Update TRENDS in Genetics Vol.21 No.8 August 2005

Research Focus

Hox clusters as models for vertebrate genome evolution

Simone Hoegg and Axel Meyer

Lehrstuhl für Zoologie und Evolutionsbiologie, Department of Biology, University of Konstanz, 78457 Konstanz, Germany

The surprising variation in the number of Hox clusters and the genomic architecture within vertebrate lineages, especially within the ray-finned fish, reflects a history of duplications and subsequent lineage-specific gene loss. Recent research on the evolution of conserved non-coding sequences (CNS) in Hox clusters promises to reveal interesting results for functional and phenotypic diversification.

Hox genes – quo vadis?

Hox genes are arranged in clusters on chromosomes and, as transcription factors, have a crucial role during development. They determine the positional specification of the anterior–posterior axis and are, in most cases, expressed in a 'colinear' fashion (i.e. genes that are anterior in the Hox clusters are expressed early and in the anterior part of the embryo, whereas genes that are posterior in the clusters are expressed later and towards the posterior of the embryo).

Derived vertebrates have multiple clusters: there are four in tetrapods, up to eight in ray-finned fish and ~14 in tetraploid salmonid species [1]. They originated by duplication of a single ancestral cluster during two rounds (the 2R hypothesis) of genome-duplication events that occurred early in the evolution of chordates and vertebrates.

Comparative studies on Hox cluster evolution among the ~25 000 species of fish have, so far, mainly focussed on gene numbers obtained through PCR-based screens. These studies revealed important insights, and originally suggested super-numeral (relative to the expected number of four) Hox clusters. But there is more to Hox genes than just numbers of genes and clusters. Recently, data from genome projects [2,3], in addition to studies that employ large-insert genomic libraries (i.e. BACs and PACs) [4–6], permitted analyses of significant genomic stretches that included introns and intergenic non-coding sequences in Hox clusters. Comparisons of this 'non-coding' DNA showed that it contains a surprising number of putative conserved regulatory elements. We would like to draw attention to the insights that these comparative genomic analyses offer.

Hox-cluster evolution in vertebrates

Although all known tetrapod clusters consist of genes that can be assigned to 13 paralogy groups (PGs), a recent study found evidence for the existence of *Hox*14 genes in the *Hox*A and *Hox*D clusters in shark and coelacanth [6–8] (Figure 1). Because shark *Hox*D14 and the coelacanth *Hox*A14 genes are more similar to each other than to any other Hox gene, it can be assumed that *Hox*14 genes were lost independently in the tetrapod-stem lineage after the divergence of the coelacanth and in the lineage that led to ray-finned fish. Analyses of complete *Hox*A clusters from derived vertebrates failed to detect an additional gene between *Evx1* (encoding even-skipped homeobox homolog 1) and *Hox*A13 [6,9] (Figure 1).

It had been assumed that the land vertebrates (the Hox clusters in human and mouse served as incomplete evidence for this) were identical in terms of numbers of clusters (i.e. they have four clusters), their architecture and total gene content. However, unpublished results from the frog genome (*Xenopus tropicalis*, http://genome.jgi-psf.org/Xentr3/Xentr3.home.html) showed that some variation exists because it lacks two genes (*Hox*B13 and *Hox*D12) that are present in mouse and human (Figure 1).

The fish-specific genome duplication (3R) and Hox-cluster evolution

Recent data from genome projects on ray-finned fish (zebrafish, medaka and two species of pufferfish), which are at various stages of completion, have shown that they have more Hox clusters than tetrapods (Figure 1). These extranumeral Hox clusters result from a genome duplication event that is specific for the fish (actinopterygian) lineage: the fish-specific genome duplication (FSGD or 3R). In zebrafish (*Danio rerio*), a set of seven Hox clusters have been described: two *Hox*A, two *Hox*B, two *Hox*C and one *Hox*D cluster [10]. Seven clusters were subsequently described in two pufferfish species (*Takifugu rubripes* and *Tetraodon nigroviridis*); however, it has been suggested that *T. rubripes* contains a third *Hox*A cluster [3,4]. In contrast to the situation in zebrafish, both pufferfish have duplicated *Hox*D clusters but only a single copy of the *Hox*C cluster (Figure 1). In addition, data from medaka (*Oryzias latipes*) show evidence of one *Hox*C cluster and duplicated *Hox*A, *Hox*B and *Hox*D clusters [11]. The loss of the second *Hox*C cluster might be a shared feature of the Neoteleostei, the 'modern' ray-finned fish that comprise most of the fish model systems (e.g. pufferfish, medaka, cichlids, platies and swordtails, but not zebrafish). More data will show if this hypothesis is correct. Studies of Hox genes in a basal actinopterygian fish, for example, in the bichir (*Polypterus senegalus*) showed that its genome is in

Corresponding author: Meyer, A. (axel.meyer@uni-konstanz.de).
Available online 20 June 2005

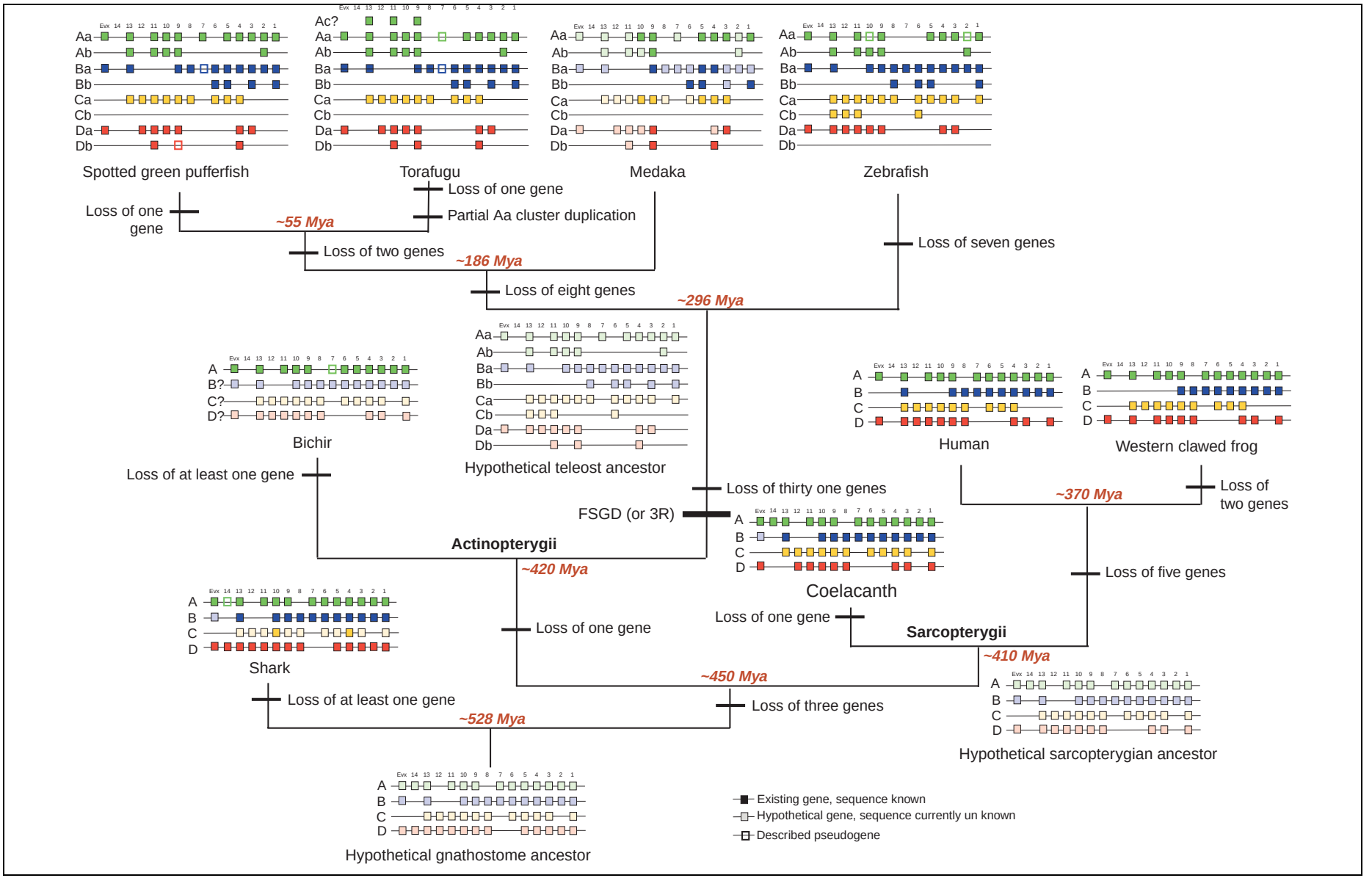

Figure 1. The hypothesis on the inferred Hox cluster evolution within the jawed vertebrates (including data from Refs [3,4,6]). A hypothetical gnathostome ancestor with four clusters [including genes from PGs 1–14 and even-skipped homeobox homologs (*Evx1*)]; the most likely deduced architecture is shown. The gene content of the eight Hox gene clusters of the inferred hypothetical teleost ancestor and the four Hox clusters of the hypothetical sacropterygian are shown (all three hypothetical ancestral genomic states are shown in faded colours). Sharks, tetrapods and basal ray-finned fish such as bichirs (and most likely also sturgeons, gars and bowfins) still maintained a four-cluster state, whereas more derived teleost fish (including the osteoglossomorphs Ref. [13]) underwent an additional duplication (FSGD or 3R), initially resulting in eight Hox clusters. This probably occurred shortly after the FSGD individual Hox genes were lost, which led to a total of seven clusters in most modern fish with different gene content. Closed squares indicate genes that have been previously described and open squares indicate reported pseudogenes. Shaded squares are genes that have not been sequenced yet, but probably are present in the cluster. This is the case for the complete *Hox*B, *Hox*C and *Hox*D clusters of the bichir, which have not been described yet, but do exist based on data from a PCR screen [12]. Data from medaka (*Oryzias latipes*) are based on a combination of PCR screen and mapping results [11]. Therefore, linkage was determined but the complete sequences still have not been published. Abbreviation: Mya, million years ago.

a presumed pre-3R pre-duplication condition, both in terms of the number of Hox genes that were identified by a PCR screen [12] and the structure of the *Hox*A cluster [5]. The 3R duplication is likely to have occurred after polypterids branched off from the actinopterygian fish-stem lineage. Therefore, not all recent ray-finned fish are derived from a fish ancestor whose genome was duplicated. The more exact phylogenetic timing of the FSGD was deduced from data sets of other duplicated genes [13], suggesting that the genome duplication occurred later in the fish lineage. Interestingly, all of the basal lineages of fish that branched off from the fish stem-lineage before the 3R event are 'species-poor'. This observation and earlier analyses led to the suggestion that the FSGD and biodiversity of fish might be causally related ([13]; and references therein). More complete studies of Hox clusters in basal actinopterygian lineages such as bichir, bowfin and osteoglossomorphs are required and will help in the reconstruction of major genomic events early in the evolution of fish and tetrapods.

Evolution of non-coding sequences in gnathostome Hox clusters

Hox clusters provide a good model system for genomic comparisons of vertebrates, because they define a specific stretch of DNA as a result of their highly conserved cluster structure. Rearrangements and gene loss complicate studies in non-Hox gene families, but a complete genome analysis of the *Tetraodon* genome increases support for FSGD [3]. However, not only is the structure of Hox clusters evolutionarily conserved, and possibly constraint, but also there appears to be strong selection against the invasion or spreading of repetitive elements [e.g. short interspersed nuclear elements (SINEs), long interspersed nuclear elements (LINEs), long terminal repeats (LTRs) and DNA transposons] in Hox clusters [3,14]. Gene loss, however, is also often accompanied by the invasion of those repetitive elements. In invertebrates, Hox-cluster structure is less conserved and there is no difference between the number and length of the repetitive sequences within a cluster and those in the surrounding sequences [14].

The compactness of the clusters made it possible to compare, for example, the available *Hox*A-cluster sequences from shark with those of tetrapods and several other teleost species. The first studies of this kind on Hox clusters used an algorithm based on multiple sequence alignments, and showed not only that previously known regulatory elements can be identified, but also that many more conserved non-coding sequences (CNS) can be identified, at least some of which are probably novel cis-regulatory elements [9,15] (Box 1). This technique of identifying conserved non-coding elements by comparing homologous sequences from different species is called 'phylogenetic footprinting'.

Recently, new software (Tracker) has been developed by Prohaska and colleagues [16] that can identify corresponding footprints in long sequences from multiple species. Testing this software on the data set of Hox genes, used in a previous study [15], they [16] determined that Tracker can identify the almost complete list of phylogenetic footprint clusters (PFCs), and that it is much faster than the previous web-based tools. Tracker has also been used to compare the *Hox*N cluster of the shark *Heterodontus francisci* with the Hox clusters of other known vertebrates (human, rat and pufferfish) [17]. Interestingly, the shark *Hox*N cluster has the greatest length of shared PFCs compared with the *Hox*D clusters of other species, which indicates a homology relationship that was impossible to make based on the similarities of the amino acid sequences of the Hox proteins alone. Another study involving this new program involves the *Hox*A cluster of the bichir (*Polypterus senegalus*) – the most basal extant ray-finned fish [5]. The analysis of co-occurring PFCs in bichir, shark, human and in duplicated teleost A-clusters suggests that bichir has only four clusters. Conserved non-coding nucleotides (CNCNs), as identified by Tracker, can also be used for estimates of evolutionary rates [18]. A tetrapod comparison showed a constant evolutionary rate within the mammals, whereas the western clawed frog (*Xenopus tropicalis*) had an increased rate of modifications of CNCN positions. In fish, duplicated clusters have different evolutionary rates that are consistent in genes and their surrounding non-coding sequences [19].

Box 1. Definitions of conserved elements

Phylogenetic footprints (PFs): short blocks of non-coding DNA sequences (≥6 bp), which are conserved in taxa that have an additive evolutionary time of at least 250 million years [20].

Phylogenetic footprint clusters (PFCs): two-to-thirteen PFs that are located within 200 bp of each other [21] (Figure I).

Conserved non-coding nucleotides (CNCNS): concatenated sequences of PFs from a comparison of two outgroup species. This implies conservation over a larger evolutionary distance.

Conserved non-coding sequences (CNS): these sequences have ≥70% identity over at least 100 bp in human and mouse genomes [22], (for more details, see Ref. [9]).

```
HsA7-6-a  ATGGGGAAAGGGTCATAAATCCGTTGTT-G
HfA7-6-a  ATGGGGAAATG-TCATAAATCCGTTGTT-G
MsA7-6-a  ------------TCATAAATCCGTTGTTCG
```

TRENDS in Genetics

Figure I. Conserved sequence in the intergenic region between *Hoxa*7 and *Hoxa*6 from human (Hs), shark (Hf) and striped bass (Ms). Data are from Ref. [21].

Concluding remarks

The newly determined genomes combined with new analytical tools for identifying conserved elements from multiple clusters provides many new possibilities for the evaluation of genomic data from different organisms. This is especially true with respect to the testing of models of regulatory evolution (e.g. subfunctionalization) following duplication events. The comparative study of the evolution and function of conserved non-coding sequences in Hox clusters promises to yield important insights for the functional and phenotypic diversification of vertebrate genomes more generally.

Acknowledgements
We thank the Deutsche Forschungsgemeinschaft for financial support. S.H. was supported by a grant of the Landesgraduiertenförderung Baden-Württemberg.

References

1 H.K. Moghadam *et al.* Organization of Hox clusters in rainbow trout (*Oncorhynchus mykiss*): a tetraploid model species. *J. Mol. Evol.* (in press)
2 Aparicio, S. *et al*. (2002) Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. *Science* 297, 1301–1310
3 Jaillon, O. *et al*. (2004) Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. *Nature* 431, 946–957
4 Amores, A. *et al*. (2004) Developmental roles of pufferfish Hox clusters and genome evolution in ray-fin fish. *Genome Res.* 14, 1–10
5 Chiu, C-h. *et al*. (2004) Bichir *Hox*A cluster sequence reveals surprising trends in ray-finned fish genomic evolution. *Genome Res.* 14, 11–17
6 Powers, T.P. and Amemiya, C.T. (2004) Evolutionary plasticity of vertebrate Hox genes. *Curr. Genomics* 5, 459–472
7 Powers, T.P. and Amemiya, C.T. (2004) Evidence for a *Hox*14 paralog group in vertebrates. *Curr. Biol.* 14, R183–R184
8 Garcia-Fernàndez, J. (2005) Hox, ParaHox, ProtoHox: facts and guesses. *Heredity* 94, 145–152
9 Santini, S. *et al*. (2003) Evolutionary conservation of regulatory elements in vertebrate hox gene clusters. *Genome Res.* 13, 1111–1122
10 Amores, A. *et al*. (1998) Zebrafish hox clusters and vertebrate genome evolution. *Science* 282, 1711–1714
11 Naruse, K. *et al*. (2000) A detailed linkage map of medaka, *Oryzias latipes*: comparative genomics and genome evolution. *Genetics* 154, 1773–1784
12 Ledje, C. *et al*. (2002) Characterization of Hox genes in the bichir, *Polypterus palmas*. *J. Exp. Zool.* 294, 107–111
13 Hoegg, S. *et al*. (2004) Phylogenetic timing of the fish-specific genome duplication correlates with the diversification of teleost fish. *J. Mol. Evol.* 59, 190–203
14 Fried, C. *et al*. (2004) Exclusion of repetitive DNA elements from gnathostome Hox clusters. *J. Exp. Zoolog. B Mol. Dev. Evol.* 302, 165–173
15 Chiu, C-h. *et al*. (2002) Molecular evolution of the *Hox*A cluster in the three major gnathostome lineages. *Proc. Natl. Acad. Sci. U. S. A.* 99, 5492–5497
16 Prohaska, S.J. *et al*. (2004) Surveying phylogenetic footprints in large gene clusters: applications to Hox cluster duplications. *Mol. Phylogenet. Evol.* 31, 581–604
17 Prohaska, S.J. *et al*. (2004) The shark *Hox*N cluster is homologous to the human *Hox*D cluster. *J. Mol. Evol.* 58, 212–217
18 Wagner, G.P. *et al*. (2004) Divergence of conserved non-coding sequences: rate estimates and relative rate tests. *Mol. Biol. Evol.* 21, 2116–2121
19 Wagner, G.P. *et al.* Molecular evolution of duplicated ray-finned fish *Hox*A clusters: increased synonymous substitution rate and asymmetrical co-divergence of coding and non-coding sequences. *J. Mol. Evol.* (in press)
20 Tagle, D.A. *et al*. (1988) Embryonic epsilon and gamma globin genes of a prosimian primate (*Galago crassicaudatus*). Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. *J. Mol. Biol.* 203, 439–455
21 Chiu, C-h. *et al*. (2002) Molecular evolution of the *Hox*A cluster in the three major gnathostome lineages. *Proc. Natl. Acad. Sci. U. S. A.* 99, 5492–5497
22 Loots, G.G. *et al*. (2000) Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science* 288, 136–140

doi:10.1016/j.tig.2005.06.004

Discovering functional relationships: biochemistry versus genetics

Sharyl L. Wong, Lan V. Zhang and Frederick P. Roth

Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, 250 Longwood Ave, Boston, MA, 02115 USA

Biochemists and geneticists, represented by Doug and Bill in classic essays, have long debated the merits of their methods. We revisited this issue using genomic data from the budding yeast, *Saccharomyces cerevisiae*, and found that genetic interactions outperformed protein interactions in predicting functional relationships between genes. However, when combined, these interaction types yielded superior performance, convincing Doug and Bill to call a truce.

Introduction

For more than ten years, Doug, a retired biochemist, and Bill, a retired geneticist, have lived on a hill overlooking a car factory, debating their strategies for reverse engineering a car (see: http://www2.biology.ualberta.ca/locke.hp/dougandbill.htm). Doug advocated rolling up his sleeves, getting under the hood and determining how the parts fit together. Bill preferred tying the hands of a different car-factory worker each morning, then relaxing with a cup of coffee and later examining the cars that emerged from the factory.

One day, Doug and Bill strolled over the next hill. In the midst of debate, they encountered Sharyl, a graduate student in computational genomics. Having overheard their debate, she interjected, 'I don't know much about cars, but I detect an analogy to biochemistry and genetics. I'm trying to discover functional relationships between genes and proteins in yeast and I wonder which of your strategies would work best.'

Corresponding author: Roth, F.P. (fritz_roth@hms.harvard.edu).
Available online 27 June 2005