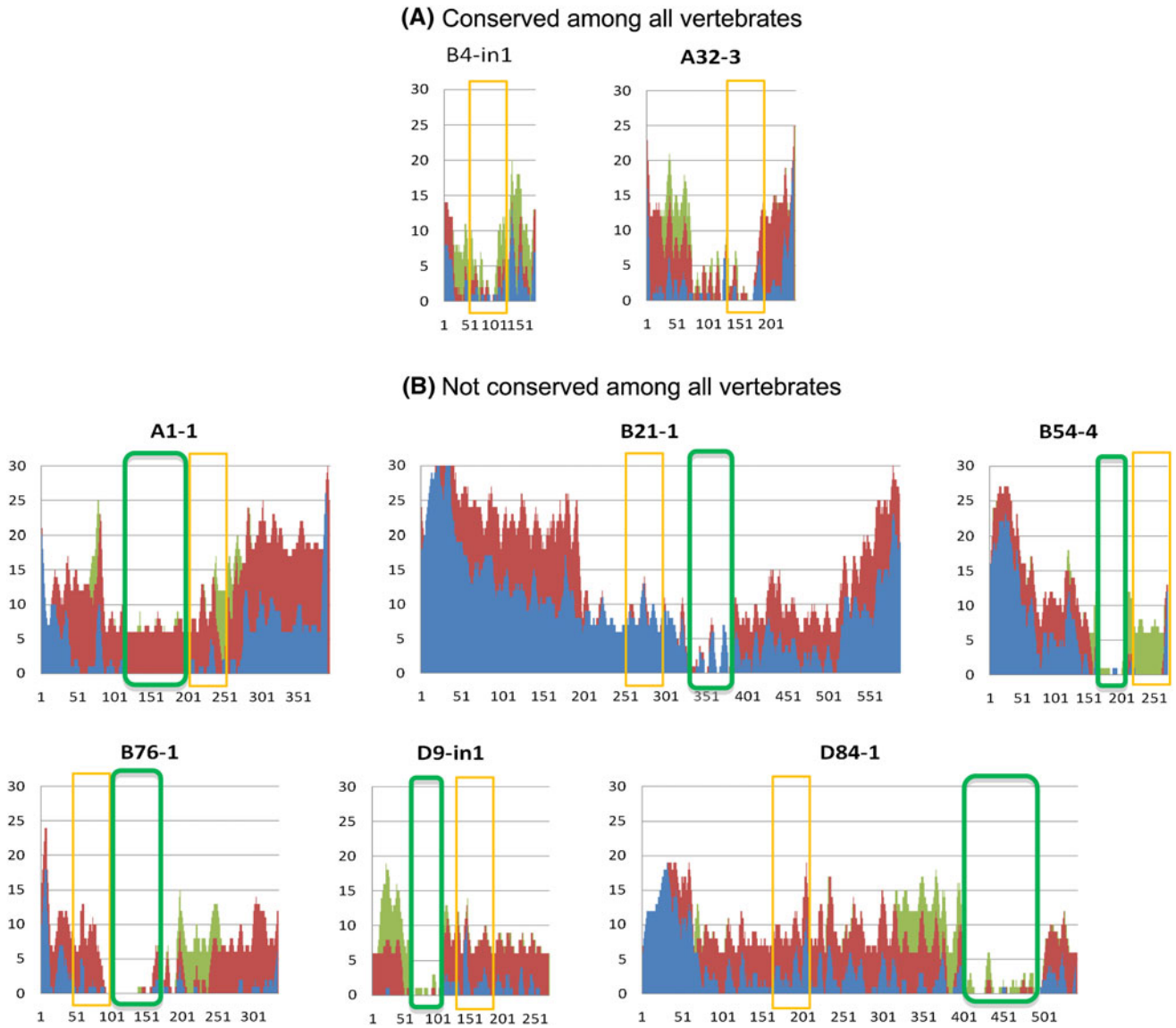

Fig. 4 The phylogenetic footprinting analysis within chordates. The authors compared the vertebrate CNSs with amphioxus–human CNSs (amphiCNSs). The results of phylogenetic footprinting are described. Each *orange* box corresponds to amphiCNS. **a** CNSs conserved among all the vertebrates. **b** CNSs not conserved among all the vertebrates. Each *green* box represents highly conserved region among vertebrates identified by phylogenetic footprinting. Each axis and color is the same as Supplementary Fig. 2 (Color figure online)

cluster could be an artifact. Because the intergenic sequence data of HoxC cluster is the poorest, the authors cannot detect the intergenic conservation from several species accurately. If the sequence data of Hox clusters are complete, then the abundance of CNS of each Hox cluster may not be so different.

The number of the CNSs located at anterior region is higher than that of CNSs located at posterior region. The divergence of posterior paralogous Hox genes are more rapid compared with other paralogous Hox genes, called "posterior flexibility" (Ferrier et al. 2000). For example, because posterior genes of the HoxD cluster are regulated not only by each gene regulatory element but also by the global control regulatory element located 240 kb upstream of the cluster (Spitz et al. 2003), the intergenic region of the posterior HoxD cluster might have poor conservation. The posterior HoxA genes show similar expression pattern with the posterior HoxD genes. Therefore, this tendency applies to the HoxA cluster. The HoxA cluster also have global control enhancers located at upstream of the cluster (Lehoczky and Innis 2008).

The DP CNSs have many putative TFBSs (Table 2 and Supplementary Fig. 3). The homeobox binding motifs are especially abundant. This suggests that DP CNSs are important for the auto regulatory mechanism of the four vertebrate Hox clusters. Each Hox protein may bind to

cis-regulatory regions of other Hox genes and controls the expression patterns. E-box is the motif related to the HLH (helix-loop-helix) transcription factor. HLH and homeobox proteins mainly regulate the expression pattern of Hox genes. The DP7 CNSs bear the conserved TATA-box. This suggests that the DP7 CNSs have promoter function as the authors described.

The authors identified three paralogous regions conserved among the four Hox clusters. One of them, FCS, was previously reported (Kim et al. 2000). Surprisingly, many RNAs are transcribed in this area. Different directional transcripts are started in the HoxA cluster. FCS of the HoxB cluster corresponds with TSS of the HoxB6 gene. In the HoxC cluster, FCS is the TSS of HoxC6, HoxC5 and HoxC4 coding transcripts. In the HoxD cluster, FCS might control different directional transcripts. Not only FCS but also other paralogous CNSs (DP7) between HoxA and HoxB clusters overlap with TSS and alternative exons (Fig. 3). Experimental approach revealed long non-coding antisense RNA started from this HoxB cluster region (Dinger et al. 2008). Because RNA data are insufficient to detect all the cluster transcripts, some of these transcripts are partial and were found only in human and/or mouse. It is probable that these paralogous CNSs play important roles in alternative transcription in other tetrapod species.

The other two TP CNSs (TP2 and TP3) include the RARE (Mainguy et al. 2003). Their functions are experimentally confirmed (Morrison et al. 1997). Retinoids are thought to exert their activities at the transcriptional level, acting as ligands to activate nuclear receptors. These nuclear receptors recognize DNA sequences closely related to 5′-(A/G)G(G/T)TCA-3′. Previous studies suggested that retinoic acids contribute to the expressions of Hox genes (Dubrulle and Pourquié 2004). TP2 and TP3 have type11 and type3 RAREs, respectively. A conserved sequence, TP3, downstream of HoxC4 gene was newly detected in this study. This sequence is located more than 20 kb away of the HoxC4 gene and corresponds to type3 RAREs. Amphioxus also has RARE in this intergenic region (Wada et al. 2006). However, the authors could not detect this element in this study. Only one motif conservation is difficult to detect using this method. Other motifs of those paralogous CNSs might function as *cis*-regulatory element that cooperates with RAREs.

It is possible that these TP CNSs are key components of cluster organization. The motifs within them might have already existed in the ancestor of vertebrates who had only one Hox cluster. Because other motifs are not conserved within the orthologous region of invertebrates but conserved in the paralogous region of vertebrates, they were acquired after the emergence of vertebrates.

Pascual-Anaya et al. (2008) reported 75 amphiCNSs which might include CNSs that are not conserved among all the vertebrates but conserved only between human and amphioxus. To remove these CNSs and to increase statistical significance, the authors compared multiple orthologous vertebrate sequences. The authors found that two amphiCNSs are overlapped and conserved in vertebrate CNSs. Ancestral DNA sequences of these CNSs have probably been under strong selective constraint throughout the chordate evolution, though their conservation is detected in only one Hox cluster. Other amphiCNSs might not be conserved among all the vertebrates. However, the authors should deal with this problem carefully, for only two amphioxus genomes were used to detect CNSs conserved among chordates. More information of the Hox cluster from non-vertebrate chordate genome is necessary to obtain the complete picture of chordate CNSs.

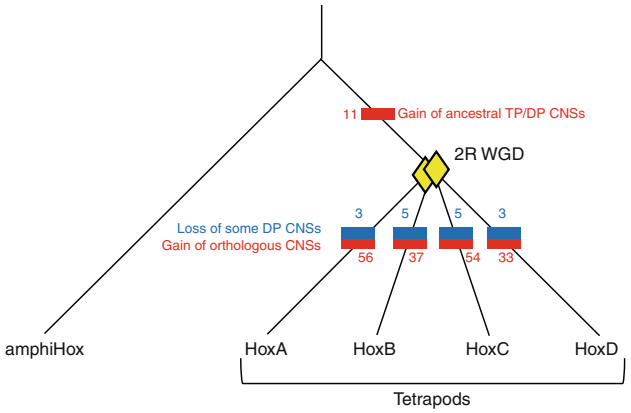

Fig. 5 The loss and gain of Hox CNSs during the chordate evolution. The Numbers of gain and loss of CNSs, shown in *red* and *blue* colors, respectively, are apportioned to the known Hox gene tree (Color figure online) (Color figure online) (Color figure online) (Color figure online)

The loss and gain of Hox CNSs are shown in the Fig. 5. After the 2R WGD, the massive gains of CNSs were occurred. In contrast, the conservation of non-coding regions in the invertebrate genomes is low. This difference on the Hox clusters may be related with the evolution of various unique features of vertebrates. When vertebrates acquired the more complex morphogenesis, the Hox clusters may become more conservative. To solve why these highly conserved CNS were appeared, it is necessary to consider the relationship between the non-coding functions and evolutionary conservations.

In summary, the authors efficiently detected orthologous CNSs of vertebrates. The authors identified three paralogous CNSs, and one of them bears a newly detected RARE motif. These CNSs are conserved among all the paralogous Hox clusters, and might contribute to Hox cluster organization and gene expression patterns.

Acknowledgments The authors thank Drs. Kiyoshi Ezawa, Hiroki Kokubo, Kazuho Ikeo, Toshihiko Shiroishi, and Hiroyuki Sasaki for their many helpful discussions, suggestions, and comments. The

authors appreciate English polishing by Ms Rumiko Suzuki, Ms Mahoko Takahashi, and Mr. Tim Jinam. This study was supported partly by The Graduate University for Advanced Studies (Sokendai) to M.M. and by Grant-in-Aid for scientific research from the Ministry of Education, Culture, Sports, Science, and Technology of Japan to N.S.

References

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

Amemiya CT, Prohaska SJ, Hill-Force A, Wasserscheid J, Ferrier DEK, Pascual-Anaya J, Garcia-Fernàndez J, Dewar K, Stadler PF (2008) The amphioxus Hox cluster: characterization, comparative genomics, and evolution. J Exp Zool (Mol Dev Evol) 310B:465–477

Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, Haussler D (2004) Ultraconserved elements in the human genome. Science 304:1321–1325

Carroll SB (2001) Chance and necessity: the evolution of morphological complexity and diversity. Nature 409:1102–1109

Chiu CH, Amemiya C, Dewar K, Kim CB, Ruddle F, Wagner GP (2002) Molecular evolution of the HoxA cluster in three major gnathostome lineages. Proc Natl Acad Sci USA 99:5492–5497

Dehal P, Boore JL (2005) Two rounds of whole genome duplication in the ancestral vertebrate. PLoS Biol 3:1700–1708

Dinger ME, Amaral PP, Mercer TR, Pang KC, Bruce SJ, Gardiner BB, Askarian-Amiri ME, Ru K, Soldà G, Simons C, Sunkin SM, Crowe ML, Grimmond SM, Perkins AC, Mattick JS (2008) Long noncoding RNAs in mouse embryonic stem cell pluripotency and differentiation. Genome Res 18:1433–1445

Dubrulle J, Pourquié O (2004) Coupling segmentation to axis formation. Development 131:5783–5793

Ferrier DE, Minguillon C, Holland PW, Garcia-Fernàndez J (2000) The amphioxus Hox cluster: deuterostome posterior flexibility and Hox14. Evol Dev 2:284–293

Fitch WM (1971) Towards defining the course of evolution: minimum change for a specific tree topology. Syst Zool 20:406–416

Garcia-Fernàndez J (2005) The genesis and evolution of homeobox gene clusters. Nat Rev Genet 6:881–892

Gibson TJ, Spring J (2000) Evidence in favour of ancient octaploidy in the vertebrate genomes. Biochem Soc Trans 28:259–264

Hancock JM, Shaw PJ, Bonneton F, Dover GA (1999) High sequence turnover in the regulatory regions of the developmental gene hunchback in insects. Mol Biol Evol 284:1083–1094

Holland PW, Garcia-Fernàndez J, Williams NA, Sidow A (1994) Gene duplications and the origins of vertebrate development. Dev Suppl 120(SUPPL.):125–133

Hughes AL, da Silva J, Friedman R (2001) Ancient genome duplications did not structure the human Hox-bearing chromosomes. Genome Res 11:771–780

Juan AH, Ruddle FH (2003) Enhancer timing of Hox gene expression: deletion of the endogenous Hoxc8 early enhancer. Development 130:4823–4834

Kim CB, Amemiya C, Bailey W, Kawasaki K, Mezey J, Miller W, Minoshima S, Shimizu N, Wagner GP, Ruddle F (2000) Hox cluster genomics in the horn shark, *Heterodontus francisci*. Proc Natl Acad Sci USA 97:1655–1660

Lehoczky JA, Innis JW (2008) BAC transgenic analysis reveals enhancers sufficient for Hoxa13 and neighborhood gene expression in mouse embryonic distal limbs and genital bud. Evol Dev 10:421–423

Lehoczky JA, Williams ME, Innis JW (2004) Conserved expression domains for genes upstream and within the HoxA and HoxD clusters suggests a long-range enhancer existed before cluster duplication. Evol Dev 6:423–430

Lemons D, McGinnis W (2006) Genomic evolution of Hox gene clusters. Science 313:1918–1922

Ludwig MZ, Palsson A, Alekseeva E, Bergman CM, Nathan J, Kreitman M (2005) Functional evolution of a cis-regulatory module. PLoS Biol 3:588–598

Mainguy G, In der Rieden PM, Berezikov E, Woltering JM, Plasterk RH, Durston AJ (2003) A position-dependent organization of retinoid response elements is conserved in the vertebrate *Hox* clusters. Trends Genet 19:476–479

Mainguy G, Koster J, Woltering J, Jansen H, Durston A (2007) Extensive polycistronism and antisense transcription in the mammalian Hox clusters. PLoS One 2:e356

McEwen GK, Woolfe A, Goode D, Vavouri T, Callaway H, Elgar G (2006) Ancient duplicated conserved noncoding elements in vertebrates: a genomic and functional analysis. Genome Res 16:451–465

Morrison A, Ariza-McNaughton L, Gould A, Featherstone M, Krumlauf R (1997) HOXD4 and regulation of the group 4 paralog genes. Development 124:3135–3146

Murphy WJ, Pevzner PA, O'Brien SJ (2004) Mammalian phylogenomics comes of age. Trends Genet 20:631–639

Ohno S (1970) Evolution by gene duplication. Springer Verlag, New York

Pascual-Anaya J, D'Aniello S, Garcia-Fernàndez J (2008) Unexpected number of conserved noncoding regions within the ancestral chordate Hox cluster. Dev Genes Evol 218:591–597

Pearson JC, Lemons D, McGinnis W (2005) Modulating Hox gene functions during animal body patterning. Nat Rev Genet 6:893–904

Prohaska SJ, Fried C, Flamm C, Wagner GP, Stadler PF (2004) Surveying phylogenetic footprints in large gene clusters: application to Hox cluster duplications. Mol Phylogenet Evol 31:581–604

Putnam NH, Butts T, Ferrier DEK, Furlong RF, Hellsten U, Kawashima T, Robinson-Rechavi M, Shoguchi E, Terry A, Yu JK, Benito-Gutie′rrez E, Dubchak I, Garcia-Fernàndez J, Grigoriev IV, Horton AC, de Jong PJ, Jurka J, Kapitonov V, Kohara Y, Kuroki Y, Lindquist E, Lucas S, Osoegawa K, Pennacchio LA, Salamov AA, Satou Y, Sauka-Spengler T, Schmutz J, Shin-I T, Toyoda A, Gibson-Brown JJ, Bronner-Fraser M, Fujiyama A, Holland LZ, Holland PWH, Satoh N, Rokhsar DS (2008) The amphioxus genome and the evolution of the chordate karyotype. Nature 453:1064–1071

Ray P, Shringarpure S, Kolar M, Xing EP (2008) CSMET: comparative genomic motif detection via multi-resolution phylogenetic shadowing. PLoS Comput Biol 4:e1000090

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

Spitz F, Gonzalez F, Peichel C, Vogt TF, Duboule D, Zákány J (2001) Large scale transgenic and cluster deletion analysis of the HoxD complex separate an ancestral regulatory module from evolutionary innovations. Genes Dev 15:2209–2214

Spitz F, Gonzalez F, Duboule D (2003) A global control region defines a chromosomal regulatory landscape containing the HoxD cluster. Cell 113:405–417

Thomopson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673–4680

Wada H, Escriva H, Zhang S, Laudet V (2006) Conserved RARE localization in amphioxus Hox clusters and implications for Hox

code evolution in the vertebrate neural crest. Dev Dyn 235:1522–1531

Werauch MT, Hughes TR (2010) Conserved expression without conserved regulatory sequence: the more things change, the more they stay the same. Trends Genet 26:66–74

Woolfe A, Goodson M, Goode DK, Snell P, McEwen GK, Vavouri T, Smith SF, North P, Callaway H, Kelly K, Walter K, Abnizova I, Gilks W, Edwards YJ, Cooke JE, Elgar G (2005) Highly conserved non-coding sequences are associated with vertebrate development. PLoS Biol 3:116–130

Yang Z (2007) PAML 4: phylogenic analysis by maximum likelihood. Mol Biol Evol 24:1586–1591

Yekta S, Shih IH, Bartel DP (2004) MicroRNA-directed cleavage of HOXB8 mRNA. Science 304:594–596