Breakup of a homeobox cluster after genome duplication in teleosts

John F. Mulley*, Chi-hua Chiu†, and Peter W. H. Holland*‡

*Department of Zoology, University of Oxford, South Parks Road, Oxford, OX1 3PS, United Kingdom; and †Rutgers University, Department of Genetics, Life Sciences Building, Piscataway, NJ 08854

Edited by Eric H. Davidson, California Institute of Technology, Pasadena, CA, and approved May 11, 2006 (received for review January 13, 2006)

Several families of homeobox genes are arranged in genomic clusters in metazoan genomes, including the Hox, ParaHox, NK, Rhox, and Iroquois gene clusters. The selective pressures responsible for maintenance of these gene clusters are poorly understood. The ParaHox gene cluster is evolutionarily conserved between amphioxus and human but is fragmented in teleost fishes. We show that two basal ray-finned fish, *Polypterus* and *Amia*, each possess an intact ParaHox cluster; this implies that the selective pressure maintaining clustering was lost after whole-genome duplication in teleosts. Cluster breakup is because of gene loss, not transposition or inversion, and the total number of ParaHox genes is the same in teleosts, human, mouse, and frog. We propose that this homeobox gene cluster is held together in chordates by the existence of interdigitated control regions that could be separated after locus duplication in the teleost fish.

Amia | gene cluster | genome evolution | ParaHox | *Polypterus*

The ParaHox gene cluster, discovered in the cephalochordate amphioxus (1), comprises three physically linked homeobox genes: *Gsx*, *Xlox*, and *Cdx*. Each of these homeobox gene families has been found in a diversity of protostomes and deuterostomes, implying that the genes, and by extrapolation the ParaHox gene cluster, date at least to the base of the Bilateria (2). ParaHox genes are closely related to Hox genes and have been implicated in a range of developmental processes, including brain patterning (Gsx genes; refs. 1, 3, and 4); specification of the vertebrate pancreas and adjacent endodermal regions (Xlox, also called *pdx1* or *Ipf1*; refs. 1 and 5); and development of the anus, posterior gut, and posterior neural tube (Cdx or *caudal*; refs. 1, 6, and 7).

The existence of a mammalian ParaHox gene cluster was first inferred from chromosomal mapping data in human and mouse (8) and later confirmed and described through analysis of complete genome sequences (9). Preliminary analysis of the *Xenopus tropicalis* genome (http://genome.jgi-psf.org/Xentr4/Xentr4.home.html) also reveals a ParaHox cluster (data not shown). The human ParaHox gene cluster maps to chromosome 13 at position 13q12.1 and comprises the human *GSH1*, *IPF1* (=*PDX1*), and *CDX2* homeobox genes, orthologous to the *Gsx*, *Xlox*, and *Cdx* genes of amphioxus. The ParaHox gene clusters of mouse and *Xenopus* have precisely the same constitution. In addition, *CDX1* and *CDX4* (two paralogues of *CDX2*), plus *GSH2* (a paralogue of *GSH1*), occur on other chromosomes in human, mouse, and *Xenopus*. Although unlinked, these are evolutionary remnants of three other ParaHox gene clusters (8). Within the true vertebrate ParaHox gene cluster, the gene order and gene orientations are identical to amphioxus, although intergenic distances differ considerably. The presence of a ParaHox gene cluster in amphioxus, humans, mouse, and *Xenopus* implies there has been a strong selective pressure to maintain physical linkage of the three homeobox genes; otherwise, inversions and translocations would have dispersed these genes during the half-billion years since the divergence of cephalochordates and vertebrates. Despite this selective pressure, which must have been active throughout vertebrate history, we show that in the ray-finned fish clade, the ParaHox gene cluster was lost in the evolution of teleosts after divergence from more basal ray-finned fish.

Results

We searched the emerging genome sequences of zebrafish (*Danio rerio*; www.sanger.ac.uk/Projects/D_rerio) and two pufferfish [*Tetraodon nigroviridis* (10) and *Takifugu rubripes* (11)] for DNA sequences assignable to the Gsx, Xlox, and Cdx gene families. Comparison between species gave a consistent picture and assurance that we have identified all ParaHox genes in these teleost species. Consistent with an earlier report (12), we find that ParaHox genes are dispersed across the teleost genomes.

To determine whether loss of the cluster in teleosts occurred by inversions, translocations, or gene loss, we determined the synteny relations of chromosomal locations containing ParaHox gene(s) between teleosts and human. This analysis involved molecular phylogenetic analysis of gene families in the vicinity of each ParaHox gene and of the ParaHox genes themselves (Figs. 6–10, which are published as supporting information on the PNAS web site). The results for *Tetraodon*, for which the most complete genome sequence is available, are shown in Fig. 1. Similar results were obtained for *Danio* and *Takifugu* (see Figs. 11 and 12, which are published as supporting information on the PNAS web site). We find that *Tetraodon* possesses two chromosomal regions homologous to human 13q12.1 (the location of the human ParaHox cluster), but each contains only a single ParaHox gene: the *gsh1* or the *pdx1* gene, respectively. Contrary to an earlier report (12), both copies of *cdx2* are lost, but this may be compensated for by two copies of *cdx1* on other chromosomes (Figs. 1, 7, 11, and 12). The total number of homeobox genes in the Gsx, Xlox, and Cdx families is therefore the same in teleosts, human, mouse, and *Xenopus tropicalis*.

To determine the timing of cluster breakup in ray-finned fish phylogeny, we examined the genomic arrangement of ParaHox genes in two basal lineages, the bichir, *Polypterus senegalus*, the most basal extant ray-finned fish (13); and the bowfin, *Amia calva*, the immediate outgroup to the teleost fish (14) (Figs. 2 and 3). From a genomic bacterial artificial chromosome library of *Polypterus senegalus*, we identified a clone containing the three clustered ParaHox genes. Partial sequence of the clone revealed linkage of *Pdx1* and *Cdx2*, as well as a homeobox sequence of a Gsx gene. Using PCR, we also cloned a second Gsx gene and two additional Cdx genes from *Polypterus* whole-genomic DNA.

Conflict of interest statement: No conflicts declared.

This paper was submitted directly (Track II) to the PNAS office.

Data deposition: The sequences reported in this paper have been deposited in the GenBank database (accession nos. DQ504303 and DQ522897–DQ522900).

‡To whom correspondence should be addressed. E-mail: peter.holland@zoo.ox.ac.uk.

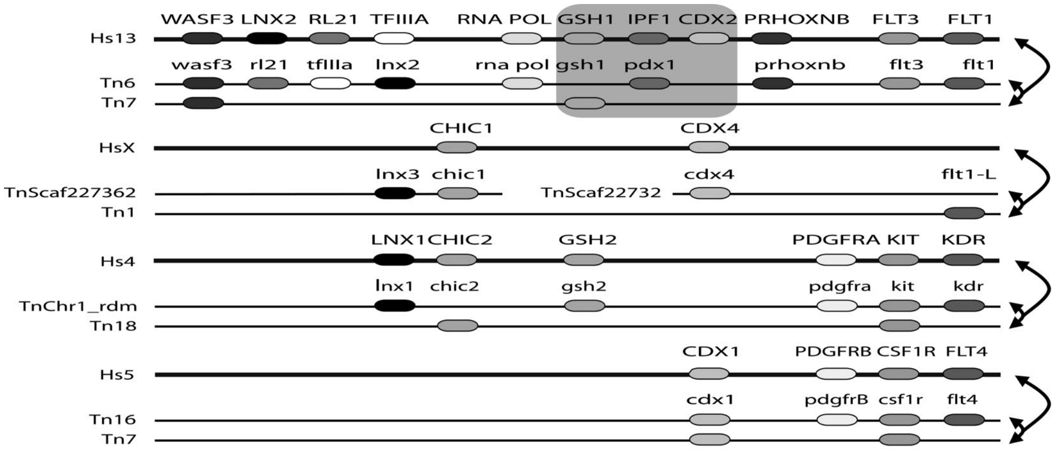

Fig. 1. Organization of human and *Tetraodon* ParaHox genes. The human ParaHox-bearing chromosomes are each homologous to a duplicated pair of chromosomal regions in *Tetraodon*. The shaded area designates the intact ParaHox cluster in humans and the broken cluster of *Tetraodon.*

From a genomic fosmid library of *Amia*, we identified a clone containing a complete ParaHox gene cluster. We determined the full sequence of this clone, inferring the coding sequences of the three homeobox genes, their transcriptional orientations, the intergenic DNA sequences, and intergenic distances (Fig. 2). Molecular phylogenetic analyses (Figs. 6 and 7) indicate that the three clustered *Amia* homeobox genes are orthologues of human *GSH1*, *IPF1* (*PDX1*), and *CDX2*; hence, this gene cluster is the orthologue of the human ParaHox gene cluster. Using PCR, we also cloned one additional Gsx gene (*Gsh2*) and two additional Cdx genes (*Cdx1* and *Cdx4*) from *Amia*, as also found in human, mouse (8), and *Xenopus*.

To investigate patterns of conservation of noncoding ParaHox sequences in tetrapod vs. fish lineages, we used MVISTA analysis with the fully sequenced *Amia* ParaHox gene cluster as the reference sequence (Fig. 4). This analysis revealed highly conserved blocks of noncoding sequences 5′ of *Gsh1* and in the intergenic region between *Gsh1* and *Pdx1* orthologues of human, mouse, frog, pufferfish (*Takifugu* and *Tetraodon*), and zebrafish.

Discussion

The intact ParaHox gene clusters of *Polypterus* and *Amia* reveal a sharp contrast between basal ray-finned fish (nonteleost actinopterygians) and teleosts such as *Danio*, *Fugu*, and *Tetraodon.* Our result on the timing of ParaHox cluster duplication in actinopterygians is consistent with evidence from Hox genes that both *Polypterus* and *Amia* (13, 14) diverged from the actinopterygian lineage before a whole-genome duplication event occurred close to the base of the teleost fish radiation (10). Hence, the breakup of the ParaHox gene cluster occurred after genome duplication and is secondarily derived in teleost fishes. Using synteny mapping and phylogenetic analyses, we show that in teleosts, ParaHox cluster breakup is because of gene loss, not transposition or inversion, and the total number of ParaHox genes is the same in teleosts, human, mouse, and *Xenopus*. These findings suggest that the selective pressure that held the ParaHox gene cluster together through much of vertebrate evolution, still active today in mammals, amphibian, and basal ray-finned fish, was lost or relaxed in the teleosts.

We propose a plausible explanation (Fig. 5) that explains this intriguing evolutionary pattern. In the model, the ParaHox gene cluster was retained intact for hundreds of millions of years, because the constituent homeobox genes had overlapping and/or shared regulatory elements; the three genes were interdigitated in the genome. In this situation, inversions and translocations would be selectively deleterious, because they would always disrupt a gene function, whereas gene losses would remove a gene function. After genome duplication at the base of the teleost fish, gene losses could instantly be tolerated because of genetic redundancy of both genes and regulatory elements, increasing the probability of breakup (Fig. 5). A prediction of the model is the presence of regulatory elements in the intergenic regions; consistent with this, we detected a large conserved noncoding region sited between *Gsh1* and *Pdx1*; however, experimental tests are needed to determine the function of this and other sequences. This model explains why the ParaHox gene cluster was strongly conserved in vertebrate lineages but broke apart in the common ancestor of zebrafish and pufferfish. In contrast, homeobox gene clusters that did not fragment in teleost fish, notably the Hox clusters, are more likely to be regulated by sequential activation, where cluster integrity is always necessary for correct gene control.

The wider implication of these results is that they call into question the suitability of zebrafish and pufferfish as models of

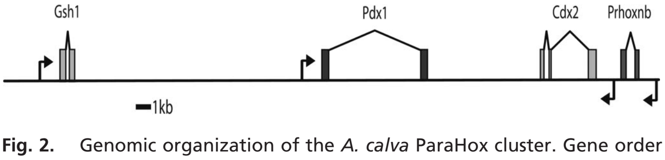

Fig. 2. Genomic organization of the *A. calva* ParaHox cluster. Gene order and transcriptional orientation of the clustered ParaHox genes, and neighboring PRHOXNB, are identical between human and *Amia*.

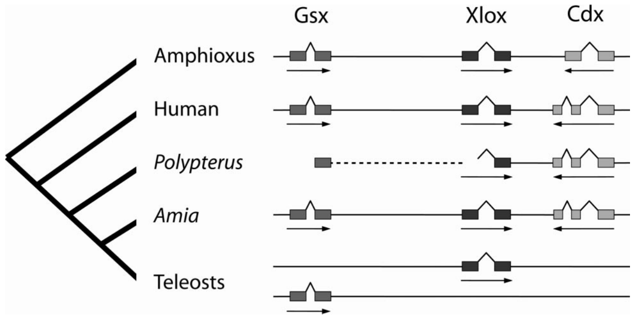

Fig. 3. Evolution of the vertebrate ParaHox cluster. From a single cluster of three genes in amphioxus, an intact cluster of ParaHox genes is conserved between amphibians and mammals and is also present in the basal actinopterygians *Polypterus* and *Amia.* The ParaHox cluster is broken in teleosts after whole-genome duplication.

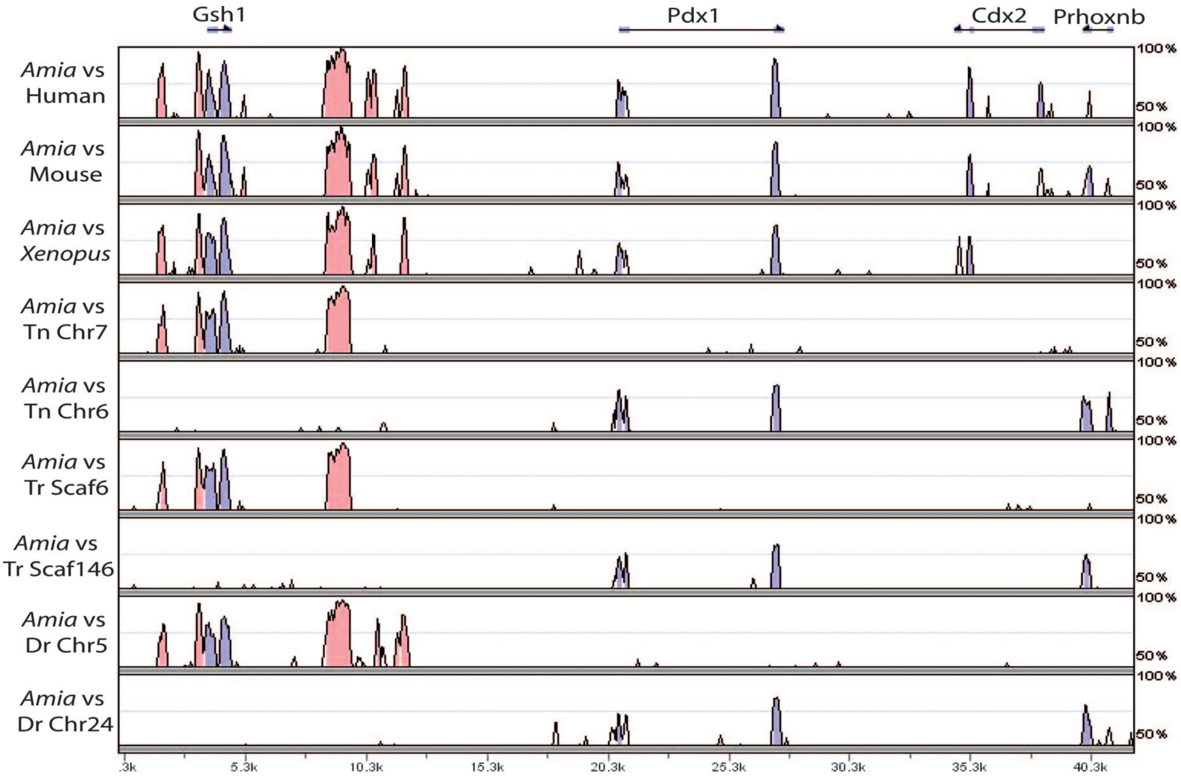

Fig. 4. MVISTA analysis comparing the *A. calva* ParaHox cluster to intact and broken clusters in other vertebrates. Exons are colored blue, and conserved noncoding sequences are pink. A highly conserved element is located between Gsh1 and Xlox in intact clusters and on the Gsh1-bearing chromosome of teleosts. *Xenopus*, *X. tropicalis*; Tn, *T. nigroviridis*; Tr, *T. rubripes*; and Dr, *D. rerio*. This analysis, based on sequence similarity, does not identify all functional regulatory sequences (15).

vertebrate genome organization. Our data suggest that the probability of breakup of linked pairs of genes, coregulated arrays of genes, and gene clusters is likely to have increased after whole-genome duplication in the ancestor of teleost fish. As a consequence, the organization of teleost fish genomes is expected to be more derived than in most other lineages of vertebrates, complicating efforts to identify "ancestral" genome arrangements and coordinately regulated genes by using teleost data. Ray-finned fish that diverged before the whole-genome duplication, such as *A. calva*, *Polypterus* sp., or *Lepisosteus* sp., would be more likely to retain ancestral gene linkages and to be more useful for this purpose. Nonetheless, teleost fish remain powerful models for addressing the relation between gene duplication and phenotypic evolution.

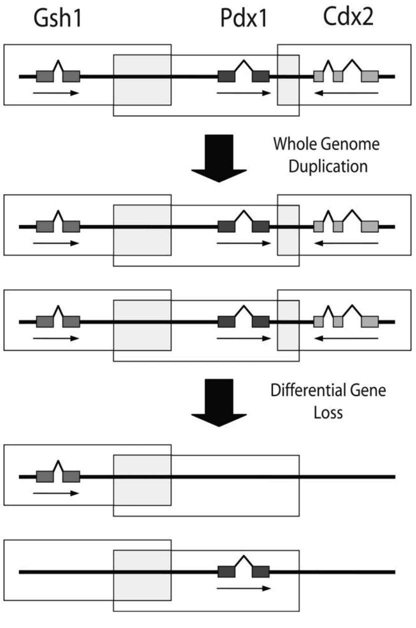

Fig. 5. A model for maintenance of ParaHox gene clustering, based on interdigitated and/or shared enhancers. The clear rectangles represent the proposed regulatory regions for each gene. After whole-genome duplication in the ancestor of teleost fish, redundancy of regulatory elements allows genomic breakup of the ParaHox cluster without affecting regulation of gene expression.

Materials and Methods

Searches for ParaHox genes were carried out in zebrafish (v5), *Tetraodon* (Tetraodon 7), and *Takifugu* (v4) genomes.

MVISTA (16) was used to compare the intact ParaHox clusters of *Amia*, human, mouse, and *X. tropicalis* with the broken clusters of teleosts using the AVID program (17).

A 1.5× *A. calva* genomic fosmid library (in CopyControl pCC1FOS) was screened for the *Cdx2* genes using the primers AcCdx3-1f (5′-ACAGGGTCTGTTTTACACG-3′), AcCdx3-2f (5′-GCTGGAGTTTATGGCTAGAG-3′), and AcCdx3-1r (5′-TGGGAAAAGGGAAACAC-3′). A single clone was isolated (AcF-5F2) and found to contain Gsx and Xlox genes by degenerate PCR using the primers SO1 (5′-ARYTNGARAARGARTT-3′) and SO2 (5′-CKNCKRTTYTGRAACCA-3′). The clone was sequenced by using a combination of shotgun subcloning, PCR, and direct sequencing.

A screen of a *Polypterus senegalus* bacterial artificial chromosome library (www.RZPD.de, library 640) for Hox gene clones (13) identified a clone containing *Pdx1* (147-018). PCR with degenerate primers revealed the presence of a Cdx and Gsx gene,

Mulley *et al.* PNAS | July 5, 2006 | vol. 103 | no. 27 | 10371

and the clone was partially sequenced using a combination of subcloning and PCR.

We thank John Postlethwait, Bill Cresko, and Angel Amores (University of Oregon, Eugene) for the *Amia* fosmid library and Robin Overstreet (Gulf Coast Research Laboratory, Ocean Springs, MS) for *Amia* tissue for genomic DNA. P.W.H.H. acknowledges support from the Biotechnology and Biological Sciences Research Council and the Elizabeth Hannah Jenkinson Fund. C.-h.C. acknowledges support from National Science Foundation (Grant MCB 0447478).

1. Brooke, N. M., Garcia-Fernàndez, J. & Holland, P. W. H. (1998) *Nature* **392,** 920–922.
2. Ferrier, D. E. K. & Minguillón, C. (2003) *Int. J. Dev. Biol.* **47,** 605–611.
3. Valerius, M. T., Li, H., Stock, J. L., Weinstein, M., Kaur, S., Singh, G. & Potter, S. S. (1995) *Am. J. Anat.* **203,** 337–351.
4. Hsieh-Li, H. M., Witte, D. P., Szucsik, J. C., Weinstein, M., Li, H. & Potter, S. S. (1995) *Mech. Dev.* **50,** 177–186.
5. Offield, M. F., Jetton, T. L., Labosky, P. A., Ray, M., Stein, R. W., Magnuson, M. A., Hogan, B. L. & Wright, C. V. E. (1996) *Development (Cambridge, U.K.)* **122,** 983–995.
6. Moreno, E. & Morata, G. (1999) *Nature* **400,** 873–877.
7. Reece-Hoyes, J. S., Keenan, I. D. & Isaacs, H. V. (2002) *Dev. Dyn.* **223,** 134–140.
8. Pollard, S. L. & Holland, P. W. H. (2000) *Curr. Biol.* **10,** 1059–1062.
9. Ferrier, D. E. K., Dewar, K., Cook, A., Chang, J. L., Hill-Force, A. & Amemiya, C. (2005) *Curr. Biol.* **15,** R820–R822.
10. Jaillon, O., Aury, J.-M., Brunet, F., Petit, J.-L., Stange-Thomann, N., Mauceli, E., Bouneau, L., Fishcer, C., Ozouf-Costaz, C., Bernot, A., *et al.* (2004) *Nature* **431,** 946–957.
11. Aparicio, S., Chapman, J., Stupka, E., Putnam, N., Chia, J. M., Dehal, P., Christoffels, A., Rash, S., Hoon, S., Smit, A., *et al.* (2002) *Science* **297,** 1301–1310.
12. Prohaska, S. J. & Stadler, P. F. (2006) *J. Exp. Zool.*, in press.
13. Chiu, C.-H., Dewar, K., Wagner, G. P., Takahashi, K., Ruddle, F., Ledje, C., Bartsch, P., Scemama, J.-L., Stellwag, E., Fried, C., *et al.* (2004) *Genome. Res.* **14,** 11–17.
14. Crow, K. D., Stadler, P. F., Lynch, V. J., Amemiya, C. & Wagner, G. P. (2006) *Mol. Biol. Evol.* **23,** 121–136.
15. Fisher, S., Grice, E. A., Vinton, R. M., Bessling, S. L. & McCallion, A. S. (2006) *Science* **312,** 276–279.
16. Frazer, K. A., Pachter, L., Poliakov, A., Rubin, E. M. & Dubchak, I. (2004) *Nucleic Acids Res.* **32,** W273–W279.
17. Bray, N., Dubchak, I. & Pachter, L. (2003) *Genome Res.* **13,** 97–102.