A molecular timescale for vertebrate evolution

Sudhir Kumar & S. Blair Hedges

Department of Biology and Institute of Molecular Evolutionary Genetics, 208 Mueller Laboratory, Pennsylvania State University, University Park, Pennsylvania 16802, USA

A timescale is necessary for estimating rates of molecular and morphological change in organisms and for interpreting patterns of macroevolution and biogeography[1–9]. Traditionally, these times have been obtained from the fossil record, where the earliest representatives of two lineages establish a minimum time of divergence of these lineages[10]. The clock-like accumulation of sequence differences in some genes provides an alternative method[11] by which the mean divergence time can be estimated. Estimates from single genes may have large statistical errors, but multiple genes can be studied to obtain a more reliable estimate of divergence time[1,12–13]. However, until recently, the number of genes available for estimation of divergence time has been limited. Here we present divergence-time estimates for mammalian orders and major lineages of vertebrates, from an analysis of 658 nuclear genes. The molecular times agree with most early (Palaeozoic) and late (Cenozoic) fossil-based times, but indicate major gaps in the Mesozoic fossil record. At least five lineages of placental mammals arose more than 100 million years ago, and most of the modern orders seem to have diversified before the Cretaceous/Tertiary extinction of the dinosaurs.

Molecular clocks are first calibrated with a known time of divergence and then used to estimate divergence times of other species. The divergence of birds and mammals provides a reliable calibration point with which to anchor molecular clocks. The earliest ancestors of mammals (synapsids) and birds (diapsids) are lizard-like and first appear in the Carboniferous period, at ~310 million years (Myr) ago[10,14] (Fig. 1a). The fact that the fossil record[10,14] documents a morphological transition from lobe-finned fishes to stem tetrapods at 370–360 Myr ago, and records the appearance of stem amphibians at 338 Myr ago, indicates that the time of the diapsid–synapsid split (within amniotes) is unlikely to be a considerable underestimate. Alternatively, multiple calibration points based on the mammalian fossil record may be used, but this might result in substantial underestimates of divergence time[12,15].

We used 658 genes, representing 207 vertebrate species, to estimate divergence times by two methods (see also Fig. 1b; Methods; Supplementary Information). Taxonomic biases in the sequence databases resulted in a predominance of mammalian sequences studied. For estimates derived from large numbers of genes, distributions of divergence-time estimates are approximately normal (Fig. 2a–i; Methods). These distributions show considerable dispersion around the peak, as reflected in high coefficients of variation. On average, standard errors of ~10% were obtained with 10 genes, 5% with 50 genes, and 3% with 100 genes. Multigene time estimates obtained without rate-constancy tests were nearly identical to those obtained with stringent rate testing (Fig. 2j–l). This indicates that there are probably no underlying directional biases in the data.

Molecular times for the origin of the major lineages of vertebrates in the Palaeozoic and early Mesozoic eras are similar to those that are based on the fossil record[14] (Fig. 3). The molecular time estimate for the marsupial–placental split, 173 Myr ago, corresponds well with the fossil-based estimate (178–143 Myr ago)[16]. The bird–crocodilian divergence is slightly younger than the earliest fossils suggest[10], at 240 Myr ago, but this difference is less than one standard error. The molecular estimate of the lissamphibian–amniote divergence at 360 Myr ago also agrees with the fossil-based estimate[10,14]. Fewer genes are available to time the earliest divergences among vertebrates, but molecular times (of 564 and 528 Myr ago) are consistent with the Late Cambrian fossil record for the first appearance of vertebrates (at 514 Myr ago)[14].

A striking pattern revealed by our molecular divergence times is the Cretaceous origin of all modern orders of mammals examined (Fig. 3). Earlier molecular[12] and fossil[15] studies found evidence that at least some mammalian divergences occurred in the Cretaceous, leaving open the possibility of a gradual diversification of orders into the early Cenozoic. Molecular times now indicate that at least five major lineages of placental mammals (Edentata, Hystricognathi, Sciurognathi, Paenungulata, Archonta + Ferungulata) may have arisen in the Early Cretaceous, >100 Myr ago, and that most mammalian orders were involved in a Cretaceous radiation that predated the Cretaceous/Tertiary extinction of the dinosaurs (Fig. 3). The origin of most mammalian orders seems not to be tied to the filling of niches left vacant by dinosaurs, but is more likely to be related to events in Earth history[12]. Similarly, a mid-Cretaceous divergence was obtained for the bird orders Anseriformes and Galliformes[12,17].

Multigene divergence times within several orders of mammals compare closely with fossil-based estimates. For example, molecular divergence times from humans to chimpanzees, gorillas, gibbons, and Old World monkeys are close to currently accepted dates from the fossil record[4,10,18]. The orangutan molecular divergence

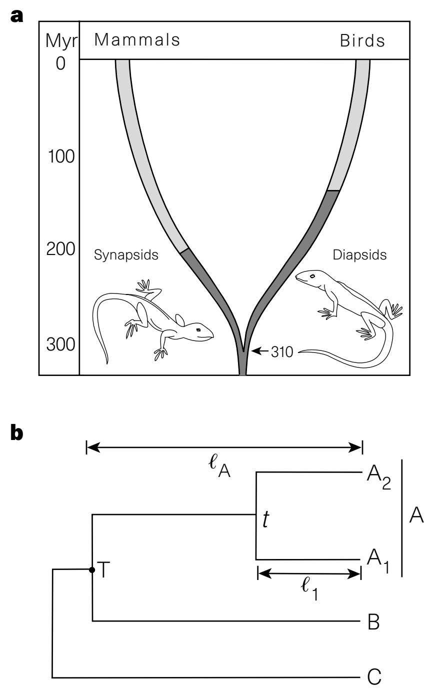

Figure 1 Estimation of divergence times. **a**, Calibration. The arrow marks the first appearance of synapsids (ancestors of mammals) and diapsids (ancestors of birds) in the fossil record at 310 Myr ago. Reconstructions of an early synapsid (*Varanosaurus*) and stem diapsid (*Hylonomus*) are shown. The dark shading represents the reptilian portion and the lighter shading represents the avian and mammalian portion of the phylogeny. **b**, Two methods for dating the unknown divergence time (t) between A_1 and A_2 when A and B diverged at calibration time T (Myr ago). In the average distance method, $t = d_{12}/(2r)$, where $r = (d_{1B} + d_{2B})/(4T)$ is the rate of change for lineages A and B, and d_{ij} is the number of substitutions per site between sequences i and j. In the lineage-specific method, $t_{12} = d_{12}/(2r_A)$ (where $r_A = \ell_A/T$); alternatively, t is based on the length of one of the two lineages; that is, $t = \ell_1/r$ or $t = \ell_1/r_A$, where ℓ_A and ℓ_1 are estimated by the ordinary least squares method (C = outgroup).

time (8.2 Myr ago) corresponds with the age of the unique skull of the fossil hominoid *Sivapithecus*[19], which is usually placed on the orangutan lineage, but is about 4 Myr younger than the earliest teeth and jaw fragments assigned to *Sivapithecus*[19]. The *Sivapithecus*–orangutan association itself has been questioned[20,21]. Molecular time estimates among cetartiodactyls (whales and artiodactyls) and for the catarrhine–platyrrhine and feliform–caniform divergences are close to, or only slightly older than, fossil-based estimates.

In contrast, molecular divergence times among sciurognath rodents (Fig. 3) are roughly four times older than their fossil-based estimates[22], as was found previously[1,7,23]. Because these times were estimated from many genes (343) and did not change when a lineage-specific method (Fig. 1b) was used (e.g., a divergence time of 41 Myr ago was obtained for the mouse–rat divergence), the difference cannot be attributed to stochastic error or increased rate of substitution in rodents. Furthermore, increased stringency of the rate-constancy test resulted in similar time estimates (Fig. 2j–l).

Overall, fossil-based and molecular times are in relatively close agreement (Fig. 4a), except for the origin of placental orders and the early history of rodents. The average difference between molecular and fossil-based dates for Mesozoic comparisons is large (30%) (Fig. 4b). An interpretation of this gap in the Mesozoic fossil record is that molecular times are overestimates. However, this is unlikely as many earlier (Palaeozoic) and later (Cenozoic) fossil and molecular dates show close agreement, especially those dates involving taxa with well documented fossil records and where transitional forms are known. Recent findings of 85-Myr-old placental fossils from Central Asia[15–16] are also evidence for the presence of mammalian fossils in this Mesozoic gap.

Our molecular timescale for vertebrate evolution will be useful in calibrating local molecular clocks and in estimating intraordinal divergence times more reliably, especially in groups with poor fossil records. Molecular times also provide an independent measure of the tempo and mode of morphological change. For example, the sudden appearance (in the Early Tertiary fossil record) of mammalian and avian orders, which show large morphological differences, has been taken to imply rapid rates of morphological change at that time[14,24]. Now, the possibility of 20–70 Myr of prior evolutionary history relaxes that assumption and suggests a greater role for Earth history in the evolution of terrestrial vertebrates[12,25]. An accurate knowledge of divergence times can help to direct the search for 'missing' fossils and test hypotheses of macroevolution. □

Methods

Sequence retrieval and tests of molecular clocks. Amino-acid sequences of nuclear genes were obtained from the HOVERGEN[26] database (Genbank Release 97) and all 5,050 gene families were manually examined. Alignments were retrieved whenever data were available to time at least one divergence. Genes under strong positive selection (for example, major histocompatibility complex genes) and sequences with ambiguous orthologies and extensive alignment gaps were excluded. Gene phylogenies were scrutinized further and genes (or sequences) showing extensive rate variation among lineages

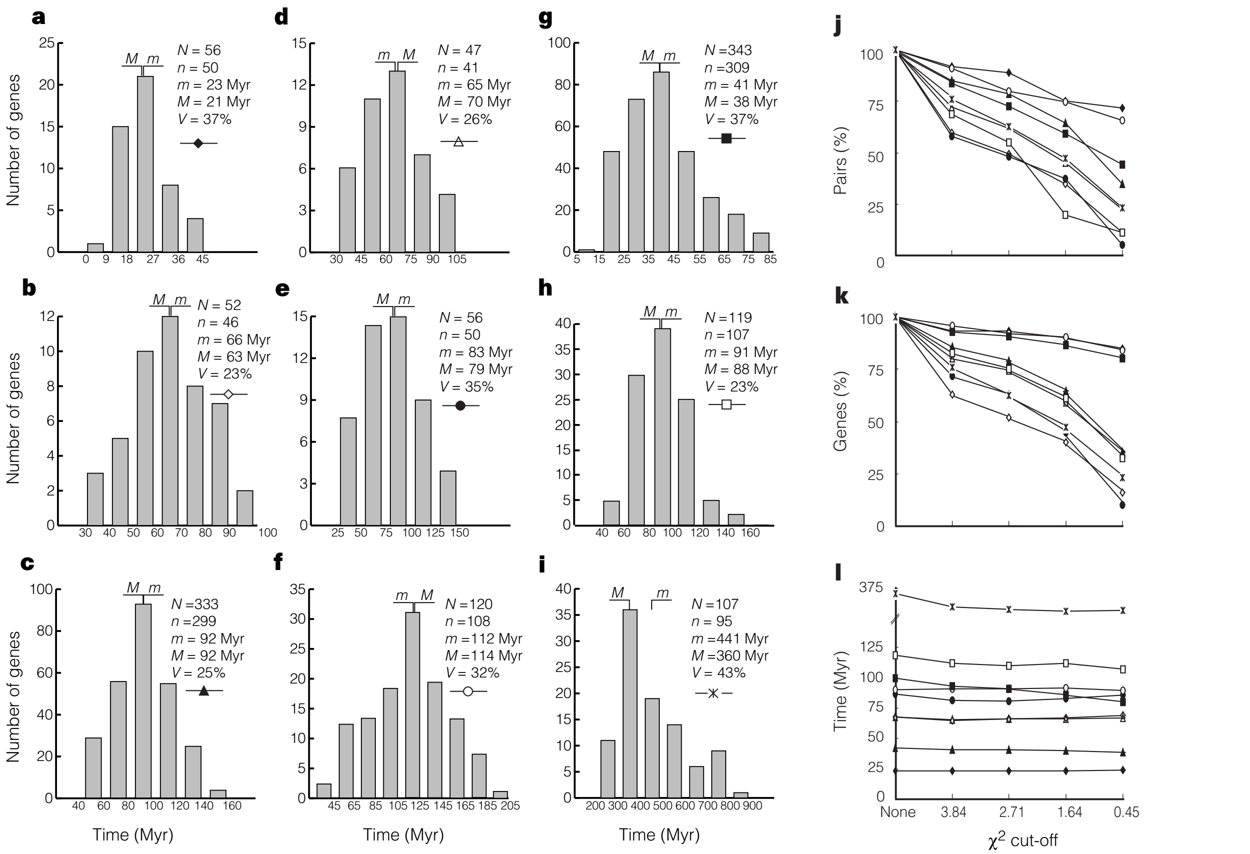

Figure 2 Histograms (**a**–**i**) of distributions of single-gene divergence times for nine multigene time estimates, and graphs (**j**–**l**) of the effects of increased stringency of the rate-constancy test (corresponding to areas of 5% (recommended), 10%, 20%, and 50%, of the χ^2 rejection curve) for the same divergences. Divergence of: **a**, Hominoidea and Cercopithecoidea; **b**, Muridae and Cricetidae; **c**, Archonta and Ferungulata; **d**, Ruminantia and Suidae; **e**, Carnivora+Perissodactyla and Cetartiodactyla; **f**, Rodentia and (Archonta + Ferungulata+Paenungulata); **g**, mouse and rat; **h**, Primates and Lagomorpha; and **i**, Amphibia and Amniota. **j**, Percentage of pairwise comparisons not rejected. **k**, Percentage of genes not rejected. **l**, Time estimates for divergences a–i. Symbols for histograms: *M*, mode; *m*, mean; *N*, total number of genes; *n*, number of genes used after removal of outliers; *V*, coefficient of variation. The locations of *m* and *M* are shown. Myr, millions of years ago.

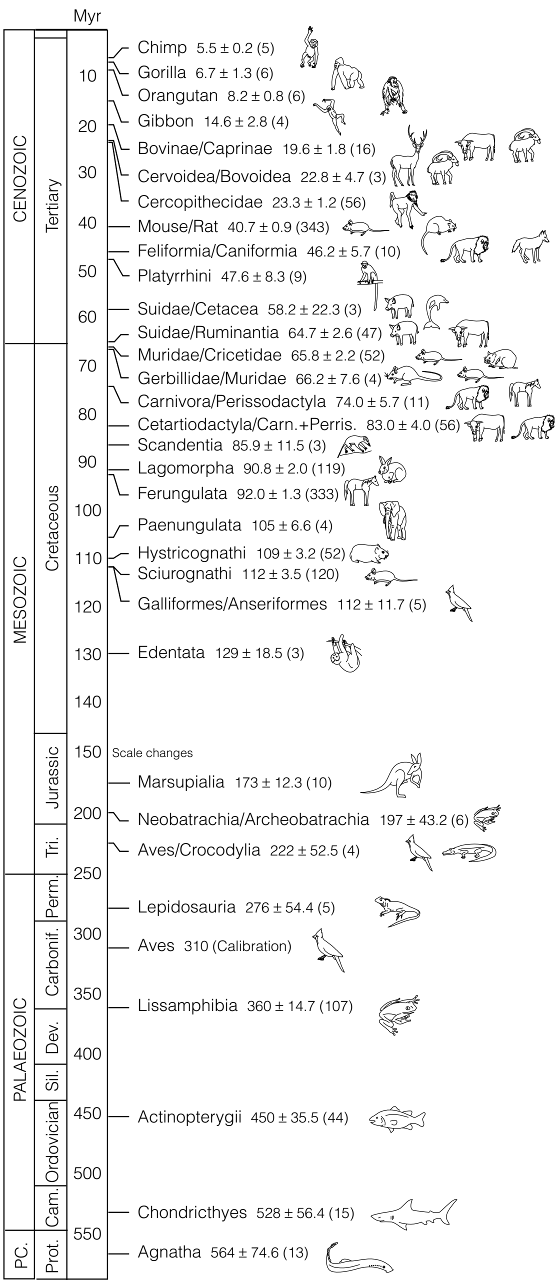

Figure 3 A molecular timescale for vertebrate evolution. All times indicate Myr separating humans (or the largest sister group containing humans) and the group shown, except when the comparative groups are separated by a slash (/). Time estimates are shown with ±1 s.e.m. and the number of genes used is given in parentheses. Three groups of mammalian orders are: Archonta (Primates, Scandentia, Dermoptera, Chiroptera, and Lagomorpha), Ferungulata (Carnivora, Cetartiodactyla, and Perissodactyla), and Paenungulata (Hyracoidea, Proboscidea, Sirenia). Cam, Cambrian; Carbonif, Carboniferous; Dev, Devonian; PC, Precambrian; Perm, Permian; Prot, Proterozoic; Sil, Silurian; Tri, Triassic.

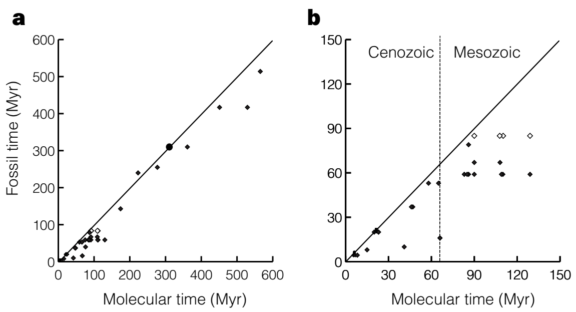

Figure 4 Comparison of fossil-based and molecular estimates of divergence time in vertebrates. **a**, Plot of all values (correlation coefficient = 99%). The solid line indicates a 1:1 relationship; the closed circle represents the calibration point. **b**, Close-up of the region from 0 to 150 Myr ago. Open diamonds show fossil dates based on the 85-Myr placental fossils from Asia[15,16], under the assumption that some represent stem (or basal) ferungulates and archontans. Fossil-based estimates[4,10,14,16,18,20,22,29] were calculated using a method described elsewhere[30]. Myr, millions of years ago.

were removed. The 1DN test[27] was conducted for all relevant triplets of sequences, and pairs in which rate constancy was rejected in any of these comparisons were excluded (22% of pairs out of 7,943 pairs were rejected).

Estimating evolutionary rates for protein-clock calibration. Gene-specific evolutionary rates were estimated by $r_{B\text{–}M} = d_{avg}/(2 \times 310)$ substitutions per site per Myr, where d_{avg} is the average pairwise sequence divergence between bird (B) (*Gallus gallus* (Gga)) and mammal (M) (*Homo sapiens* (Hsa), *Mus musculus* (Mmu), and/or *Rattus norvegicus* (Rno)) sequences. In the absence of bird sequences, we used a mammalian calibration, based indirectly on the bird–mammal calibration, where the divergence time of rodents and primates (or artiodactyls)[12] of 110 Myr ago was obtained from an analysis of 108 genes; d_{avg} was computed by comparing Mmu, Rno, and Hsa (or *Bos taurus*) sequences. The divergence times estimated using the poisson correction distance (presented here) were similar to those estimated using the gamma distance (shape parameter = 2) (ref. 28).

Estimating average divergence time from multiple genes and species. For each gene, the divergence time between two groups was estimated by taking the average of divergence times from all constant-rate species pairs belonging to those groups (Fig. 1b). This procedure was repeated for every gene, and an average multigene time estimate was calculated. Whenever five or more genes were present, the upper and lower 5% of these estimates were excluded (before averaging) to minimize the influence of outliers (at least the highest and lowest time estimates were excluded). Multigene distributions for the amphibian–amniote (Fig. 2i) and actinopterygian–tetrapod divergences included a higher number of early time estimates. As gene orthology was more difficult to determine for those basal groups, we interpreted such unusually high time estimates as representing paralogous comparisons. Therefore, we used the mode to measure the central value because the mean is sensitive to such outliers. We discarded all estimates that were based on only one or two genes.

Received 14 November 1997; accepted 2 February 1998.

1. Wilson, A. C., Carlson, S. S. & White, T. J. Biochemical evolution. *Annu. Rev. Biochem.* **46,** 573–639 (1977).
2. Nei, M. *Molecular Evolutionary Genetics* (Columbia Univ. Press, New York, 1987).
3. Novacek, M. J. Mammalian phylogeny: shaking the tree. *Nature* **356,** 121–125 (1992).
4. Martin, R. D. Primate origins: plugging the gaps. *Nature* **363,** 223–234 (1993).
5. Avise, J. C. *Molecular Markers, Natural History and Evolution* (Chapman & Hall, New York, 1994).
6. Hallam, A. *An Outline of Phanerozoic Biogeography* (Oxford Univ. Press, New York, 1994).
7. Easteal, S., Collet, C. & Betty, D. *The Mammalian Molecular Clock* (R. G. Landes, Austin, TX, 1995).
8. Gerhart, J. & Kirschner, M. *Cells, Embryos, and Evolution* (Blackwell Scientific, Malden, Massachusetts, 1997).
9. Gee, H. *Before the Backbone* (Chapman & Hall, New York, NY, 1996).
10. Benton, M. J. *The Fossil Record 2* (Chapman & Hall, London, 1993).
11. Zuckerkandl, E. On the molecular evolutionary clock. *J. Mol. Evol.* **26,** 34–46 (1987).
12. Hedges, S. B., Parker, P. H., Sibley, C. G. & Kumar, S. Continental breakup and the ordinal diversification of birds and mammals. *Nature* **381,** 226–229 (1996).
13. Takezaki, N., Rzhetsky, A. & Nei, M. Phylogenetic test of the molecular clock and linearized tree. *Mol. Biol. Evol.* **12,** 823–833 (1995).
14. Benton, M. J. *Vertebrate Paleontology* (Chapman & Hall, New York, 1997).
15. Archibald, J. D. Fossil evidence for a late Cretaceous origin of "hoofed" mammals. *Science* **272,** 1150–1153 (1996).
16. Kielan-Jaworowska, Z. Interrelationships of Mesozoic mammals. *Hist. Biol.* **6,** 185–202 (1992).
17. Cooper, A. & Penny, D. Mass survival of birds across the Cretaceous-Tertiary boundary: molecular evidence. *Science* **275,** 1109–1113 (1997).
18. Kay, R. F., Ross, C. & Williams, B. A. Anthropoid origins. *Science* **275,** 797–804 (1997).
19. Ward, S. in *Function, Phylogeny, and Fossils* (eds Begun, D. R., Ward, C. V. & Rose, M. D.) 269–290 (Plenum, New York, 1997).
20. Pilbeam, D. Genetic and morphological records of the Hominoidea and hominid origins: a synthesis. *Mol. Phyl. Evol.* **5,** 155–168 (1996).
21. Benefit, B. R. & McCrossin, M. L. Earliest known Old World monkey skull. *Nature* **388,** 368–371 (1997).
22. Jaeger, J.-J., Tong, H. & Denys, C. The age of the *Mus-Rattus* divergence: paleontological data compared with the molecular clock. *C. R. Acad. Sci. Paris* **302,** 917–922 (1986).
23. Parker, P. H. *An Improved Estimate of the Mouse-Rat Divergence Time and Rates of Amino Acid Substitution in Mammals and Birds* Thesis, Pennsylvania State Univ. (1996).
24. Carroll, R. L. *Vertebrate Paleontology and Evolution* (W. H. Freeman and Co., New York, 1988).

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.

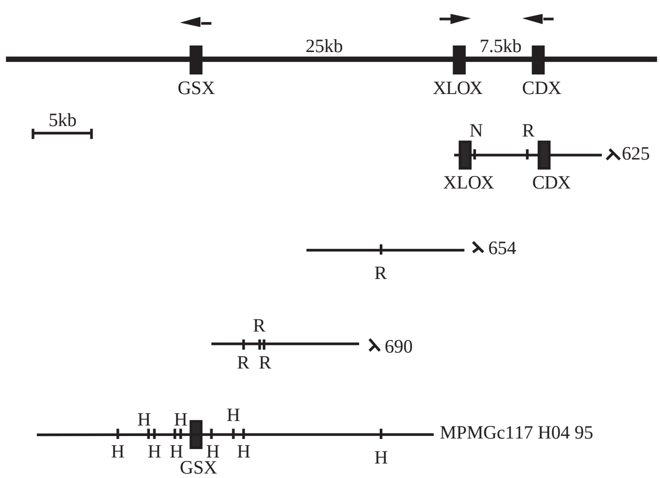

Figure 2 Genomic organization of amphioxus Gsx, Xlox and Cdx genes, showing genomic clones used in walking. Arrows denote transcriptional orientation. R, *Eco*RI, site and N, *Not* I site, mapped in bacteriophage clones only; H, *Hin*dIII sites mapped in cosmid only.