Exclusion of Repetitive DNA Elements from Gnathostome *Hox* Clusters

Claudia Fried [a], Sonja J. Prohaska [a], Peter F. Stadler [a,b]

[a] *Bioinformatics Group, Department of Computer Science, University of Leipzig Kreuzstraße 7b, D-04103 Leipzig, Germany.*
{studla,claudia,sonja}@bioinf.uni-leipzig.de

[b] *Institut für Theoretische Chemie und Molekulare Strukturbiologie, Universität Wien, Währingerstraße 17, A-1090 Wien, Austria*

Abstract

The *Hox* gene clusters of gnathostomes have a strong tendency to exclude repetitive DNA elements. In contrast, no such trend can be found in the *Hox* gene clusters of protostomes. Repeats "invade" the gnathostome *Hox* clusters from the 5' and 3' ends while the core of the clusters remains virtually free of repetitive DNA.

Key words: Hox gene clusters, repetitive DNA elements

1 Introduction

The *Hox* genes code for homeodomain containing transcription factors that are essential for embryonic patterning (McGinnis and Krumlauf, 1992). In many species they are organized in tightly linked clusters although in some cases the clusters have been broken up, see Tab. 1.

The homology of the vertebrate *Hox* genes with the genes in the Drosophila homeotic gene clusters was demonstrated already a decade ago (Akam, 1989; Schubert *et al.*, 1993). The common ancestor of all recent gnathostomes (sharks, bony fish, and tetrapods) had four clusters homologous to the mammalian ones (Holland and Garcia-Fernández, 1996; Prohaska *et al.*, 2003a). The two agnathan lineages, lampreys and hagfish, also exhibit multiple *Hox* clusters which, however, arose through duplication events independent of those leading to the mammalian clusters (Irvine *et al.*, 2002; Force *et al.*, 2002; Fried *et al.*, 2003; Stadler *et al.*, 2003). In contrast, protostomes and invertebrate deuterostomes (echinodermata, hemichordata, urochordata, and cephalochordata) have a single cluster (Martinez *et al.*, 1999; Pendleton *et al.*, 1993; Dehal *et al.*, 2002; Garcia-Fernández and Holland, 1994).

Table 1
Well-studied Hox clusters for which at least partial information on physical linkage is known.

Species	#	Size (kb)	Ref.
Vertebrates			
Homo sapiens	4	107, 199, 116, 94	[1]
Mus musculus	4	163, 173, 115, 108	[2]
Rattus norvegicus	4	165, 109, 116, 111	[3]
Xenopus laevis	4	100, ≥57, ≥54, ≥29	[4]
Latimeria menadoensis	4	?, ?, ?, ?	[5]
Heterodontus francisci	4	106, ?, ?, ≥67	[6]
Danio rerio	6	120, 38; 83, 74; 135	[7]
Takifugu rubripes	7	70, 28; 158, 14; 66, ?; 40	[8]
Petromyzon marinus	≥ 3	fragmented?	[9,10]
Other Deuterostomes			
Branchiostoma floridae	1	370	[11,12]
Ciona intestinalis	1	5 fragments	[13,14]
Strongylocentrotus purpuratus	1	∼ 500	[15,16]
Protostomes			
Drosophila melanogaster	1	274+248	[17]
Anopheles gambiae	1	1052	[18,19]
Tribolium castaneum	1	≥ 300	[20]
Schistocerca gregaria	1	≥ 700	[21]
Caenorhabditis elegans	1	403+207+138	[22,23,24]

References: [1] The Human Genome International Sequencing Consortium (2001), [2] Mouse genome project `http://www.sanger.ac.uk/Projects/M_musculus/`, [3] The Rat Genome Sequencing Consortium `http://www.hgsc.bcm.tmc.edu/projects/rat/`, [4] JGI Xenopus Genome Project `www.jgi.doe.gov/xenopus/`, [5] Chris T. Amemiya and Thomas P. Powers, pers. comm. (2003), see also Koh *et al.* (2003), [6] Kim *et al.* (2000), [7] Amores *et al.* (1998), [8] Amores *et al.* (2003), [9] Force *et al.* (2002), [10] Irvine *et al.* (2002), [11] Garcia-Fernández and Holland (1994), [12] Ferrier *et al.* (2000), [13] Dehal *et al.* (2002), [14] Spagnuolo *et al.* (2003), [15] Martinez *et al.* (1999), [16] Cameron *et al.* (2000), [17] von Allmen *et al.* (1996), [18] Powers *et al.* (2000), [19] Devenport *et al.* (2000), [20] Brown *et al.* (2002), [21] Ferrier and Akam (1996), [22] Burglin and Ruvkun (1993), [23] Aboobaker (2003), [24] The *C. elegans* Sequencing Consortium (1998).

The most striking difference between the *Hox*-cluster of *Drosophila melanogaster* and *Hox*-clusters of the gnathostomes is the fact that in the fly tandem duplications of of *Hox* genes and even non-*Hox*-genes are interspersed in the cluster (von Allmen *et al.*, 1996; Adams *et al.*, 2000; Negre *et al.*, 2003). While invertebrates have *Hox*-clusters with large intergenic distances that vary considerable among different species, one observes highly conserved distances between orthologous *Hox* genes in species as different as humans and sharks, see Table 1 for a summary and references. These facts suggest that the gnathostome *Hox* clusters have to satisfy much tighter organizational constraints than their invertebrate counterparts.

In order to corroborate this hypothesis we investigate here the distribution of repetitive DNA elements within and in the vicinity of *Hox* clusters. It has been mentioned in passing in the literature that repetitive DNA elements are depleted in the contiguous vertebrate *Hox* clusters (Hart *et al.*, 1987; Kim *et al.*, 2000; Wagner *et al.*, 2003). On the other hand, transposable elements have been reported close to *Hox* genes in organisms with fragmented *Hox* clusters: The *Pm18* fragment of the lamprey *Petromyzon marinus* around the *HoxW10a* contains a Tcl-like transposon. A reverse transcriptase gene has been predicted close to the *Hox-1* gene in the *Ciona intestinalis* genome (Dehal *et al.*, 2002). An enhanced frequency of transposon-mediated inversions in *Drosophila* (Casals *et al.*, 2003) was proposed as a possible cause for the fragmentation of the *Drosophila Hox*-cluster (Lewis *et al.*, 2003).

If gnathostome *Hox* clusters are indeed constrained to maintaining intergenic distances there should be a selection pressure against the invasion of repetitive DNA elements. A second argument for the exclusion of mobile DNA elements is based on their regulatory activities. Alu elements, for instance, often function as RNA polymerase III promotors. In some cases the regulatory abilities of mobile DNA elements are used by the host and are now central in control/enhancement of transcription (Britten, 1996; Stenger *et al.*, 2001). In general, however, we can expect that any interference with the the cross-regulatory network of a *Hox* cluster will be detrimental to its function. Hence there should be a strong selection pressure against mobile DNA elements in gene clusters with a high degree of cross regulation and small intergenic distances.

We therefore expect to observe a reduced density of repeats within the *Hox* clusters. We will show here that this is indeed the case in gnathostomes.

2 Methods

Hox cluster sequences were retrieved from Genbank for *Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, *Polypterus senegalus*, *Morone saxatilis*, *Drosophila melanogaster*, *Anopheles gambiae* and *Caenorhabditis elegans*. The sequences *Takifugu rubripes* were taken from web server of the Fugu Genome Project [1], the *Danio rerio* sequences are taken from the web server of the Danio rerio Sequencing Project [2] and Genbank. The sequences for the latter two organism are identical to those use in (Prohaska *et al.*, 2003b) for the analysis of phylogenetic footprints. Accession numbers are listed in the appendix.

Repetitive elements within *Hox* cluster sequences and in the adjacent 100kb segments of genomic DNA were determined by means of the `censor` server [3] web interface using `repbase 8.9` database (Jurka *et al.*, 1996; Jurka, 2000). Similar results, albeit with a significantly smaller number of detected repetitive elements, were obtained using `repeat masker` based on `repbase 7.4` [4] A graphical representation of the repeat distribution in a few *Hox* clusters is given in Fig. 1.

We report here both numbers n and total lengths $\mathcal{L}$ of repetitive elements. We define the "inside" of a *Hox* cluster as the intergenic regions between the most 5' and the most 3' *Hox* gene of the cluster. In the case of the fragmented clusters of *Drosophila melanogaster* and *Caenorhabditis elegans* we use all intergenic regions adjacent to a *Hox* gene. For comparison we use the genomic DNA adjacent to the *Hox* clusters in order to account for potential large scale variations in repeat densities. Data are normalized by the length ℓ of the analyzed sequence. The significance of the estimates for n is estimated assuming a Poisson distribution of repeats. The variance of the total length of repetitive sequences, $\sigma_{\mathcal{L}}$, can then be estimated by

$$\sigma^2 = \bar{n}\sigma_L^2 + \bar{L}^2\sigma_n^2 = \bar{n}(\sigma_L^2 + \bar{L}^2) \tag{1}$$

where $\bar{L}$ and σ_L are the mean and standard deviations of the distribution length distribution of the repeats.

[1] Version 3.0, `http://genome.jgi-psf.org/fugu6/fugu6.home.html`
[2] `http://www.sanger.ac.uk/Projects/D_rerio/`
[3] `http://www.girinst.org/`
[4] Smit, A.F.A. & Green, P.: RepeatMasker.
URL: `http://ftp.genome.washington.edu/RM/RepeatMasker.html`.

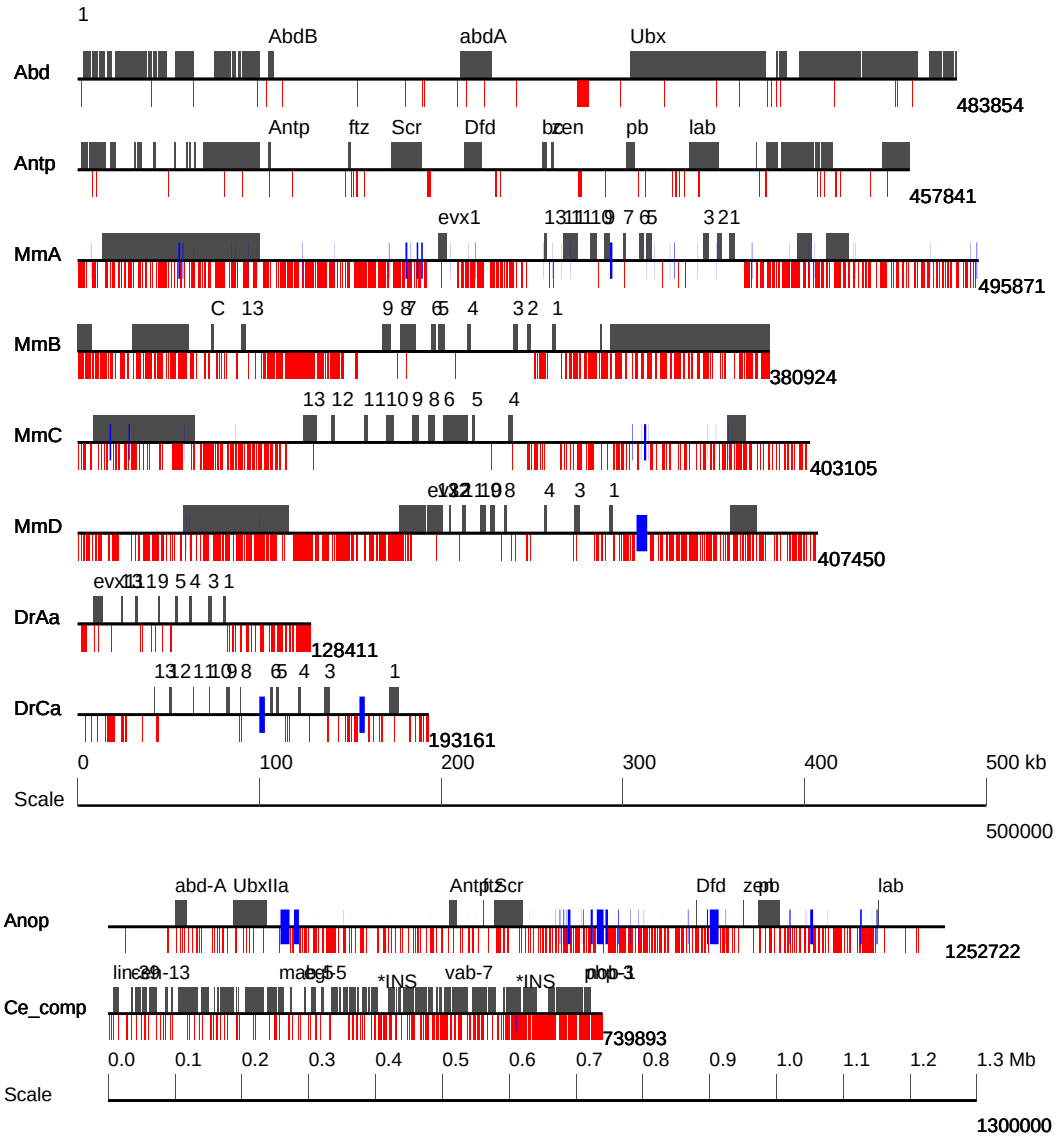

Fig. 1. Distribution of repetitive elements in some *Hox* clusters.
Boxes above the line represent coding regions and predicted genes. Gaps in the sequences are indicated by (blue) bars across the line, complex repetitive DNA elements are indicated below the line. The breaks in the *Caenorhabditis elegans* cluster are indicated by *INS.

3 Results

Number and length densities of repetitive elements for the Hox clusters are compiled in Table 2. We find that the repeat densities are 1-2 orders of magnitude smaller in gnathostome *A*, *C*, and *D* clusters. Surprisingly, in mammalian *B* clusters the reduction is only about 20-30%. When the intergenic region between the *Hox-B13* gene and its downstream neighbor is excluded, however, the ratio increases dramatically. The available sequence of the *Danio rerio* Ba

Table 2
Density of complex repetitive DNA elements inside and adjacent to *Hox* clusters.

Cl.	Repeats per 10000nt					Fraction of repetitive sequence				
	within		outside		ratio	within		outside		ratio
	n/ℓ	$\pm$	n/ℓ	$\pm$		$\mathcal{L}/\ell$	$\pm$	$\mathcal{L}/\ell$	$\pm$	
HsA	0.48	0.24	8.83	1.27	0.054	0.007	0.004	0.160	0.026	0.044
HsB	13.32	0.86	15.87	0.86	0.839	0.297	0.023	0.351	0.021	0.846
HsB'	3.29	0.66	17.58	0.75	0.187	0.059	0.013	0.394	0.019	0.150
HsC	0.86	0.30	10.62	0.84	0.081	0.013	0.005	0.236	0.026	0.055
HsD	2.44	0.55	13.34	1.09	0.183	0.048	0.012	0.349	0.036	0.138
MmA	0.83	0.34	15.15	0.74	0.055	0.008	0.003	0.270	0.022	0.030
MmB	10.36	0.85	14.57	1.35	0.712	0.165	0.017	0.199	0.021	0.829
MmB'	2.81	0.64	15.81	1.01	0.177	0.037	0.009	0.239	0.018	0.155
MmC	0.13	0.13	11.08	0.71	0.012	0.001	0.001	0.155	0.012	0.006
MmD	2.20	0.53	14.83	0.84	0.148	0.032	0.009	0.307	0.034	0.104
RnA	1.10	0.37	14.67	0.67	0.075	0.012	0.004	0.243	0.022	0.049
RnB	11.48	0.87	15.72	0.79	0.731	0.151	0.013	0.235	0.025	0.643
RnB'	4.14	0.73	17.43	0.89	0.237	0.054	0.010	0.229	0.013	0.236
RnC	0.36	0.21	12.43	0.73	0.029	0.002	0.001	0.200	0.019	0.010
RnD	2.32	0.56	14.98	1.03	0.155	0.034	0.009	0.234	0.022	0.145
DrAa	1.64	0.58	9.04	1.18	0.181	0.038	0.016	0.265	0.055	0.143
DrAb	3.61	1.20	6.30	1.00	0.573	0.089	0.031	0.181	0.046	0.492
DrBa'	1.13	0.43	5.28	1.21	0.214	0.031	0.014	0.099	0.029	0.313
DrBb	1.65	0.95	9.43	1.31	0.175	0.039	0.023	0.175	0.029	0.223
DrCa	2.76	0.49	6.19	1.03	0.445	0.060	0.014	0.139	0.029	0.432
PsA	0.18	0.18	1.86	0.32	0.097	0.002	0.002	0.050	0.012	0.040
HfA:h	0.23	0.16	1.04	0.74	0.222	0.001	0.001	0.015	0.011	0.067
HfA:z	0.12	0.12	0.52	0.52	0.223	0.001	0.001	0.003	0.003	0.333
HfD:h	0	0	1.23	0.56	0	0.000	0.000	0.018	0.009	0
HfD:z	0	0	0.50	0.35	0	0.000	0.000	0.010	0.007	0
Dm	0.97	0.16	0.86	0.22	1.122	0.032	0.016	0.008	0.002	4.000
Ag	4.09	0.22	0.50	0.16	8.180	0.129	0.013	0.010	0.006	12.90
Ce	7.12	0.89	16.08	0.67	0.443	0.105	0.015	0.203	0.011	0.517
Ce'	7.12	0.89	7.93	0.56	0.896	0.105	0.015	0.124	0.001	0.847

Species Abbreviations: Hs *Homo sapiens*, Mm *Mus musculus*, Rn *Rattus norvegicus*, Dr *Danio rerio*, Ps *Polypterus senegalus*, Hf *Heterodontus francisci*, Dm *Drosophila melanogaster*, Ag *Anopheles gambiae*, Ce *Caenorhabditis elegans*. Cluster designations: A, B, C, D: homologs to the four mammalian clusters; Aa, Ab, Ba, Bb, Ca, D for duplicated teleost clusters; Ant and Abd for the two pieces of the *Drosophila* clusters. B': fraction of the mammalian B-cluster (teleost Ba-cluster) from *Hox9* to the cluster-end only, the region from *HoxB13* to *HoxB9* is treated as an "outside" sequence. The shark (Hf) clusters have been analyzed with the human (:h) and zebrafish (:z) repeat databases. For Ce' we count only the sequences between the cluster fragments as "outside" sequence.

cluster is incomplete, spanning only the region from *Hox-B9* to *Hox-B1*.

The genome of the pufferfish *Takifugu rubripes* contains only a very few repetitive elements; this fact was one of the reasons to select the pufferfish for a

genome sequencing project (Aparicio *et al.*, 2002). Our data (not shown) are consistent with a reduced density of repeats also in the pufferfish. No repetitive elements were found in the *Hox-A10* to *Hox-A4* region of the striped bass *Morone saxatilis* sequences by Snell *et al.* (1999). For the bichir *Polypterus senegalus*, a basal actinopterygian fish, only the (unduplicated) *HoxA* cluster is available at present (Chiu *et al.*, 2003). Exclusion of repeats is clearly demonstrated. No dedicated data set of repetitive DNA is available for the hornshark *Heterodontus francisci*; we therefore analyze the repeats that match repeats from human or zebrafish. With both data sets we find an at least five-fold reduction of the repeat density within the *HoxM* and *HoxN* clusters, which are homologous to the mammalian *HoxA* and *HoxD* clusters, respectively. All available data thus show unambiguously that repetitive DNA is strongly excluded from the *Hox* clusters of gnathostomes.

In contrast, no significant exclusion of repeats has been detected in protostomes. The two insect sequences, *Drosophila melanogaster* and *Anopheles gambiae* even exhibit an over-representation of repetitive DNA within the cluster, while the reduction in the fraction of repetitive sequence in *Caenorhabditis elegans* is less than a factor of two.

In order to further characterize the distribution of repetitive elements within the gnathostome *Hox* cluster we analyzed each intergenic region separately. The corresponding data for the fraction of repetitive sequence are summarized in Figure 2. The most striking fact is that the density of repeats in the intergenic regions between *Hox-B13* and *Hox-B9* is almost the same as in the regions adjacent to the cluster. The "invasion" of repetitive elements also clearly visible at the 5'-end of the *Hox-A* and the 3'-side of the *Hox-C* and *Hox-D* clusters. It is interesting to note that there seems to be more in-clusters repeats in zebrafish sequences than in mammals. The central regions of the gnathostome *Hox* clusters, however, are almost entirely free of repetitive DNA sequences. In contrast, the protostome sequences do not exhibit virtually repeat-free regions.

In the clusters with highly reduced repeat density the repeats are also shorter. Figure 3 shows the ratio of total length of repetitive sequence inside and adjacent to the clusters is even smaller than the number densities, an effect that becomes more pronounced in clusters that exclude repeats more efficiently. This implies that selection pressure to exclude repetitive sequence also leads to a reduction in the length of the remaining repetitive elements.

The pressure against repetitive DNA does not distinguish significantly between different types of repeats. While the relative abundance of ALUs, non-ALU SINEs, LINEs, DNA transposons, and LTRs that are detected by `censor` differs widely between the different species considered here, there are only small variations between different regions in the same species.

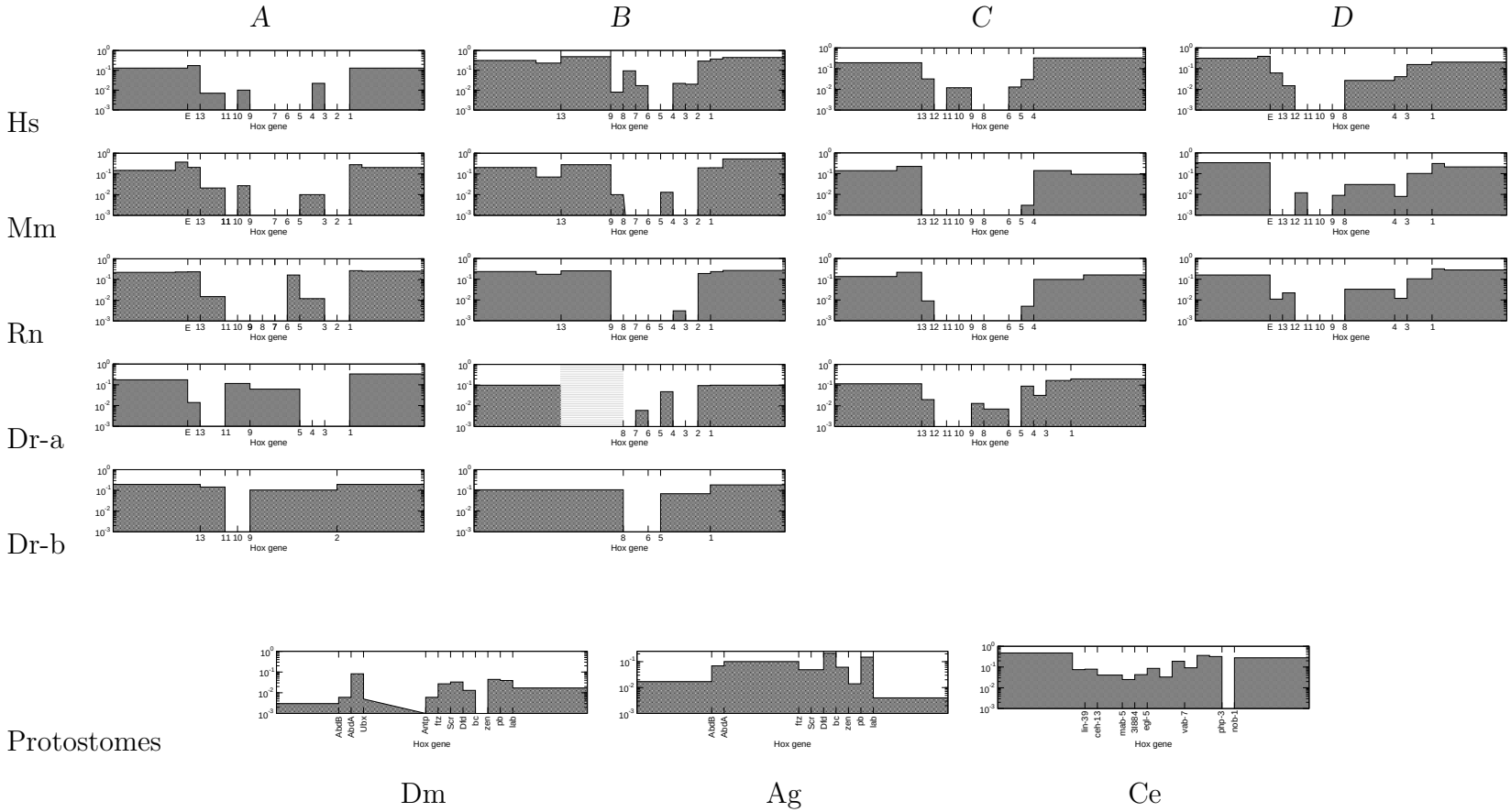

Fig. 2. Distribution of the fraction repetitive elements in the intergenic regions of gnathostome *Hox* clusters. The density of repeats if shown averaged over 100000nt up- and downstream of the cluster and for the intergenic regions between the indicated *Hox* genes. A gray block indicates missing data. The "invasion" of repeats from the cluster ends in the mammalian clusters (Hs, Mm, Rn) and the zebrafish (Dr) are clearly visible. The few repeat-free intergenic regions in the protostome clusters are probably counting artifacts since the corresponding sequence intervals are very short.

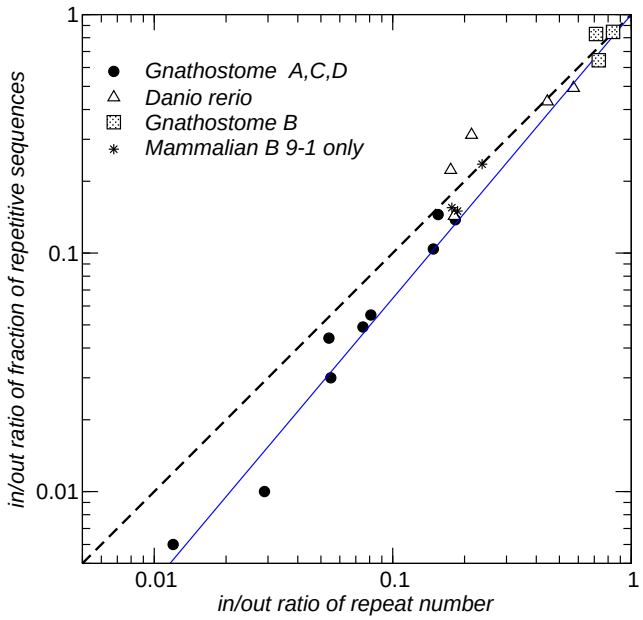

Fig. 3. Comparison of the ratio of number and total lengths of repetitive elements inside and adjacent to the *Hox* clusters from Tab 2.

4 Discussion

We have shown here that repetitive sequence elements are strongly excluded from gnathostome *Hox* clusters, while no such effect is detectable in protostomes. In the gnathostome *Hox* clusters we find that repetitive elements predominately accumulates in regions where *Hox* genes have been lost: the IGR between *HoxB13* and *HoxB9*, the 3' end of the *HoxC* and *HoxD* clusters. The *HoxAb* and *HoxBb* clusters of the zebrafish show this effect quite dramatically, Figure 2. Scemama *et al.* (2002) reported the invasion of repetitive sequence in the *HoxB3-HoxB2* region of the striped bass *Morone saxatilis* while the corresponding regions of the zebrafish is virtually repeat free. It is likely that the duplication of the *Hox* clusters in teleost fishes reduced the constraints on the structural integrity of cluster, thereby allowing repetitive elements to accumulate in the intergenic regions. More data will be necessary, however, to determine whether the slow disintegration of the clusters is an ongoing process.

The exclusion of repeats appears to be independent of the type of the repetitive elements. Furthermore, the few repeats that have invaded the "core" of a *Hox* are reduced in length.

A plausible explanation for these finding is that the selection against repeats is a consequence of the need to maintain intergenic distances within narrow bounds. This in turn can be explained by the the high density of regulatory

sequence motifs that are located in the intergenic regions of *Hox* clusters (Chiu *et al.*, 2002; Santini *et al.*, 2003; Prohaska *et al.*, 2003b,a; Chiu *et al.*, 2003). The activity of regulatory sequences depends on their exact distance from the the transcription start and from other regulatory sequences. Hence insertions should be selected against in most parts of the *Hox* cluster. Loss of genes from within the clusters would reduce the on length conservation in the vicinity of the deletion since most of regulatory sequence elements in this region will become non-functional as a consequence. As a consequence, there might be less resistance to the invasion of repetitive elements in such a region. This model is consistent with the distribution of repeats within mammalian clusters and explains the fact that the zebrafish *Hox* clusters are less efficient in excluding repeats: subsequent to the last duplication events a large number of genes were lost.

Our analysis suggests that the exclusion of repetitive sequence elements from *Hox* clusters may in fact be a gnathostome innovation since a significant reduction of repetitive sequences can be observed only in gnathostome lineages. The three available protostome *Hox* clusters do not exclude repetitive sequences. The lower deuterostomes seem to have a tendency toward fragmented *Hox* clusters, as exemplified by lampreys (Irvine *et al.*, 2002) and tunicates (Dehal *et al.*, 2002; Spagnuolo *et al.*, 2003). Even when the clusters are contiguous, as in amphioxus (Garcia-Fernández and Holland, 1994) and in sea urchins (Martinez *et al.*, 1999), they are comparable in length to the protostome rather than gnathostome *Hox* clusters. The question when exactly in early chordate evolution the organizational constraints on the *Hox* clusters tightened will be answered only when the complete sequences for amphioxus and the sea urchin *Hox* clusters become available.

Acknowledgments

Stimulating discussions with Dieter Schweizer and Günter P. Wagner, as well as financial support by the DFG Bioinformatics Initiative BIZ-6/1-2, are gratefully acknowledged.

References

Aboobaker AA, 2003. Hox gene loss during dynamic evolution of the nematode cluster. Curr Biol 13:37–40.

Adams MD, Celniker SE, Holt RA, 192 co-authors, 2000. The genome sequence of *Drosophila melanogaster*. Science 287:2185–2195.

Akam M, 1989. *Hox* and *HOM*: homologous gene clusters in insects and vertebrates. Cell 57:347–349.

Amores A, Force A, Yan YL, Joly L, Amemiya C, Fritz A, Ho RK, Langeland J, Prince V, Wang YL, Westerfield M, Ekker M, Postlethwait JH, 1998. Zebrafish hox clusters and vertebrate genome evolution. Science 282:1711–1714.

Amores A, Suzuki T, Yan YL, Pomeroy J, Singer A, Amemiya C, Postlethwait J, 2003. Developmental roles of pufferfish *Hox* clusters and genome evolution in ray-fin fish. Genome Res In press.

Aparicio S, Chapman J, Stupka E, Putnam N, Chia Jm, Dehal P, Christoffels A, Rash S, Hoon S, Smit A, Gelpke MDS, Roach J, Oh T, Ho IY, Wong M, Detter C, Verhoef F, Predki P, Tay A, Lucas S, Richardson P, Smith SF, Clark MS, Edwards YJK, Dogget N, Zharkikh A, Tavtigian SV, Pruss D, Barstead M, Evans C, Baden H, Powell J, Glusman G, Rowen L, Hood L, H. TY, Elgar G, Hawkins T, Venkatesh B, Rokhsar D, Brenner S, 2002. Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. Science 297:1301–1310.

Britten RJ, 1996. DNA sequence insertion and evolutionary variation in gene regulation. Proc Natl Acad Sci USA 93:9374–9377.

Brown SJ, Fellers JP, Shippy Teresa D. Richardson EA, Maxwell M, Stuart JJ, Denell RE, 2002. Sequence of the *Tribolium castaneum* homeotic complex: The region corresponding to the *Drosophila melanogaster* antennapedia complex. Genetics 160:1067–1074.

Burglin TR, Ruvkun G, 1993. The *Caenorhabditis elegans* homeobox gene cluster. Curr Opin Genet Dev 3:615–620.

Cameron RA, Mahairas G, Rast JP, Martinez P, Biondi TR, Swartzell S, Wallace JC, Poustka AJ, Livingston BT, Wray GA, Ettensohn CA, Lehrach H, Britten RJ, Davidson EH, Hood L, 2000. A sea urchin genome project: Sequence scan, virtual map, and additional resources. Proc Natl Acad Sci USA 97:9514–9518.

Casals F, Caceres M, Ruiz A, 2003. The foldback-like transposon *Galileo* is involved in the generation of two different natural chromosomal inversions of *Drosophila buzzatii*. Mol Biol Evol 20:674–685.

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

Chiu Ch, Dewar K, Wagner GP, Takahashi K, Ruddle FH, Ledje C, Bartsch P, Scemama JL, Stellwag E, Fried C, Prohaska SJ, Stadler PF, Amemiya CT, 2003. Bichir hoxa cluster sequence reveals surprising trends in ray-finned fish genomic evolution. Genome Res In press.

Dehal P, Satou Y, Campbell RK, Chapman J, Degnan B, De Tomaso A, Davidson B, Di Gregorio A, Gelpke M, Goodstein DM, Harafuji N, Hastings KEM, Ho I, Hotta K, Huang W, Kawashima T, Lemaire P, Martinez D, Meinertzhagen IA, Necula S, Nonaka M, Putnam N, Rash S, Saiga H, Satake M, Terry A, Yamada L, Wang HG, Awazu S, Azumi K, Boore J, Branno M, Chin-bow S, DeSantis R, Doyle S, Francino P, Keys DN, Haga S, Hayashi H, Hino K, Imai KS, Inaba K, Kano S, Kobayashi K, Kobayashi M, Lee

BI, Makabe KW, Manohar C, Matassi G, Medina M, Mochizuki Y, Mount S, Morishita T, Miura S, Nakayama A, Nishizaka S, Nomoto H, Ohta F, Oishi K, Rigoutsos I, Sano M, Sasaki A, Sasakura Y, Shoguchi E, Shin-i T, Spagnuolo A, Stainier D, Suzuki MM, Tassy O, Takatori N, Tokuoka M, Yagi K, Yoshizaki F, Wada S, Zhang C, Hyatt PD, Larimer F, Detter C, Doggett N, Glavina T, Hawkins T, Richardson P, Lucas S, Levine YKM, Satoh N, Rokhsar DS, 2002. The draft genome of *Ciona intestinalis*: Insights into chordate and vertebrate origins. Science 298:2157–2167.

Devenport MP, Blass C, Eggleston P, 2000. Characterization of the *Hox* gene cluster in the malaria vector mosquito, *Anopheles gambiae*. Evol Dev (pp. 326–339).

Ferrier DEK, Akam M, 1996. Organization of the *Hox* gene cluster in the grasshopper, *Schistocerca gregaria*. Proc Natl Acad Sci USA 93:13024–13029.

Ferrier DEK, Minguillón C, Holland PWH, Garcia-Fernández J, 2000. The amphioxus Hox cluster: deuterostome posterior flexibility and *Hox14*. Evol Dev 2:284–293.

Force A, Amores A, Postlethwait JH, 2002. Hox cluster organization in the jawless vertebrate *Petromyzon marinus*. J Exp Zool Mol Dev Evol 294:30–46.

Fried C, Prohaska SJ, Stadler PF, 2003. Independent hox-cluster duplications in lampreys. J Exp Zool Mol Dev Evol 299B:18–25.

Garcia-Fernández J, Holland PW, 1994. Archetypal organization of the amphioxus hox gene cluster. Nature 370:563–566.

Hart CP, Bogarad LD, Fainsod A, Ruddle FH, 1987. Polypurine/polypyrimidine sequence elements of the murine homeo box loci, hox-1, -2 and -3. Nucl Acids Res 15:5495.

Holland PW, Garcia-Fernández J, 1996. Hox genes and chordate evolution. Dev Biol 173:382–395.

Irvine SQ, Carr JL, Bailey WJ, Kawasaki K, Shimizu N, Amemiya CT, Ruddle FH, 2002. Genomic analysis of Hox clusters in the sea lamprey, *Petromyzon marinus*. J Exp Zool Mol Dev Evol 294:47–62.

Jurka J, 2000. Repbase update: a database and an electronic journal of repetitive elements. Trends Genet 9:418–420.

Jurka J, Klonowski P, Dagman V, Pelton P, 1996. Censor — a program for identification and elimination of repetitive elements from DNA sequences. Comput Chem 20:119–122.

Kim CB, Amemiya C, Bailey W, Kawasaki K, Mezey J, Miller W, Minosima S, Shimizu N, P. WG, Ruddle F, 2000. Hox cluster genomics in the horn shark, *heterodontus francisci*. Proc Natl Acad Sci USA 97:1655–1660.

Koh EGL, Lam K, Christoffels A, Erdmann MV, Brenner S, Venkatesh B, 2003. *Hox* gene clusters in the indonesian coelacanth, *Latimeria menadoensis*. Proc Natl Acad Sci USA 100:1084–1088.

Lewis EB, Pfeiffer BD, Mathog DR, Celniker SE, 2003. Evolution of the homeobox complex in the diptera. Current Biology 13:587–588.

Martinez P, Rast JR, Arena-Mena C, Davidson EH, 1999. Organization of an echinoderm *Hox* gene cluster. Proc Natl Acad Sci USA 96:1469–1471.

McGinnis W, Krumlauf R, 1992. Homeobox genes and axial patterning. Cell 68:283–302.

Negre B, Ranz JM, Casals F, Cáceres M, Ruiz A, 2003. A new split of the hox gene complex in drosophila: Relocation and evolution of the gene labial. Mol Biol Evol Epub 2003 Aug 29.

Pendleton J, Nagai BK, Murtha MT, Ruddle FH, 1993. Expansion of the Hox gene family and the evolution of chordates. Proc Natl Acad Sci USA 90:6300–6304.

Powers TP, Hogan J, Ke Z, Dymbrowski K, Wang X, Collins FH, Kaufman TC, 2000. Characterization of the hox cluster from the mosquito *Anopheles gambiae* (Diptera: Culicidae). Evol Dev 2:311–325.

Prohaska SJ, Fried C, Amemiya CT, Ruddle FH, Wagner GP, Stadler PF, 2003a. The shark HoxN cluster is homologous to the human HoxD cluster. J Mol Evol In press.

Prohaska SJ, Fried C, Flamm C, Wagner G, Stadler PF, 2003b. Surveying phylogenetic footprints in large gene clusters: Applications to Hox cluster duplications. Mol Phyl Evol In press; doi: 10.1016/j.ympev.2003.08.009.

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

Scemama JL, Hunter M, McCallum J, Prince V, Stellwag E, 2002. Evolutionary divergence of vertebrate *Hoxb2* expression patterns and transcriptional regulatory loci. J Exp Zool Mol Dev Evol 294:285–299.

Schubert FR, Nieselt-Struwe K, Gruss P, 1993. The antennapedia-type homeobox genes have evolved from three precursors separated early in metazoan evolution. Proc Natl Acad Sci USA 90:143–147.

Snell EA, Scemama JL, Stellwag EJ, 1999. Genomic organization of the hoxa4-hoxa10 region from *Morone saxatilis*: implications for hox gene evolution among vertebrates. J Exp Zool Mol Dev Evol 285:41–49.

Spagnuolo A, Ristoratore F, Di Gregorio A, Aniello F, Branno M, Di Lauro R, 2003. Unusual number and genomic organization of *Hox* genes in the tunicate *Ciona intestinalis*. Gene 309:71–79.

Stadler PF, Fried C, Prohaska SJ, Bailey WJ, Misof BY, Ruddle FH, Wagner GP, 2003. Evidence for independent *Hox* gene duplications in the hagfish lineage: A PCR-based gene inventory of *Eptatretus stoutii*. Mol Phylog Evol Submitted.

Stenger JE, Lobachev KS, Gordenin D, Darden T, Jurka J, Resnick MA, 2001. Biased distribution of inverted and direct Alus in the human genome: implications for insertion, exclusion, and genome stability. Genome Res 11:12–27.

The *C. elegans* Sequencing Consortium, 1998. Genome sequence of the nematode *C. elegans*: A platform for investigating biology. Science 282:2012–2018.

The Human Genome International Sequencing Consortium, 2001. Initial se-

quencing and analysis of the human genome. Nature 409:860–921.
von Allmen G, Hogga I, Spierer A, Karch F, Bender W, Gyurkovics H, Lewis E, 1996. Splits in fruitfly *Hox* gene complexes. Nature 380:116.
Wagner GP, Amemiya C, Ruddle F, 2003. Hox cluster duplication and the genetics of evolutionary novelties. Proc Natl Acad Sci In press.

Appendix: Accession Numbers

HsA: AC004080 r.c. (reverse complement), AC010990 r.c. (overlaps 200nt with AC004080), and AC004079 (pos. 75001-end, r.c., overlaps 200nt with AC010990), as in (Chiu *et al.*, 2002); HsB: NT_010783 (pos. 931646-1263780); HsC: NT_009563 (pos. 580371-708054 r.c); HsD: NT_037537 (pos. 4075338-end);

HfM: AF479755; HfN: AF224263;

PsA: AC132195, AC12632, as in (Chiu *et al.*, 2003);

DrAa: AC107365; DrAb: AC107364; *Morone saxatilis* MsAa: AF089743;