The Amphioxus *Hox* Cluster: Characterization, Comparative Genomics, and Evolution

Chris T. Amemiya [a,b], Sonja J. Prohaska [c,d,e], Alicia Hill-Force [a], April Cook [f], Jessica Wasserscheid [i] David E. K. Ferrier [g], Juan Pascual Anaya [h], Jordi Garcia-Fernàndez [h], Ken Dewar [i], Peter F. Stadler [j,k,d,c,*]

[a] *Benaroya Research Institute at Virginia Mason, 1201 Ninth Avenue, Seattle, WA 98101 USA*
camemiya@benaroyaresearch.org,ahill-force@bastyr.edu

[b] *Department of Biology, University of Washington, 106 Kincaid Hall, Seattle, WA 98195 USA*

[c] *Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, NM 87501, USA*
sonja@santafe.edu

[d] *Department of Theoretical Chemistry, University of Vienna, Währingerstraße 17, A-1090 Wien, Austria*

[e] *Department of Biomedical Informatics, School of Computing and Informatics, Arizona State University, Tempe, PO-Box 878809, AZ 85287, USA*

[f] *Broad Institute of MIT and Harvard, 320 Charles Street, Cambridge, Massachusetts 02141, USA*
cook@broad.mit.edu

[g] *The Gatty Marine Laboratory, University of St Andrews, St Andrews, Fife, KY16 8LB, Scotland, UK*
dekf@st-andrews.ac.uk

[h] *Departament de Genètica, Facultat de Biologia, Universitat de Barcelona, Av. Diagonal, 645, E-08028 Barcelona, Spain*
jordigarcia@ub.edu,jpascual@ub.edu

[i] *McGill University and Génome Québec Innovation Centre, 740 Avenue Doctor-Penfield, Montreal, Québec H3A 1A4, Canada*
ken.dewar@mcgill.ca,jessica.wasserscheid@mail.mcgill.ca

[j] *Bioinformatics Group, Dept. of Computer Science, Interdisciplinary Center for Bioinformatics, University of Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany*
studla@bioinf.uni-leipzig.de

[k] *RNomics Group, Fraunhofer Institut für Zelltherapie und Immunologie, Deutscher Platz 5e, D-04103 Leipzig, Germany*

Manuscript *19 January 2008*

Abstract

The amphioxus *Hox* cluster is often viewed as "archetypal" for the chordate lineage. Here we present a descriptive account of the 448kb region spanning the *Hox* cluster of the amphioxus *Branchiostoma floridae* from *Hox14* to *Hox1*. We provide complete coding sequences of all 14 previously described amphioxus sequences and give a detailed analysis of the conserved noncoding regulatory sequence elements. We find that the posterior part of the *Hox* cluster is so highly derived that even the complete genomic sequence is insufficient to decide whether the posterior *Hox* genes arose by independent duplications or whether they are true orthologs of the corresponding gnathostome paralog groups. In contrast, the anterior region is much better conserved. The amphioxus *Hox* cluster strongly excludes repetitive elements with the exception of two repeat islands in the posterior region. Repeat exclusion is also observed in gnathostomes, but not protostome *Hox* clusters. We thus hypothesize that the much shorter vertebrate *Hox* clusters are the result of extensive resolution of the redundancy of regulatory DNA following the genome duplications rather than the consequence of a selection pressure to remove non-functional sequence from the *Hox* cluster.

Key words: *Hox* cluster, amphioxus, *Branchiostoma floridae*

1 Introduction

The *Hox* cluster has been a subject of extreme interest to both evolutionary and developmental biologists due to both its highly conserved organization in terms of gene composition, gene structure, and gene order and gene orientation, and its intimate involvement in developmental patterning and formation of the bauplan (Gellon and McGinnis, 1998; Capecchi, 1997; Holland and Garcia-Fernàndez, 1996; Zakany and Duboule, 2007). *Hox* gene expression is typically organized such that the genes at the 3' end of the cluster are active first and/or at the anterior end of the embryo, followed by progressive activation and/or staggered expression along the anterior-posterior axis of the embryo of the genes along the Hox cluster, until the genes at the 5' end of the cluster are activated last and/or at the posterior end of the embryo. This phenomenon that the order of genes along the chromosome is related to the order of their expression along the anterior-posterior axis has is termed *colinearity* (Krumlauf, 1992; Duboule and Deschamps, 2004).

Hox genes are also known to have numerous other roles in vertebrates, including contributions to hematopoiesis and lymphomagenesis (Abramovich and Humphries, 2005; Eklund, 2006) and development of reproductive organs (Lynch *et al.*, 2004; Wagner and Lynch, 2005; Podlasek *et al.*, 2002). The conservation of *Hox* genes has permitted routine PCR surveys from a wide array of metazoans, allowing coarse determination of the *Hox* composition of species for which little genomic information is available, including a wide

range of invertebrates (de Rosa *et al.*, 1999; Lee *et al.*, 2003; Fritzsch *et al.*, 2007) and vertebrates (Longhurst and Joss, 1999; Stadler *et al.*, 2004). PCR approaches can then used to isolate full-length *Hox* cDNAs, further enabling characterization of expression patterns during development. This general approach has been very successful for many protostome and deuterostome taxa, see e.g. (Hara *et al.*, 2006; Manuel *et al.*, 2006).

However, because the genes are not isolated in their genomic context, relationships with regard to co-linearity of physical arrangement and expression patterns during embryonic and larval development are not known or can only be inferred. The importance of obtaining entire *Hox* cluster sequences is further emphasized in light of recent findings that demonstrate:

(1) striking conservation of presumptive regulatory elements within and among *Hox* clusters (Chiu *et al.*, 2002; Frasch *et al.*, 1995; Gould *et al.*, 1997; Spitz *et al.*, 2003);
(2) conservation of certain microRNAs within the *Hox* clusters that may have regulatory activities (Pearson *et al.*, 2005; Tanzer *et al.*, 2005; Yekta *et al.*, 2004); and
(3) clear disintegration of *Hox* clusters within certain metazoan lineages (Ikuta and Saiga, 2005; Seo and *al.*, 2004), reviewed in (Monteiro and Ferrier, 2006; Prohaska *et al.*, 2006).

The availability of bacterial artificial chromosome (BAC) libraries and the improved efficiency of high-throughput shotgun sequencing are now enabling the targeted sequencing of many metazoan *Hox* clusters for further comparative studies.

The sequence of the *Hox* cluster of the cephalochordate, amphioxus, is of particular interest due to the phylogenetic position of the *Cephalochordata* as an outgroup to the vertebrates (Delsuc *et al.*, 2006). Previous work based on lambda-phage chromosomal walking, physical mapping and fragmentary DNA sequencing has shown that amphioxus possesses a single *Hox* cluster whose homeobox composition and organization bear a clear relationship to *Hox* clusters of higher vertebrates (e.g., mouse) (Garcia-Fernández and Holland, 1994). This has led to the suggestion that amphioxus comprises an "archetypal" *Hox* cluster relative to those in the duplicated genomes of vertebrates (Garcia-Fernández and Holland, 1994).

However, its cluster was found to be comparatively larger than those of mammals, possessed an extra gene at its 5' end (*AmphiHox14*) and seemingly exhibited uneven rates of molecular evolution (Ferrier *et al.*, 2000). The phenomenon whereby the posterior *Hox* genes have apparently evolved faster in deuterostomes than in protostomes has been termed *"Deuterostome Posterior Flexibility"* by Ferrier *et al.* (2000). In most phylogenetic analyses, the pos-

terior *AmphiHox* genes neither group unambiguously with the corresponding paralog groups (PG) of vertebrates nor clearly support independent duplication events, see e.g. (Ferrier *et al.*, 2000; Campos *et al.*, 2004; Ferrier, 2004; Peterson, 2004; Cameron *et al.*, 2006). Despite the discovery of PG14 genes in some vertebrates (Powers and Amemiya, 2004), the question is still open how exactly the posterior genes *AmphiHox14-AmphiHox10* are related to the vertebrate PG14-PG10 *Hox* genes.

In an important set of experiments, Peter Holland and coworkers demonstrated that noncoding fragments from the 3' end of the amphioxus *Hox* cluster could effectively drive transcription of minimal promoter constructs in vertebrate-specific structures (neural crest, placodes) in chick and mouse embryos (Manzanares *et al.*, 2000). This is significant in that it implied that noncoding elements in amphioxus were conserved enough to direct regulatory activities in a vertebrate assay system and that perhaps it would be possible to delineate what these elements were and how they evolved in both sequence and function.

In order to address questions germane to vertebrate *Hox* gene and cluster evolution, and the divergence of their regulatory control elements, it is imperative to obtain not only the sequences of the *Hox* genes but also the intervening noncoding DNA. This is a prerequisite for a detailed analysis of organization, phylogenetic footprint signatures, repeat abundance and molecular evolution.

In this paper we therefore report the *Hox* sequence of the Florida lancelet, *Branchiostoma floridae*, based on a regional assembly of selected BAC (bacterial artificial chromosome) and PAC (P1 artificial chromosome) clones that span the region from about 7kb upstream of *AmphiHox14* to 41kb downstream of *AmphiHox1*.

The carefully annotated sequence of the nearly complete amphioxus *Hox* cluster should serve to direct empirical investigations into the evolution and divergence of vertebrate developmental gene regulation as well as provide a suitable outgroup for future studies in the comparative genomics of chordate *Hox* clusters. Extensive supplemental data are provided in electronic form at `http://www.bioinf.uni-leipzig.de/Publications/SUPPLEMENTS/07-029/`.

2 Materials and Methods

2.1 Specimen Procurement and Genomic Libraries

Adult amphioxus specimens were purchased from Gulf Specimen Marine Lab (Panacea, Florida). Six eviscerated specimens were pulverized in liquid nitrogen using a Waring blender. The powdered material was embedded in Incert™ agarose (FMC), processed for high molecular weight DNA, and a PAC library (VMRC2) was generated using methods described in (Osoegawa *et al.*, 1998; Amemiya *et al.*, 1996). The library utilized MboI partial digests and pCYPAC7 vector (GenBank ***DQ092493***); it is arrayed in fifty-four 384-well microtiter dishes and comprises approximately 5× coverage of the roughly 500Mb amphioxus genome.

A BAC library that was constructed from a single specimen at the BACPAC facility at Children's Hospital of Oakland Research Institute (*CHORI-302 amphioxus BAC library* [1]) was also used for this study.

2.2 Isolation of the Amphioxus Hox *Cluster*

High-density colony filters of the amphioxus PAC library were screened with probes encompassing non-homeobox regions from *AmphiHox1* to *AmphiHox14*. PAC clones were assessed as to their gene content using PCR with primers designed specifically to amphioxus *Hox* genes Clones were sized by excising the inserts with *NotI* and electrophoresing on pulsed field gels. Based on insert sizes and gene composition, a minimal spanning path was generated; these clones were selected for DNA sequencing. A gap that encompassed *AmphiHox9-AmphiHox7* was subsequently filled with a BAC isolated from the CHORI amphioxus library using hybridization; this BAC was also sequenced.

2.3 DNA Sequencing and Assembly

PAC and BAC clones were sequenced using standard high-throughput techniques and strategies (as described in International Human Genome Sequencing Consortium (2001)) PAC and BAC DNAs were physically sheared into $2-4$kb random fragments, subcloned into compatible plasmids, and maintained in laboratory strains of *E. coli* in arrays in 96-well or 384-well microtiter plates. Subcloned DNAs were purified and sequenced in each orientation. The

[1] `http://bacpac.chori.org/amphiox302.htm`

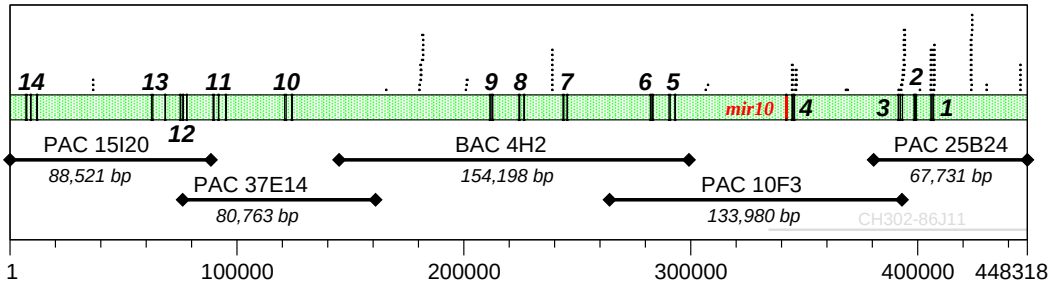

Fig. 1. *Hox* cluster organization of *Branchiostoma floridae*. The 448 318bp region subjected to DNA sequencing and regional assembly is displayed. The overlapping PAC and BAC clones used to construct the contig are shown below the map, with their names, sizes, and GenBank accession numbers also given. *Hox* genes 1 to 14 and exons are shown, with all genes being transcribed from left to right. Exons are shown as boxes. In addition, the location of the putative microRNA *mir-10* gene (Tanzer *et al.*, 2005) is shown. Above, the Branchiostoma ESTs available in `dbEST` are summarized. Most of the non-*Hox* ESTs arise from a variety transposons. In the anterior part, several ESTs indicate the presence of non-coding transcripts in the IGRs of *AmphiHox4* and *AmphiHox3* (***BW949805*** and ***BW891939***) and *AmphiHox2* and *AmphiHox1* (***BW842892***, ***BW918623***, ***BW900614***, ***BW918623***, ***BW860661***).

resultant sequences and paired-end information were then assembled to reproduce the sequence of the original PAC or BAC, after which residual sequencing gaps or ambiguous sequences were corrected following PCR directed closure and sequencing. Sequence of the four PAC clones, 15I20 (***AC129909***), 37E14 (***AC129910***), 10F3 (***AC124817***), 25B24 (***AC124805***), and BAC clone 4H2 (***AC214474***) were deposited in Genbank.

As with other species that comprise large outbreeding populations, amphioxus exhibit high levels of polymorphisms. Since six different individuals were used for the PAC library and another individual was used for the BAC library, each of the four pairs of overlapping clones is likely to stem from different alleles. The differences in the overlapping regions ($\sim 12-28$kb) were used to assess genetic variation in terms of nucleotide substitutions and indels.

In order to obtain a unique reference sequence for further analysis we arbitrarily defined that the longer of the two clones takes precedence. For example, where PAC 37E14 and BAC 4H2 overlap ($\sim$ 20kb), the sequence from the BAC was used since the BAC clone contained a larger insert. Using this criterion, the entire sequence build comprises 448 318 base pairs. A map showing the spanning path across the entire amphioxus *Hox* cluster is given in Fig. 1.

A BAC clone independently sequenced and encompassing *AmphiHox4* to 120kb downstream of *AmphiHox1* was identified through database searches of GenBank (***AC150428***). This BAC clone was not included in our *Hox* sequence

Table 1
Gene prediction software used for annotation

Program	URL	Reference
GrailEXP	http://compbio.ornl.gov/grailexp	Uberbacher *et al.* (1996)
GeneID	http://www1.imim.es/geneid.html	Parra *et al.* (2000)
GeneMark	http://opal.biology.gatech.edu/GeneMark	Besemer and Borodovsky (2005)
GenScan	http://genes.mit.edu/GENSCAN.html	Burge and Karlin (1997)
GenomeScan	http://genes.mit.edu/genomescan.html	Burge and Karlin (1998)

assembly (Fig. 1a), however, it was used for pairwise comparison in order to assess the degree of nucleotide polymorphism in the sequenced regions (see below).

2.4 Annotation of the Amphioxus Hox *Cluster*

The complete amphioxus *Hox* cluster sequence was annotated by two methods:

(1) Comparison with known *AmphiHox* sequences. A list of previously published complete coding sequences is given in the Supplemental Material. For *AmphiHox9* we obtained a partial cDNA sequence, and structure of *AmphiHox14* was determined by comparison with the corresponding genomic sequence of *Branchiostoma lanceolatum* (J. Garcia-Fernàndez, unpublished). In addition, publicly available EST sequences (Yu *et al.*, 2007) were mapped to our contig using `blastn`.
(2) *Ab initio* gene prediction was performed using the programs listed in Table 1. Of these five, only `GenomeScan` utilizes a user-defined training set of Hox protein sequences in order to specifically predict *Hox* gene models; it thereby proved to be the most reliable. All annotations were entered manually using VectorNTI™ software, version 8 (Informax-Invitrogen). Translations of all 14 *Hox* genes and the melded and annotated Hox peptide sequences are provided in the Supplemental Material.

2.5 HOX Alignments and Analysis of Phylogenetic Footprints

For global alignments and their visualization we employed both `PipMaker`[2] (Schwartz *et al.*, 2000) and `VISTA`[3] (Mayor *et al.*, 2000); an example is shown in the electronic supplement. Due to the large sequence divergence and size

[2] `http://pipmaker.bx.psu.edu/pipmaker`
[3] `http://genome.lbl.gov/vista`

discrepancy between the amphioxus *Hox* cluster and other vertebrate *Hox* sequences, this method was not optimal for detecting conserved sequence tracks.

A more sensitive method, `tracker` (Prohaska *et al.*, 2004), was utilized to detect phylogenetic footprints between the amphioxus *Hox* gene cluster and the following gnathostome *Hox* clusters: *Heterodontus francisci* (horn shark, Hf) A, B and D cluster, and all four clusters of *Latimeria menadoensis* (Lm, Indonesian coelacanth; CTA unpublished), *Monodelphis domestica* (Md, South American opossum; UCSC: monDom1), *Canis familiaris* (Cf, domestic dog; ENSEMBL 28-02-2003), *Homo sapiens* (Hs, as in (Prohaska *et al.*, 2004), *Mus musculus* (Mm, A: ***NT_039343*** [3927927-4123797, reverse complement]; B: ***AC011194***, C: ***NT_028016***; D: ***AC_015584***. These sequences are provided in the electronic supplement. To run `tracker` on the 24 cluster sequences we had to group them in smaller sets of 4 to 7 clusters. We compiled two sets of runs: *Set 1* was composed of 4 `tracker` runs, each aligning the amphioxus *Hox* cluster with the available vertebrate clusters of the same type (i.e., HOX-A, HOX-B, HOX-C, or HOX-D). *Set 2* consists of 6 individual `tracker` runs, each aligning the amphioxus *Hox* cluster with the available 3 or 4 clusters of the same species (Hf, Lm, Md, Mm, Cf or Hs).

In brief, the `tracker` program is based on `blastz` for the initial search of all pairs of input sequences. Comparisons are optionally restricted to homologous intergenic regions. The resulting list of pairwise sequence alignments is then assembled into groups of partially overlapping regions that are subsequently passed through several filtering steps. The end result of the procedure is a collection of multiple local sequence alignments, which we will refer to as "footprints" for the purpose of this analysis. Since several local alignments may sometime contain the same sequence interval, we use the number of unique nucleotides in the amphioxus sequence that are contained in such alignments as the basis for the statistical analysis.

RARE sites were determined by exact pattern matching using a customized `perl` script. RARE sequences were taken from (Mainguy *et al.*, 2003) and (Wada *et al.*, 2006).

2.6 Repeat Content

We first attempted to use `Censor` (Kohany *et al.*, 2006) to determine repetitive elements. Since no repeat set for amphioxus is publicly available, the algorithm was run using existing vertebrate and invertebrate masks, however without significant result.

We therefore decided to use the repeat sequences provided by the JGI genome

browser[4] for the two scaffolds 206 (1155kb) and 402 (738kb), which contain the *Hox* cluster in the `brafl1` assembly. In order to obtain comparable data, we used `blastn` with $E < 10^{-10}$ to map these sequences back to both our *Hox* cluster sequence and the two `brafl1` scaffolds. Since the repeat density within the *Hox* cluster is lower than the overall repeat density of scaffolds 206 and 402, the average over the scaffolds actually underestimates the repeat density surrounding the *Hox* cluster.

2.7 Phylogenetic Analysis of Protein Sequences

Phylogenetic analyses of the amphioxus *Hox* genes based on the homeobox-containing exon-2 have been repeatedly published in the past, see e.g. (Garcia-Fernández and Holland, 1994; Popodi *et al.*, 1996; Ferrier *et al.*, 2000). In order to assess whether the additional sequence information contained in the complete *Hox* coding sequences is phylogenetically informative, we compared the exon-1 sequences with both the protein and the nucleic acid section of `Genbank` using `blastx` and `tblastx`, respectively. Significant hits were found only for the anterior sequences *AmphiHox1-AmphiHox5*. Even for these genes only short portions of the first exon are alignable with other deuterostome *Hox* sequences. Since the homology of these genes with the corresponding PG1 through PG5 genes of other deuterostomes is undisputed, we concluded that our extended coding sequences cannot contribute to a better understanding of the duplication history. Hence we have not pursued the construction of gene phylogenies in this contribution.

3 Results and Discussion

3.1 Isolation, Sequencing, and Characterization

We have sequenced and assembled a 448 318 base pair region of the *Branchiostoma floridae* genome that extends from 6 964bp upstream of the putative translation start site of *AmphiHox14* and 41 271bp downstream of the translation termination site of *AmphiHox1*. Database searches and `Genscan` analysis with these upstream and downstream sequences failed to identify additional genes, see also (Minguillón *et al.*, 2005).

A `GenScan` analysis of the region downstream of *AmphiHox1*, which is contained in a BAC clone (***AC150428***) that was independently sequenced iden-

[4] `http://genome.jgi-psf.org/Brafl1/Brafl1.home.html`

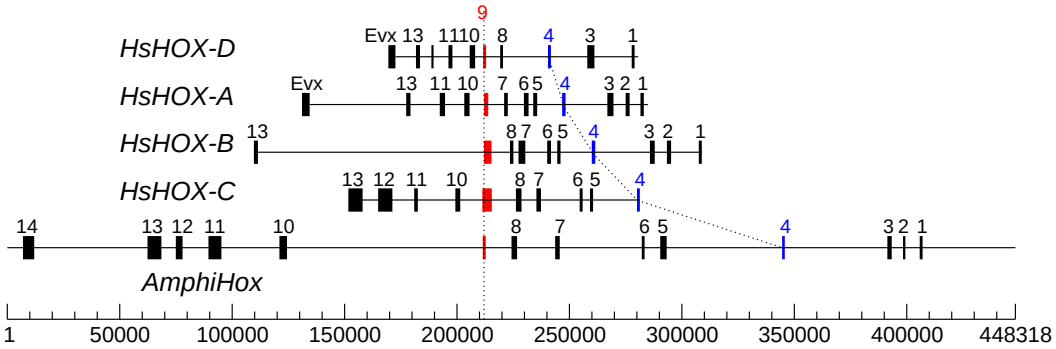

Fig. 2. Comparison of the relative sizes of the sequenced portion of the *Hox* cluster of *B. floridae* and HOX-A, -B, -C and -D of human. The HOX-A and HOX-D contigs also include *EVX* loci upstream of the *HoxA13* and *D13* genes.

tified a *Metaxin2* gene. This gene is also found immediately downstream of human HOX-D cluster on chromosome 2, suggesting that the extensive syntenic blocks at the vertebrate *Hox* loci reported in (Lee *et al.*, 2006) were present in last common ancestor of cephalochordates and vertebrates.

Sequencing of the amphioxus cluster corroborates previous data based on genomic phage chromosomal walking (Garcia-Fernández and Holland, 1994; Ferrier *et al.*, 2000) that the entire region is considerably larger than orthologous vertebrate *Hox* clusters. As shown in Fig. 2, the amphioxus *Hox* cluster is about four times larger than the respective human *Hox* clusters. This same trend holds for all of the vertebrates for which *Hox* cluster sequences have been obtained, including sharks (Kim *et al.*, 2000; Venkatesh *et al.*, 2007), various teleosts (Chiu *et al.*, 2004; Kurosawa *et al.*, 2006; Hoegg *et al.*, 2007), and chicken (Richardson *et al.*, 2007). Conversely, the amphioxus cluster is comparable in size with that of the sea urchin (558kb (Cameron *et al.*, 2006) vs. 448kb in the sequenced region of amphioxus), while it is relatively compact in comparison to several protostome clusters (data compiled e.g. in (Fried *et al.*, 2004)).

3.2 Annotation

Analysis of our contiguous sequence revealed no other genes besides the fourteen *Hox* genes, *AmphiHox1-AmphiHox14*, and corroborates that all the genes are in the same transcriptional orientation (Ferrier *et al.*, 2000; Garcia-Fernández and Holland, 1996) (Fig. 1). The most reliable method to annotate genes is to use transcribed sequences (cDNAs). Unfortunately, only five of the fourteen *Hox* genes had complete cDNA or coding sequences deposited in GenBank (*AmphiHox1-4*, *AmphiHox6*). A partial cDNA of *AmphiHox9* (see Supplemental material) allowed us to determine the sequence of this protein.

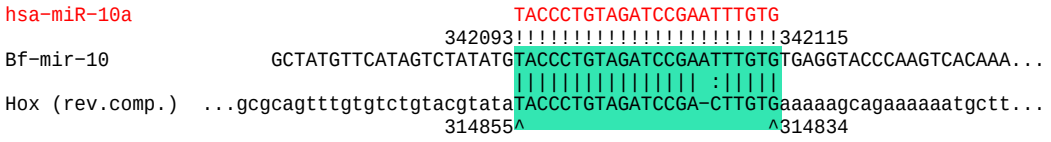

Fig. 3. Putative target site of *mir-10*. The reverse complement of the *Hox* cluster sequence is almost identical to the mature microRNA sequence, i.e., the predicted microRNA could bind to this site with a single short bulge.

For all remaining genes, only the homeobox sequences had been deposited in GenBank; this necessitated the use of *ab initio* methods to predict coding regions. We employed five different exon prediction programs as described above. All of the programs recognized coding regions correctly, however, the only one that did not add extra exons to the predictions, was `GenomeScan`. This is due to the fact that this program employs a training set of "specimen" proteins in order to derive the best possible gene models. `GenomeScan` properly delineated the gene models for the six genes for which complete coding sequences were known, whereas the other algorithms were correct only about half of the time; nonetheless they were very useful for identifying putative transcriptional start sites, exon-intron boundaries and corroborating the `GenomeScan` predictions. `GenomeScan` also properly identified three situations where the homeobox was encoded by two separate exons (*AmphiHox14*, *AmphiHox12*, and *AmphiHox11*).

The intergenic distances between *Hox* genes vary between about 87000nt (*AmphiHox10-AmphiHox9*) and 5000nt (*AmphiHox3-AmphiHox2*). Compared to the vertebrate *Hox* clusters, which are shorter than the amphioxus *Hox* cluster by about a factor of 4, inter-genic regions (IGRs) in the "inner core" (*Hox8-Hox3*) have roughly proportional lengths. In contrast, IGR lengths at both the anterior and posterior ends of the cluster can vary considerably in their relative lengths, mostly due to repeat invasion (Fried *et al.*, 2004), see also Supplemental Material.

The protein coding regions are comparable in size with the vertebrate *Hox* genes. The only prominent differences are the introns of the posterior *AmphiHox14-AmphiHox10* genes, which are much longer than the introns of posterior vertebrate *Hox* genes (Supplemental Material). Note, however, that we do not know whether the genes with the same paralog group designation in amphioxus and vertebrates are true orthologs or whether they arose from an independent expansion of the cluster (Ferrier *et al.*, 2000; Powers and Amemiya, 2004; Campos *et al.*, 2004; Ferrier, 2004). This issue may well be impossible to resolve with phylogenies due to the extent of sequence divergence between the genes in question. However, there is the possibility that more sequence data from lineages in between those of amphioxus and higher vertebrates could help. In particular the Posterior *Hox* sequences from agnathans (hagfish and

lampreys) might be useful, provided the Hox clusters of these basal vertebrate lineages have not themselves undergone independent duplications.

Homology searches of GenBank using the respective *Hox* genes/proteins were carried out to corroborate our peptide predictions, with particular attention to shared blocks of motifs in non-homeodomain regions. As described in the methods section, sequence homology in particular of exon-1 is weak between amphioxus and vertebrates. While there is easily recognizable homology for the anterior genes, only a few alignable blocks can be detected for the posterior genes. This faint sequence similarity renders the exon-1 sequences uninformative for phylogenetic analysis. In particular, they do not resolve the questions regarding the common or separate origin of the posterior *Hox* genes in cephalochordates and vertebrates. Denser taxon coverage might eventually allow us to use presense/absence patterns of such weakly conserved motifs directly for phylogenetic analysis.

Vertebrate *Hox* clusters harbor two unrelated microRNA families, *mir-10* and *mir-196* (Yekta *et al.*, 2004; Tanzer *et al.*, 2005). The putative precursor hairpin of microRNA *mir-10* is located just upstream of *AmphiHox4* (Tanzer *et al.*, 2005). While its expression in amphioxus has not yet been demonstrated, its identification based on sequence similarity is unambiguous, since this microRNA is also widely conserved in invertebrate *Hox* clusters (Tanzer *et al.*, 2005; Hertel *et al.*, 2006). In contrast, no homolog of *mir-196* was found. So far, this family has been reported only in vertebrates.

A `blast` search with the *mir-10* sequence revealed a second blast hit further upstream, between *AmphiHox5* and *AmphiHox4*. The sequence is complementary to the predicted mature *Bf-miR-10*, Fig. 3. This microRNA has turned out to be a "master regulator" within the *Hox* clusters of Drosophilids (Stark *et al.*, 2007). In particular, it is known to regulate *Scr*, the fly *Hox5* homolog (Brennecke *et al.*, 2005). Fig. 3 suggests that *mir-10* regulates *Hox* genes also in amphioxus. Unfortunately the *AmphiHox5* transcripts are not known in detail, hence it presently remains speculation that the target is located within the UTR of *AmphiHox5* mRNA.

A few of the publicly available ESTs map to intergenic regions of *Hox* cluster. Most of them are located in repetitive regions and can be identified as members of several different transposon families. A few ESTs, however map uniquely to the genome, in particular to the IGR between *AmphiHox2* and *AmphiHox1*. None of them has recognizable protein coding capacity. These ESTs indicate that the Amphioxus *Hox* cluster probably harbours *Hox*-associated noncoding transcripts just as its vertebrate counterparts (Mainguy *et al.*, 2007; Rinn *et al.*, 2007).

3.3 Polymorphisms

Pairwise comparison of $\sim$ 150kb of overlapping genomic regions for any two *AmphiHox* alleles (i.e., between different overlapping PAC and BAC clones for the same region) revealed that, in general, around 98% of the nucleotides were identical, with 2% being the result of single nucleotide polymorphisms or small indels. The overlapping region also contains part of the repeat island between *AmphiHox10* and *AmphiHox9*. There, the number and structure of the repeats also varied widely between two different alleles.

The observed level of polymorphism is high in comparison with other chordates, including *Takifugu rubripes* and *Ciona intestinalis*, and leads to substantial genome assembly problems. A comparison of our *Hox* cluster sequence with the corresponding regions of the currently available assembly of the *Branchiostoma floridae* shot gun sequencing shows major discrepancies (see Supplemental Material). We have therefore not attempted to utilize these sequences for assessing polymorphisms.

Within the coding regions for the six complete *Hox* sequences available, nonsynonymous substitutions were detected in three of the genes: *AmphiHox2*, *AmphiHox4*, and *AmphiHox6* (Supplemental Material). None of the substitutions were found within the homeobox for respective genes. The most substitutions were detected for *AmphiHox2*, where 5 and 6 amino acid replacements, respectively, were found relative to our reference *AmphiHox2* sequence.

3.4 Repetitive Elements

In addition to this high level of polymorphism, we also detected two large internally repeated structures, located between *AmphiHox9* and *AmphiHox10*, and between *AmphiHox13* and *AmphiHox14*, respectively. In these regions smaller repeat units were found in both orientations within larger repeat structures (Fig. 4, l.h.s. panel). The repeat regions consist of a diverse collection of repetitive elements, among them members of several transposon families including *Jockey*, *Penelope*, *Harbinger*, and *PiggyBac*.

Overall, the repeat density within the *Hox* cluster is substantially lower than in the surrounding areas, 3.9% versus 13% (Fig. 4, r.h.s. panel), with the bulk of the repeats concentrated in two contiguous regions. Similar to vertebrates, but in contrast to most other invertebrates, the amphioxus *Hox* cluster is thus refractory to the invasion of repetitive elements, albeit less stringently than most vertebrate genomes (Fried *et al.*, 2004; Prohaska *et al.*, 2006).

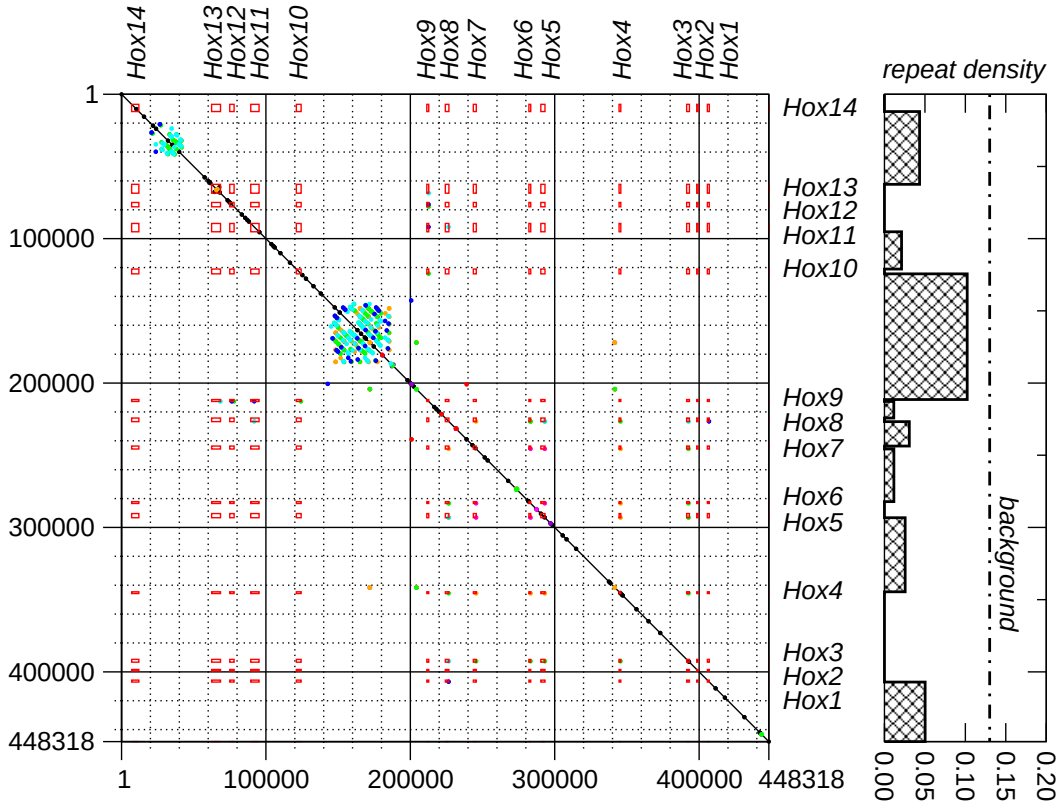

Fig. 4. Dot plot (l.h.s.) created by comparing the amphioxus *Hox* cluster against itself using `blastn` (Altschul *et al.*, 1990). The `blastn` hits are shown color-coded by their *E*-value (black 0, violet 10^{-70}, magenta 10^{-50}, red 10^{-30}, orange 10^{-20}, green 10^{-10}, cyan 1, blue 10. Regions of repetitive sequence can be clearly seen between *Hox14* and *Hox13* as well as between *Hox9* and *Hox10*. Red boxes indicate the pairs of coding regions; blast hits within these boxes correspond to the homeobox sequences. The panel on the r.h.s. displays the fraction of repetitive elements currently annotated in the JGI genome browser for each intergenic region and the background value obtained for the two scaffolds of the `Brafl1` assembly that contains the *Hox* cluster.

3.5 Phylogenetic Footprint Analysis

The fact that the single amphioxus *Hox* cluster is about four times the size of one of the gnathostome *Hox* clusters is striking. Duboule (2007) has proposed a consolidation of the vertebrate *Hox* clusters due to the evolution of long-range, global regulatory mechanisms in the vertebrate lineage. Here we suggest an additional (or alternative) hypothesis in which the initial redundancy between the vertebrate paralogous clusters after they first duplicated was resolved by subfunctionalization at the level of regulatory elements, and cluster size reduction was due to elimination of the degenerate enhancers. The results of the phylogenetic footprinting analysis that has been performed to

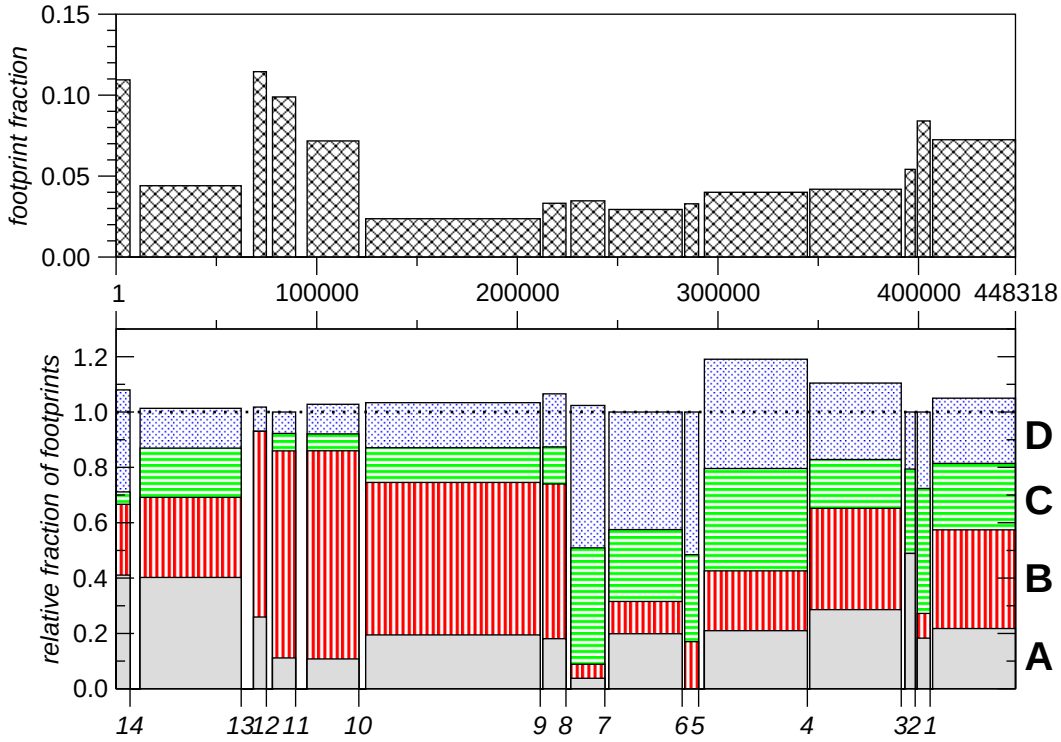

Fig. 5. Phylogenetic footprinting analysis using `tracker`. Top panel: Density of conserved noncoding DNA in the intergenic regions of the amphioxus *Hox* cluster as determined by tracker. Lower panel: Fraction of sequence in footprints that is conserved in at least one of the gnathostome HOX-A, HOX-B, HOX-C, or HOX-D clusters, respectively. Fractions do not add up to 1.0 since a few hits are conserved in more than one cluster. Note that this effect is larger in the anterior part of the cluster.

address this hypothesis are summarized in Fig. 5. Overall, there is little conservation of noncoding DNA between amphioxus and gnathostome *Hox* clusters; in total about 5% of the non-protein-coding DNA can be locally aligned with corresponding regions in at least one gnathostome *Hox* cluster, Fig. 5. Conserved elements are typically short, usually less than 40bp (see Supplemental Material for complete lists).

Surprisingly, there is very little DNA conserved between amphioxus and more than one of the four gnathostome clusters, i.e., the overwhelming majority of the footprints are conserved only in one of the four paralogous vertebrate *Hox* clusters. Rather than interpreting this as a definitive proof for the (almost) complete resolution of redundancy, we suspect that this observation could also be an artifact of the method, which operates at its sensitivity limit on this data set. This was demonstrated by running the phylogenetic footprinting method on varying combinations of cluster types and species. Resulting footprints on the amphioxus sequence had on average less than 10% overlap between different analysis runs. Furthermore, we observed no significant difference between

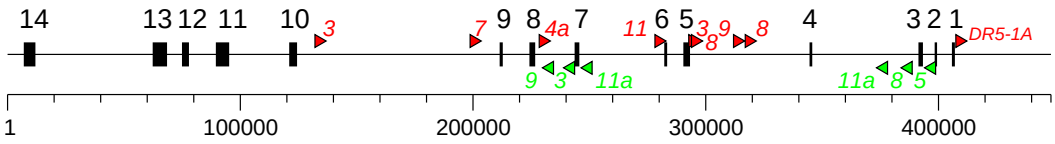

Fig. 6. Distribution of RARE sites in amphioxus *Hox* cluster sequence. Numbers in italics designate the type of the RARE sequence as defined by Mainguy *et al.* (2003), *DR5-1A* was taken from Wada *et al.* (2006). The motif *DR5-3B* described in Wada *et al.* (2006) is of type *8* in the notation of Mainguy *et al.* (2003).

the runs that compared amphioxus with the same gnathostome cluster types from different species, or with the four different clusters of the same species, respectively.

In the same vein, the increased conservation signal in the posterior region, between *Hox13* and *Hox10*, probably is an artifact that arises from the relatively long region between *HoxB13* and *HoxB10* or *HoxB9*, in which *HoxB12*, *HoxB11* and, in some lineages, also *HoxB10*, have been lost. Since `tracker` is operating at its detection limits, we are likely to find more individual signals when using a longer region for comparison. Note, however, that this is not the same as false positives: if the posterior HOX-B region is replaced by a randomly picked stretch of genomic DNA, no signals are found. This explanation is supported by the observation that the overwhelming part of the signal in this region actually comes from conservation between amphioxus and gnathostome HOX-B clusters (Fig. 5, lower panel). In addition, the large differences in AT-content ($\geq 60\%$ in amphioxus, shark, coelocanth, and frog, but $\leq 45\%$ in placental mammals) is a potential problem for the underlying alignment procedure. Since footprints are parts of larger chained alignments, however, they are very unlikely to be just random noise.

The `tracker` footprints were then used as anchors to generate a `dialign` alignments (Morgenstern *et al.*, 2006). A variant of "quartet mapping" (Nieselt-Struwe and von Haeseler, 2001) was then used to investigate whether amphioxus as outgroup can help to resolve the duplication history of the four paralogous gnathostome *Hox* clusters (Bailey *et al.*, 1997): For each species we separately counted the alignment positions supporting one of the three alternative duplication hypothesis. Even though the differences in the counts are significant, different species support different hypotheses: coelacanth supports *(AD)(BC)*, xenopus supports *(AC)(BD)*, and mammals favor *(AB)(CD)*.

We observe a systematic increase in the density of conserved DNA towards the anterior end. In this region we mostly find a fairly even distribution of conservation between the clusters. Also, most of the footprints with conservation in more than one gnathostome paralog are located here.

In chordates, the vitamin A-derived morphogen retinoic acid has a pivotal role during development, reviewed e.g. in (Marlétaz *et al.*, 2006). Fifteen presumptive retinoic-acid responsive elements (RAREs) were identified in the amphioxus *Hox* cluster based on the sequence motifs described in (Mainguy *et al.*, 2003) and (Wada *et al.*, 2006), see Figure 6. Even though RARE sites have been found to be conserved between clusters and among species (Mainguy *et al.*, 2003), none of these falls within phylogenetic footprints that are detectable by `tracker`, because there appears to be no appreciable sequence conservation surrounding the short RARE motives. Likewise, the biological activity of most of these amphioxus RARE sites will require empirical validation.

Conclusions

The amphioxus *Hox* cluster has frequently been described as "archetypal" for the chordate lineage. Indeed, it preserves the ancestral integrity of the cluster and the co-linear arrangement of the *Hox* genes also observed in vertebrates. Importantly, it shares with vertebrates a dramatically reduced density of repetitive elements, while the total size of the cluster is comparable to that of the sea urchin (Cameron *et al.*, 2006) and the few known intact *Hox* clusters of protostomes (see (Fried *et al.*, 2004) for a compilation of data). This implies that the mechanism that prohibits the invasion of repetitive elements into the *Hox* cluster pre-dates the dramatic size reduction observed in the vertebrate *Hox* clusters.

In a recent study of noncoding DNA conservation, Wang *et al.* (2007) found a few conserved noncoding DNA elements in the Pax 1/9 region, but in line with our results for the *Hox* gene cluster, there is much less conservation than among vertebrates, and little conservation that has survived into multiple paralog groups after the 1R/2R genome duplications. In conjunction with the exclusion of repetitive element, this prompts us to speculate that the amphioxus *Hox* cluster might be packed with functional sequences that have largely been distributed among the four vertebrate clusters. Unfortunately most of the noncoding sequence of the amphioxus *Hox* cluster is not alignable to the vertebrate sequences, so that a direct test of this hypothesis in not possible. It is at least consistent with the data from the `tracker` analysis, however. It is interesting to note in this context that the intron lengths at least of *AmphiHox9* and the non-posterior (*AmphiHox8-AmphiHox1*) genes remain essentially unchanged in vertebrates. Intriguingly, the RARE sites are largely concentrated in this region. Analysis of the activity of noncoding sequences and subsequent comparison with data from vertebrate *Hox* clusters will be necessary in order to assess the degree of conservation of biological function.

In general, we observe a clear trend towards more conservation at the anterior end of the *Hox* cluster. This is true for both coding and noncoding sequence. In fact, for the posterior genes *AmphiHox14-AmphiHox10* it remains uncertain whether they are true orthologs of vertebrate PG14-PG10 genes, or whether the have a different duplication history. It is worth mentioning in this context that the two large repetitive regions are also found between posterior genes (*AmphiHox14-AmphiHox13* and *AmphiHox10-AmphiHox9*, resp). The newly sequenced exon-1 data fail to help resolve this issue. Taken together, the data suggest that at least the posterior end of the amphioxus *Hox* cluster is highly derived.

Acknowledgments

We thank Sven Findeiß (U. Leipzig) for computational assistance. The *AmphiHox9* partial cDNA was prepared by JGF while in Peter Holland's lab. This work was funded, in part, from grants from the National Science Foundation to CTA (IOS-0321461, MCB-0719558), the Ministerio de Educación y Ciencia, Spain, to JGF (BFU2005-00252), the BBSRC to DEKF, and the Bioinformatics Initiative of the Deutsche Forschungsgemeinschaft to PFS. JPA holds a Generalitat de Catalunya fellowship. The `brafl1` sequence data were produced by the US Department of Energy Joint Genome Institute and were downloaded from their website `http://www.jgi.doe.gov/`.

Note Added in Proof

After our analysis was complete, we became aware that — in contrast to previous reports (Minguillón *et al.*, 2005) — there is a *AmphiHox15* gene in the region between *AmphiHox14* and *AmphiExvA/EvxB* (Holland, L.Z. *et al.* ***). This discovery does not influence our results because our analysis is almost exclusively concerned with a rather detailed comparison of the amphioxus *Hox* cluster with the vertebrate clusters, and no ortholog of *AmphiHox15* has yet been found in a vertebrate. We have therefore decided not to include any sequence data from the (as yet unpublished) Amphioxus genome project in our assembly.

References

Abramovich C, Humphries RK, 2005. Hox regulation of normal and leukemic hematopoietic stem cells. Curr Opin Hematol 12:210–216.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ, 1990. Basic local alignment search tool. J Mol Biol 215:403–410.

Amemiya CT, Ota T, Litman GW, 1996. Construction of P1 artificial chromosome (PAC) libraries from lower vertebrates. In: Lai E, Birren B, editors, Analysis of Nonmammalian Genomes, (pp. 223–256). San Diego: Academic Press.

Bailey WJ, Kim J, Wagner G, Ruddle FH, 1997. Phylogenetic reconstruction of vertebrate Hox cluster duplications. Mol Biol Evol 14:843–853.

Besemer J, Borodovsky M, 2005. `GeneMark`: web software for gene finding in prokaryotes, eukaryotes and viruses. Nucleic Acids Res 33:W451–W454.

Brennecke J, Stark A, Russell R, Cohen S, 2005. Principles of microRNA-target recognition. PLoS Biol 3:e85.

Burge CB, Karlin S, 1997. Prediction of complete gene structures in human genomic DNA. J Mol Biol 268:78–94.

Burge CB, Karlin S, 1998. Finding the genes in genomic DNA. Curr Opin Struct Biol 8:346–354.

Cameron RA, Rowen L, Nesbitt R, Bloom S, Rast JP, Berney K, Arenas-Mena C, Martinez P, Lucas S, Richardson PM, Davidson EH, Peterson KJ, Hood L, 2006. Unusual gene order and organization of the sea urchin Hox cluster. J Exp Zoolog B Mol Dev Evol 306:45–58.

Campos PRA, de Olivera VM, Wagner GP, Stadler PF, 2004. Gene phylogenies and protein-protein interactions: Possible artifacts resulting from shared protein interaction partners. J Theor Biol 231:197–202.

Capecchi MR, 1997. Hox genes and mammalian development. Cold Spring Harb Symp Quant Biol 62:273–281.

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 F, Ledje C, Bartsch P, Scemama JL, Stellwag E, Fried C, Prohaska SJ, Stadler PF, Amemiya CT, 2004. Bichir *HoxA* cluster sequence reveals surprising trends in rayfinned fish genomic evolution. Genome Res 14:11–17.

de Rosa R, Grenier Jennifer K.and Andreeva T, Cook CE, Adoutte A, Akam M, Carroll SB, Balavoine G, 1999. Hox genes in brachiopods and priapulids and protostome evolution. Nature 399:772–776.

Delsuc F, Brinkmann H, Chourrout D, Philippe H, 2006. Tunicates and not cephalochordates are the closest living relatives of vertebrates. Nature 439:923–924.

Duboule D, 2007. The rise and fall of *Hox* gene clusters. Development 134:2549–2560.

Duboule D, Deschamps J, 2004. Colinearity loops out. Dev Cell 6:738–740.

Eklund EA, 2006. The role of HOX genes in myeloid leukemogenesis. Curr Opin Hematol 13:67–73.

Ferrier DEK, 2004. Hox genes: Did the vertebrate ancestor have a Hox14? Curr Biol 14:R210R211.

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

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

Fried C, Prohaska SJ, Stadler PF, 2004. Exclusion of repetitive dna elements from gnathostome *Hox* clusters. J Exp Zool Mol Dev Evol 302B:165–173.

Fritzsch G, Böhme MU, Thorndyke M, Nakano H, Israelsson O, Stach T, Schlegel M, Hankeln T, Stadler Peter F, 2007. A pcr survey of *Xenoturbella bocki Hox* genes. J Exp Zool Mol Dev Evol In press.

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

Garcia-Fernández J, Holland PW, 1996. Amphioxus Hox genes: insights into evolution and development. Int J Dev Biol Suppl 1 (pp. 71S–72S).

Gellon G, McGinnis W, 1998. Shaping animal body plans in development and evolution by modulation of Hox expression patterns. Bioessays 20:116–125.

Gould A, Morrison A, Sproat G, White RA, Krumlauf R, 1997. Positive cross-regulation and enhancer sharing: two mechanisms for specifying overlapping Hox expression patterns. Genes Dev 11:900–913.

Hara Y, Yamaguchi M, Akasaka K, Nakano H, Nonaka M, Amemiya S, 2006. Expression patterns of hox genes in larvae of the sea lily *Metacrinus rotundus*. Dev Genes Evol 216:797–809.

Hertel J, Lindemeyer M, Missal K, Fried C, Tanzer A, Flamm C, Hofacker IL, Stadler PF, The Students of Bioinformatics Computer Labs 2004 and 2005, 2006. The expansion of the metazoan microRNA repertoire. BMC Genomics 7:15 [epub].

Hoegg S, Boore JL, Kuehl JV, Meyer A, 2007. Comparative phylogenomic analyses of teleost fish *Hox* gene clusters: lessons from the cichlid fish *Astatotilapia burtoni*. BMC Genomics 8:317.

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

Ikuta T, Saiga H, 2005. Organization of hox genes in ascidians: present, past, and future. Dev Dyn 233:382–389.

International Human Genome Sequencing Consortium, 2001. Initial sequencing and analysis of the human genome. Nature 409:860–921.

Kim CB, Amemiya C, Bailey W, Kawasaki K, Mezey J, Miller W, Minosima 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.

Kohany O, Gentles AJ, Hankus L, Jurka J, 2006. Annotation, submission and screening of repetitive elements in `Repbase`: `RepbaseSubmitter` and `Censor`. BMC Bioinformatics 7:474.

Krumlauf R, 1992. Evolution of the vertebrate *Hox* homeobox genes. Bioessays 14:245–252.

Kurosawa G, Takamatsu N, Takahashi M, Sumitomo M, Sanaka E, Yamada

K, Nishii K, Matsuda M, Asakawa S, Ishiguro H, Miura K, Kurosawa Y, Shimizu N, Kohara Y, Hori H, 2006. Organization and structure of *hox* gene loci in medaka genome and comparison with those of pufferfish and zebrafish genomes. Gene 370:75–82.

Lee AP, Koh EGL, Tay A, Brenner S, Venkatesh B, 2006. Highly conserved syntenic blocks at the vertebrate Hox loci and conserved regulatory elements within and outside Hox gene clusters. Proc Natl Acad Sci USA 103:6994–6999.

Lee PN, Callaerts P, de Couet HG, Martindale MQ, 2003. Cephalopod *Hox* genes and the origin of morphological novelties. Nature 424:1061–1065.

Longhurst TJ, Joss JM, 1999. Homeobox genes in the australian lungfish, *Neoceratodus forsteri*. J Exp Zool 285:140–145.

Lynch VJ, Roth JJ, Takahashi K, Dunn CW, Nonaka DF, Stopper GF, Wagner GP, 2004. Adaptive evolution of hoxa-11 and hoxa-13 at the origin of the uterus in mammals. Proc Biol Sci 271:2201–2207.

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

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

Manuel M, Jager M, Murienne J, Clabaut C, Le Guyader H, 2006. *Hox* genes in sea spiders (Pycnogonida) and the homology of arthropod head segments. Dev Genes Evol 216:481–491.

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

Marlétaz F, Holland LZ, Laudet V, Schubert M, 2006. Retinoic acid signaling and the evolution of chordates. Int J Biol Sci 2:38–47.

Mayor C, Brudno M, Schwartz JR, Poliakov A, Rubin EM, Frazer KA, Pachter LS, Dubchak I, 2000. `VISTA`: visualizing global DNA sequence alignments of arbitrary length. Bioinformatics 16:1046–1047.

Minguillón C, Gardenyes J, Serra E, Castro LFC, Hill-Force A, Holland PW, Amemiya CT, Garcia-Fernàndez J, 2005. No more than 14: the end of the amphioxus *Hox* cluster. Int J Biol Sci 1:19–23.

Monteiro AS, Ferrier DEK, 2006. Hox genes are not always colinear. Int J Biol Sci 2:95–103.

Morgenstern B, Prohaska SJ, Pohler D, Stadler PF, 2006. Multiple sequence alignment with user-defined anchor points. Algo Mol Biol 1:6.

Nieselt-Struwe K, von Haeseler A, 2001. Quartet-mapping, a generalization of the likelihood mapping procedure. Mol Biol Evol 18:1204–1219.

Osoegawa K, Woon PY, Zhao B, Frengen E, Tateno M, Catanese JJ, de Jong PJ, 1998. An improved approach for construction of bacterial artificial chromosome libraries. Genomics 52:1–8.

Parra G, Blanco E, Guigó R, 2000. `GeneID` in *Drosophila*. Genome Res 10:511–515.

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

Peterson KJ, 2004. Isolation of *Hox* and *Parahox* genes in the hemichordate *Ptychodera flava* and the evolution of deuterostome *Hox* genes. Mol Phylogenet Evol 31:1208–1215.

Podlasek C, Houston J, McKenna KE, McVary KT, 2002. Posterior *Hox* gene expression in developing genitalia. Evol Dev 4:142–163.

Popodi E, Kissinger JC, Andrews ME, Raff RA, 1996. Sea urchin *Hox* genes: insights into the ancestral *Hox* cluster. Mol Biol Evol 13:1078–1086.

Powers TP, Amemiya CT, 2004. Evidence for a *Hox14*, paralog group in vertebrates. Curr Biol 14:R183–R184.

Prohaska S, Fried C, Flamm C, Wagner G, Stadler PF, 2004. Surveying phylogenetic footprints in large gene clusters: Applications to Hox cluster duplications. Mol Phyl Evol 31:581–604.

Prohaska SJ, Stadler PF, Wagner GP, 2006. Evolutionary genomics of *Hox* gene clusters. In: Papageorgiou S, editor, HOX Gene Expression, (pp. 68–90). New York: Landes Bioscience & Springer.

Richardson MK, Crooijmans RP, Groenen MA, 2007. Sequencing and genomic annotation of the chicken (*Gallus gallus*) *Hox* clusters, and mapping of evolutionarily conserved regions. Cytogenet Genome Res 117:110–119.

Rinn JL, Kertesz M, Wang JK, Squazzo SL, Xu X, Brugmann SA, Goodnough LH, Helms JA, Farnham PJ, Segal E, Chang HY, 2007. Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. Cell 29:1311–1323.

Schwartz S, Zhang Z, Frazer KA, Smit A, Riemer C, Bouck J, Gibbs R, Hardison R, , Miller W, 2000. `PipMaker` — a web server for aligning two genomic DNA sequences. Genome Research 4:577–586.

Seo HC, *al.*, 2004. *Hox* cluster disintegration with persistent anteroposterior order of expression in *Oikopleura dioica*. Nature 431:67–71.

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

Stadler PF, Fried C, Prohaska SJ, Bailey WJ, Misof BY, Ruddle FH, Wagner GP, 2004. Evidence for independent *Hox* gene duplications in the hagfish lineage: A PCR-based gene inventory of *Eptatretus stoutii*. Mol Phylog Evol 32:686–692.

Stark A, Kheradpour P, Parts L, Brennecke J, Hodges E, Hannon GJ, Kellis M, 2007. Systematic discovery and characterization of fly microRNAs using 12 *Drosophila* genomes. Genome Res 17:1865–1879.

Tanzer A, Amemiya CT, Kim CB, Stadler PF, 2005. Evolution of microRNAs located within *Hox* gene clusters. J Exp Zool Mol Dev Evol 304B:75–85.

Uberbacher EC, Xu Y, Mural RJ, 1996. Discovering and understanding genes in human DNA sequence using GRAIL. Methods Enzymol 266:259–281.

Venkatesh B, Kirkness EF, Loh YH, Halpern AL, Lee AP, Johnson J, Dandona N, Viswanathan LD, Tay A, Venter JC, Strausberg RL, Brenner S, 2007. Survey sequencing and comparative analysis of the elephant shark (*Callorhinchus milii*) genome. PLoS Biol 5:e101.

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. Develop Dynamics 235:1522–1531.

Wagner GP, Lynch VJ, 2005. Molecular evolution of evolutionary novelties: the vagina and uterus of therian mammals. J Exp Zoolog B Mol Dev Evol 304:580–592.

Wang W, Zhong J, Su B, Zhou Y, Wang YQ, 2007. Comparison of *Pax1/9* locus reveals 500-myr-old syntenic block and evolutionary conserved noncoding regions. Mol Biol Evol 24:784–791.

Yekta S, Shih Ih, Bartel DP, 2004. MircoRNA-directed cleavage of *HoxB8* mRNA. Science 304:594–596.

Yu JK, Satou Y, Holland ND, Shin-i T, Kohara Y, Satoh N, Bronner-Fraser M, Holland L, 2007. Axial patterning in cephalochordates and the evolution of the organizer. Nature 445:613–617.

Zakany J, Duboule D, 2007. The role of *Hox* genes during vertebrate limb development. Curr Opin Genet Dev 17:359–366.