25. Springer, M. S. *et al.* Endemic African mammals shake the phylogenetic tree. *Nature* **388,** 61–63 (1997).
26. Duret, L., Mouchiroud, D. & Gouy, M. HOVERGEN: a database of homologous vertebrate genes. *Nucleic Acids Res.* **22,** 2360–2365 (1994).
27. Tajima, F. Simple methods for testing the molecular evolutionary clock hypothesis. *Genetics* **135,** 599–607 (1993).
28. Kumar, S., Tamura, K. & Nei, M. *MEGA: Molecular Evolutionary Genetic Analysis* (Pennsylvania State Univ., 1993).
29. Gheerbrant, E., Sudre, J. & Cappetta, H. A Paleocene proboscidean from Morocco. *Nature* **383,** 68–70 (1996).
30. Benton, M. J. Phylogeny of the major tetrapod groups: morphological data and divergence dates. *J. Mol. Evol.* **30,** 409–424 (1990).

Supplementary information is available on *Nature*'s World-Wide Web site (http://www.nature.com) or as paper copy from Mary Sheehan at the London editorial office of *Nature*.

Acknowledgements. We thank L. Poling, A. Beausang, and R. Padmanabhan for assistance with sequence data retrieval; A. Beausang for artwork; A. G. Clark, C. A. Hass, I. Jakobsen, M. Nei, C. R. Rao, and A. Walker for comments and discussion; and L. Duret for instructions on use of the HOVERGEN database. This work was supported in part by grants to M. Nei (NIH and NSF) and S.B.H. (NSF).

Correspondence and requests for materials should be addressed to S.B.H. (e-mail: sbh1@psu.edu).

The ParaHox gene cluster is an evolutionary sister of the Hox gene cluster

Nina M. Brooke*, Jordi Garcia-Fernàndez† & Peter W. H. Holland*

** School of Animal and Microbial Sciences, University of Reading, Whiteknights, PO Box 228, Reading RG6 6AJ, UK*
† Departament de Genètica, Facultat de Biologia, Universitat de Barcelona, Av. Diagonal 645, 08028 Barcelona, Spain

Genes of the Hox cluster are restricted to the animal kingdom and play a central role in axial patterning in divergent animal phyla[1]. Despite its evolutionary and developmental significance, the origin of the Hox gene cluster is obscure. The consensus is that a primordial Hox cluster arose by tandem gene duplication close to animal origins[2–5]. Several homeobox genes with high sequence identity to Hox genes are found outside the Hox cluster and are known as 'dispersed' Hox-like genes; these genes may have been transposed away from an expanding cluster[6]. Here we show that three of these dispersed homeobox genes form a novel gene cluster in the cephalochordate amphioxus. We argue that this 'ParaHox' gene cluster is an ancient paralogue (evolutionary sister) of the Hox gene cluster; the two gene clusters arose by duplication of a ProtoHox gene cluster. Furthermore, we show that amphioxus ParaHox genes have co-linear developmental expression patterns in anterior, middle and posterior tissues. We propose that the origin of distinct Hox and ParaHox genes by gene-cluster duplication facilitated an increase in body complexity during the Cambrian explosion.

Homeodomain sequence comparisons reveal that at least five classes of homeobox genes are as closely related to Hox genes as many of the latter are to each other[6]. These are the Evx, Mox, Cdx (or cad), Xlox, and Gsx homeobox classes (we term a class defined by mouse *Gsh-1* and *Gsh-2* as Gsx). The two mammalian Evx genes are each linked to the 5′ end of Hox gene clusters[6], and a cnidarian Evx-like gene is linked to a Hox-like gene[7], indicating that the close sequence relationship between Evx and Hox genes reflects tandem duplication. Mox genes may represent a similar case because the mouse *Mox-1* gene maps to chromosome 11, close to the Hoxb cluster[8]. The Cdx, Xlox and Gsx gene families are more problematic.

To investigate the evolutionary origins of Cdx, Xlox and Gsx genes, we elected to clone representatives of each gene family from amphioxus. This is because homeobox gene families in this animal are not complicated by either excessive duplication (as in vertebrates[9]) or divergence and rearrangement (as in *Drosophila* or nematode)[6,10]. Using primers directed to Hox class homeoboxes and amphioxus genomic DNA as template, amplification by polymerase chain reaction (PCR) yielded partial clones of Cdx and Xlox class homeoboxes. A fragment of amphioxus Gsx was also cloned by PCR, using primers designed from the two mammalian gene family members *Gsh-1* and *Gsh-2*. To determine the complete homeobox sequence of each gene, we isolated longer clones from amphioxus genomic libraries: only single members of each class were obtained, which we named *AmphiCdx*, *AmphiXlox* and *AmphiGsx*. Their encoded homeodomains resemble those of the *Drosophila* or vertebrate homologues (Fig. 1).

Analysis of genomic clones revealed that amphioxus Xlox and Cdx class homeoboxes were unexpectedly contained within a single bacteriophage clone. Mapping indicated that the homeoboxes were separated by just 7.5 kilobases (kb). Furthermore, using genomic walking we found that these two homeobox genes are physically linked to the *AmphiGsx* gene. The Gsx and Xlox homeoboxes are separated by just 25 kb (Fig. 2). We designate this tight cluster of three genes the ParaHox gene cluster.

The finding that amphioxus Gsx, Xlox and Cdx class genes form a novel homeobox cluster challenges the idea that these homeobox gene classes are 'dispersed' Hox genes. To reconcile linkage in

```
AmphiCdx  KDKYRVVYSDHQRLELEKEFYSNKYITIKRKVQLANELGLSERQVKIWFQNRRAKQRKMA 100%
mCdx-1    --------T----------HYSR----R--SE--AN---T--------------E--VN  77%
mCdx-2    --------T----------HFSR----R--SE--AT-----------------E--IK  78%
mCdx-4    -E------T----------HC-R----R--SE--VN-----------------E-MIK  77%
D Cad     --------T-F--------YCTSR----R--SE--QT-S---------------E-TSN  72%
Ce pal-1  A----M----Y---------HTSPF--SD--S--STM-S-T---I----------D-RDK  65%

AmphiXlox NKRTRTAYTRGQLLELEKEFHFNKYISRPRRIELAAMLNLTERHIKIWFQNRRMKWKKEQ 100%
mpdx-1    ----------A---------L----------V---V---------------------E  92%
XlHbox8   ----------A---------L----------V---V---------------------E  92%
Htr-A2    --------S-S--F--------D--------V---SS--------------------ME  85%

AmphiGsx  SRRMRTAFSSTQLLELEREFASNMYLSRLRRIEIATFLNLSEKQVKIWFQNRRVKHKKEA 100%
mGsh-1    -K------T----------------------------Y----------------------G  93%
mGsh-2    GK------T----------S----------------Y----------------------G  90%
```

Figure 1 Homeodomains of amphioxus Cdx, Xlox and Gsx genes aligned to mouse (m), *Drosophila* (D), nematode (Ce), *Xenopus* (Xl) and leech (Htr) homologues. The mouse Cdx2 gene is the probable orthologue of human CDX3. Dashes indicate identical amino acids.

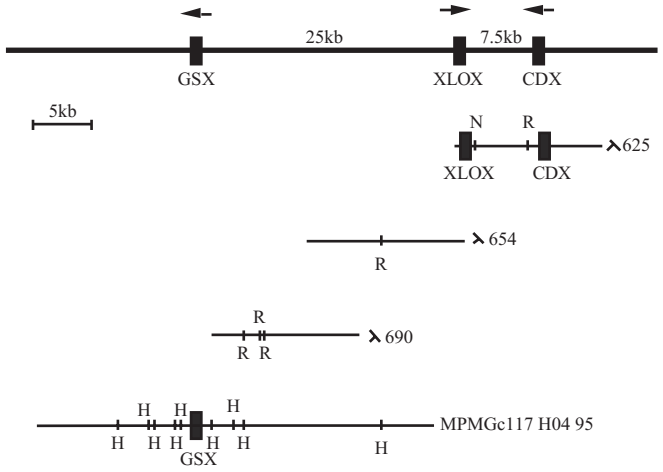

Figure 2 Genomic organization of amphioxus Gsx, Xlox and Cdx genes, showing genomic clones used in walking. Arrows denote transcriptional orientation. R, *Eco*RI, site and N, *Not* I site, mapped in bacteriophage clones only; H, *Hin*dIII sites mapped in cosmid only.

Nature © Macmillan Publishers Ltd 1998

amphioxus with an origin by transposition, Gsx, Xlox and Cdx genes must have been transposed from the evolving Hox gene cluster as a cassette of three adjacent genes. We tested this by molecular phylogenetic analysis of Hox, Gsx, Xlox and Cdx genes (Fig. 3) and found that the Gsx, Xlox and Cdx gene classes do not form a distinct branch on the tree, as would be predicted for genes that arose as neighbours within an early Hox gene cluster. The three gene classes are widely spread in the tree topology. To test the statistical significance of this finding, we estimated likelihood values for the tree shown in Fig. 3 and for 15 modified trees consistent with transposition of a cassette (Gsx, Xlox and Cdx forming a distinct clade basal to any Hox subfamily). A resampling of estimated likelihood method found that the tree shown in Fig. 3 is supported over the modified trees with a bootstrap value of 97.6%.

The tree topology suggests a probable evolutionary origin for Hox and ParaHox gene clusters. Within the tree, Hox genes divide into four subfamilies of evolutionarily related genes: 'anterior' genes (chordate PG1, PG2; insect *lab*, *pb*), 'group 3' genes (chordate PG3; insect *zen*/PG3), 'middle' genes (chordate PG4 to PG8; insect *Dfd* to *abd-A*) and 'posterior' genes (chordate PG9 to PG13; insect *Abd-B*). These subfamilies may reflect a stage in Hox gene evolution consisting of four linked genes. We find that Gsx, Xlox and Cdx class genes are grouped with different Hox gene subfamilies: the anterior, group 3 and posterior subfamilies, respectively. On the basis of this molecular phylogeny combined with the physical linkage data, we propose that Hox and ParaHox gene clusters arose by duplication of an ancestral 'ProtoHox' gene cluster. This ancestral state probably consisted of four linked genes, although more or less cannot be discounted with confidence. After duplication of the ProtoHox cluster, one set of descendant genes underwent a series of tandem duplications to yield the definitive Hox gene cluster, while the other evolved into the ParaHox gene cluster of Gsx, Xlox and Cdx (Fig. 4). We use the term ParaHox because this gene cluster is a paralogue of the definitive Hox gene cluster.

To investigate the possible functions of the ParaHox gene cluster, we examined the developmental expression patterns of *AmphiCdx*, *AmphiXlox* and *AmphiGsx*. The level of *AmphiGsx* expression is low, and is detected only in the cerebral vesicle (fore-/midbrain homologue[11]) of 18–20-h embryos (Fig. 5a). *AmphiXlox* is strongly expressed in a stripe of the archenteron wall (presumptive gut) from the neurula stage through to larvae (Fig. 5b). The region of expression is just anterior to the posterior pole at the developmental stage shown. As development proceeds, the body grows at the posterior end, so that *AmphiXlox*-expressing cells become far anterior to the anus in the swimming larvae (data not shown). There is also strong but transient expression in two cells of the

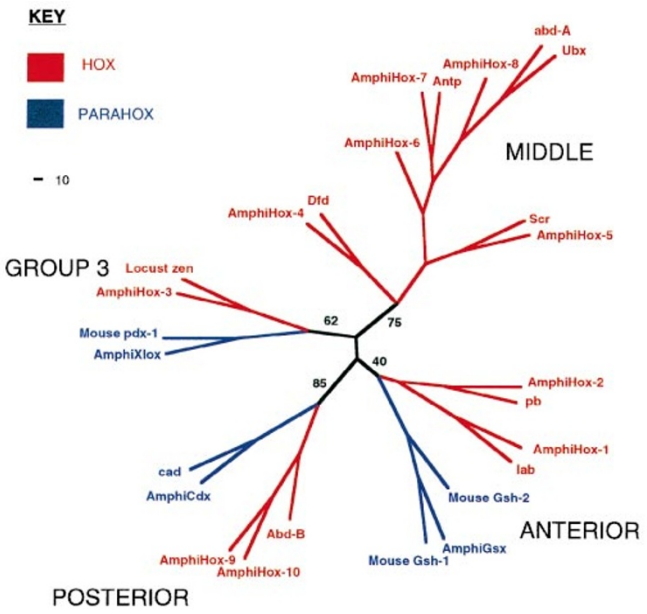

Figure 3 Evolutionary relationships of Cdx, Xlox and Gsx genes to Hox genes, as deduced by neighbour-joining analysis of homeodomains. The tree is unrooted and shows four subfamilies. Figures refer to bootstrap percentages.

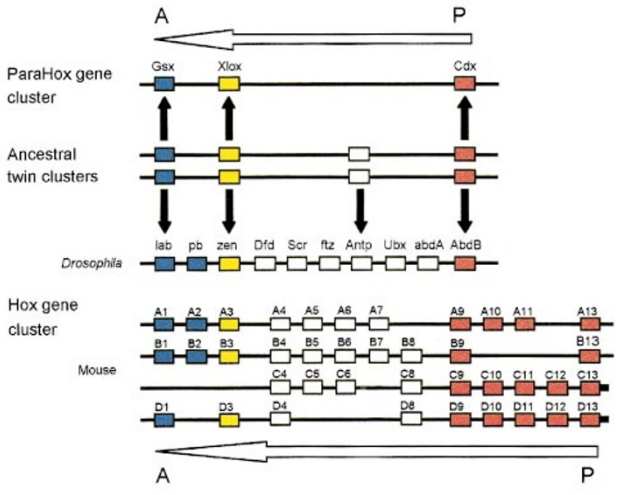

Figure 4 Origin of Hox and ParaHox gene clusters inferred from combining gene linkage and phylogenetic analyses. Hox and ParaHox gene clusters evolved from ancestral twin clusters generated by duplication. Horizontal arrows denote polarity of spatial colinearity (A, anterior; P, posterior).

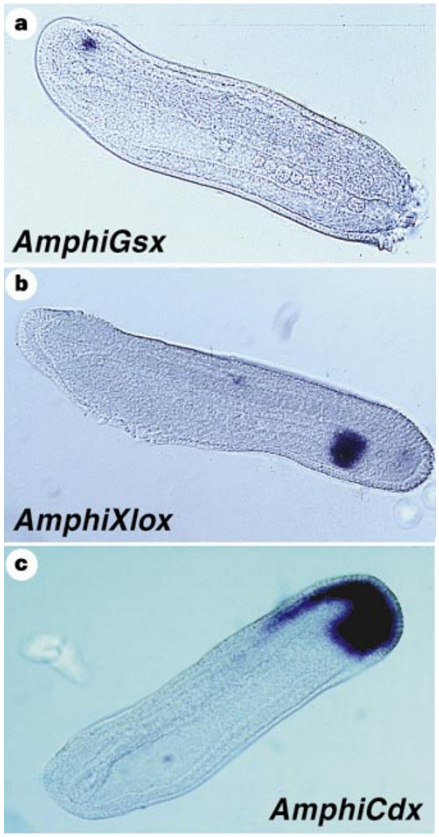

Figure 5 Whole-mount *in situ* hybridization to 20-h amphioxus embryos. Expression of **a**, *AmphiGsx*; **b**, *AmphiXlox*; **c**, *AmphiCdx* genes. Anterior is to the left.

Nature © Macmillan Publishers Ltd 1998

Nature © Macmillan Publishers Ltd 1998

neural tube opposite somite 5, at a site fated to form the first pigment spot. This has faded to a faint signal at the stage shown in Fig. 5b. In early neurulae, *AmphiCdx* is expressed posteriorly in all tissues. As development proceeds, expression persists most strongly in the posterior region of the developing gut, continuous with a gradient of expression in the most posterior part of the neural tube (Fig. 5c). At 20 h development, *AmphiCdx* expression in the gut is just posterior to that of *AmphiXlox* (Fig. 5b, c); by the larval stage, the two expression domains are clearly separated.

Data from other species argue that ParaHox gene expression sites have been conserved in evolution. Like amphioxus, both arthropods and vertebrates express Cdx genes in the hindgut[12,13]. The multiple Cdx genes of vertebrates have additional roles in other tissues but these are not shared between gene family members and may be derived. Expression of Xlox in central regions of the presumptive gut has been reported in vertebrates[14,15] and leeches[16] and is closely comparable to *AmphiXlox*. The single mammalian Xlox gene *pdx-1* is expressed in the endodermal precursor of the pancreas and is required for correct development of pancreas and rostral duodenum[15,17]. Like *AmphiGsx*, two Gsx genes previously described from mouse are expressed in the developing brain. *Gsh-2* is involved in brain development[18] and *Gsh-1* in development of the adenohypophysis[19]. This component of the pituitary develops as an ectodermal outpocketing of the oral cavity (Rathke's pouch); the homologous structure is endodermal in hagfish[20]. Hence, mouse Gsx genes have roles in both brain and anterior-gut development. We cannot be certain whether *AmphiGsx* is also expressed in anterior gut; its expression is at the limit of sensitivity of our methodology and weak expression in a few anterior gut cells would remain undetected.

In summary, the order of genes along the ParaHox gene cluster (Fig. 2) matches the spatial order of gene-expression sites along the body axis (Fig. 4). Hence, ParaHox genes, like definitive Hox genes[21], display spatial co-linearity, although the details differ. ParaHox genes obey the co-linearity rule in the neural tube and at least two, possibly three, of the genes also obey this rule in the gut. The polarity of spatial co-linearity is identical between Hox and ParaHox genes (anterior genes related to anterior genes and so on), suggesting that both genetic systems inherited spatial co-linearity from the common precursor gene cluster.

Precise dating of Hox/ParaHox origins is not yet possible, but the cluster duplication must predate the divergence of 'higher' triploblast phyla. This is because gene clusters containing definitive Hox genes are present in arthropods, echinoderms, nematodes and chordates[1,6,9,22], whereas genes belonging to the three ParaHox classes have been reported in arthropods, annelids, nematodes, echinoderms and chordates[6,13,15,17,23]. Circumstantial evidence points to retention of the ParaHox gene linkages in mammals, despite further duplication. Thus, the human *PDX1* gene maps to the same chromosomal location as one of the Cdx genes (*CDX3*) at 13q12.1 (ref. 24), and mouse *Gsh-2* maps to the region syntenic to 13q12.1 (chromosome 5 at 89 cM)[25,26]. The situation in more basal lineages of animals is unclear. Our phylogenetic analyses of cnidarian Cnox genes give ambiguous results, and more physical linkage data from cnidarians are needed to clarify the situation. Current data are consistent with the hypothesis that distinct Hox and ParaHox gene clusters originated on the triploblast stem lineage. The dramatic increase in body-plan diversity associated with the Cambrian explosion probably also occurred on this lineage[27] and predominantly affected animals of the triploblast grade. Duplication of a primordial ProtoHox gene cluster, followed by functional recruitment to different tissues, may have facilitated an increase in body complexity during the Cambrian explosion. □

Methods

Genomic cloning. Total DNA was extracted from adult *Branchiostoma floridae* from Old Tampa Bay, Florida, for PCR and library construction. Fragments of Cdx and Xlox were isolated by PCR using primers SO1 and SO2 (ref. 9). Gsx amplification used SO2 in combination with primer Sogsh (CAGCTCTTG-GARCTNGARCGNG). Genomic clones containing the Cdx and Xlox genes were isolated from a lambda FIXII library[9]. Gsx was isolated from cosmid library MPMGc117 of the RZPD, Berlin (ref. 28). Genomic walking was performed as described[9]. Overlap of phage and cosmid clones was verified by subcloning and sequencing a shared 1-kb *Hin*dIII fragment.

Phylogenetic analyses. Molecular phylogenetic trees were constructed from homeodomain protein sequences, to avoid possible distortion by codon usage differences, and utilized the neighbour-joining method on a Kimura distance matrix, implemented using PHYLIP3.572c (ref. 29). Topology robustness was assessed by bootstrap resampling. Statistical likelihood of alternative user-defined trees was assessed by a resampling of estimated likelihood method implemented with PROTML in PHYLIP3.572c (ref. 29).

Whole mount *in situ* hybridization. Hybridization to amphioxus embryos was done as described[30], except that detection time was increased to 5 days for *AmphiGsx*.

Received 4 December 1997; accepted 24 February 1998.

1. Slack, J. M. W., Holland, P. W. H. & Graham, C. F. The zootype and the phylotypic stage. *Nature* **361,** 490–492 (1993).
2. Kappen, C., Schughart, K. & Ruddle, F. H. Two steps in the evolution of Antennapedia-class vertebrate homeobox genes. *Proc. Natl Acad. Sci. USA* **86,** 5458–5463 (1989).
3. Schubert, F. R., Nieselt-Struwe, K. & Gruss, P. The Antennapedia-type homeobox genes have evolved from three precursors separated early in metazoan evolution. *Proc. Natl Acad. Sci. USA* **90,** 143–147 (1993).
4. Gehring, W. J., Affolter, M. & Bürglin, T. Homeodomain proteins. *Annu. Rev. Biochem.* **63,** 487–526 (1994).
5. Zhang, J. & Nei, M. Evolution of Antennapedia-class homeobox genes. *Genetics* **142,** 295–303 (1996).
6. Bürglin, T. in *A Guidebook for Homeobox Genes* (ed. Duboule, D.) 25–71 (Oxford University Press, Oxford, 1994).
7. Miller, D. J. & Miles, A. Homeobox genes and the zootype. *Nature* **365,** 215–216 (1993).
8. Candia, A. F. *et al*. *Mox-1* and *Mox-2* define a novel homeobox gene subfamily and are differentially expressed during early mesodermal patterning in mouse embryos. *Development* **116,** 1123–1136 (1992).
9. Garcia-Fernàndez, J. & Holland, P. W. H. Archetypal organization of the amphioxus *Hox* gene cluster. *Nature* **370,** 563–566 (1994).
10. Averof, M., Dawes, R. & Ferrier, D. Diversification of arthropod Hox genes as a paradigm for the evolution of gene functions. *Sem. Cell. Dev. Biol.* **7,** 539–551 (1996).
11. Williams, N. A. & Holland, P. W. H. Old head on young shoulders. *Nature* **383,** 490 (1996).
12. Duprey, P. *et al*. A mouse gene homologous to the *Drosophila* gene *caudal* is expressed in epithelial cells from the embryonic intestine. *Genes Dev.* **2,** 1647–1654 (1988).
13. Calleja, M., Moreno, E., Pelaz, S. & Morata, G. Visualization of gene expression in living adult *Drosophila*. *Science* **274,** 252–255 (1996).
14. Wright, C. V. E., Schnegelsberg, P. & DeRobertis, E. M. XIHbox 8—a novel *Xenopus* homeoprotein restricted to a narrow-band of endoderm. *Development* **104,** 787–794 (1988).
15. Offield, M. F. *et al*. *Pdx-1* is required for pancreatic outgrowth and differentiation of the rostral duodenum. *Development* **122,** 983–995 (1996).
16. Wysocka-Diller, J., Aisemberg, G. O. & Macagmo, E. R. A novel homeobox cluster expressed in repeated structures of the midgut. *Dev. Biol.* **171,** 439–447 (1995).
17. Jonsson, J., Carlsson, L., Edlund, T. & Edlund, H. Insulin-promoter-factor-1 is required for pancreas development in mice. *Nature* **371,** 606–609 (1994).
18. Szucsik, J. C. *et al*. Altered forebrain and hindbrain development in mice mutant for the *Gsh-2* homeobox gene. *Dev. Biol.* **191,** 230–242 (1997).
19. Li, H., Zeitler, P. S., Valerius, M. T., Small, K. & Potter, S. S. *Gsh-1*, an orphan Hox gene, is required for normal pituitary development. *EMBO J.* **15,** 714–724 (1996).
20. Gorbman, A. Early development of the hagfish pituitary gland: evidence for the endodermal origin of the adenohypophysis. *Am. Zool.* **23,** 639–654 (1983).
21. Gaunt, S. J., Sharpe, P. T. & Duboule, D. Spatially restricted domains of homeogene transcripts in mouse embryos: relation to a segmented body plan. *Development* (suppl.) **104,** 169–180 (1988).
22. Popodi, E., Kissinger, J. C., Andrews, M. E. & Raff, R. A. Sea urchin Hox genes: insights into the ancestral Hox cluster. *Mol. Biol. Evol.* **13,** 1078–1086 (1996).
23. Irvine, S. Q., Warinner, S. A., Hunter, J. D. & Martindale, M. Q. A survey of homeobox genes in *Chaetopterus variopedatus* and analysis of polychaete homeodomains. *Mol. Phylo. Evol.* **7,** 331–345 (1997).
24. Stoffel, M. *et al*. Localization of human homeodomain transcription factor insulin promoter factor-1 (*Ipf1*) to chromosome band 13q12.1. *Genomics* **28,** 125–126 (1995).
25. Kozak, C. A., Goffinet, A. & Stephenson, D. A. Mouse chromosome-5. *Mamm. Genome* **5,** s65–s78 (1994).
26. Fiedorek, F. T. & Kay, E. S. Mapping of the insulin promoter factor-1 gene (*Ipf1*) to distal mouse chromosome-5. *Genomics* **28,** 581–584 (1995).
27. Conway Morris S. Why molecular biology needs palaeontology. *Development* (suppl.) 1–13 (1994).
28. Lehrach, H. *et al*. in *Genome Analysis 1: Genetic and Physical Mapping* (eds Davies, K. E. & Tilghman, S. M.) 38–81 (Cold Spring Harbor Laboratory Press, New York, 1990).
29. Felsenstein, J. PHYLIP version 3.5c (Department of Genetics, Univ. Washington, Seattle, 1993).
30. Holland, P. W. H., Holland, L. Z., Williams, N. A. & Holland, N. D. An amphioxus homeobox gene: sequence conservation, spatial expression during development and insights into vertebrate evolution. *Development* **116,** 653–661 (1992).

Acknowledgements. We thank G. Balavoine and C. V. E. Wright for discussion, S. J. Patton and C. Burgtorf for access to the cosmid library and help with screening, and N. A. Williams and S. A. J. Thompson for advice and assistance. This research was funded by a BBSRC Earmarked Studentship (to N.M.B.) and by the DGICYT (J.G.F.), and facilitated by a grant from the Acciones Integradas of the British Council/Ministerio de Educación y Ciencia.

Correspondence and requests for materials should be addressed to P.W.H.H. (e-mail: p.w.h.holland@reading.ac.uk). The *AmphiGsx*, *AmphiXlox* and *AmphiCdx* sequences are deposited in GenBank under accession numbers AF052463 to AF052465.

Nature © Macmillan Publishers Ltd 1998