LETTERS

Minimal ProtoHox cluster inferred from bilaterian and cnidarian Hox complements

D. Chourrout[1], F. Delsuc[2], P. Chourrout[3], R. B. Edvardsen[1], F. Rentzsch[1], E. Renfer[1], M. F. Jensen[1], B. Zhu[4], P. de Jong[4], R. E. Steele[5] & U. Technau[1]

Bilaterian animals have a Hox gene cluster essential for patterning the main body axis, and a ParaHox gene cluster. Comparison of Hox and ParaHox genes has led workers to postulate that both clusters originated from the duplication of an ancient cluster named ProtoHox, which contained up to four genes with at least the precursors of anterior and posterior Hox/ParaHox genes[1–3]. However, the way in which genes diversified within the ProtoHox, Hox and ParaHox clusters remains unclear because no systematic study of non-bilaterian animals exists. Here we characterize the full Hox/ParaHox gene complements and genomic organization in two cnidarian species (*Nematostella vectensis* and *Hydra magnipapillata*), and suggest a ProtoHox cluster simpler than originally thought on the basis of three arguments. First, both species possess bilaterian-like anterior Hox genes, but their non-anterior genes do not appear as counterparts of either bilaterian central or posterior genes; second, two clustered ParaHox genes, *Gsx* and a gene related to *Xlox* and *Cdx*, are found in *Nematostella vectensis*; and third, we do not find clear phylogenetic support for a common origin of bilaterian *Cdx* and posterior genes, which might therefore have appeared after the ProtoHox cluster duplication. Consequently, the ProtoHox cluster might have consisted of only two anterior genes. Non-anterior genes could have appeared independently in the Hox and ParaHox clusters, possibly after the separation of bilaterians and cnidarians.

Recent molecular phylogenies support the contention that the non-bilaterian Cnidaria is the sister group to bilaterians[4,5] and can therefore provide information for reconstructing the early history of bilaterian homeobox gene complements. Earlier searches for cnidarian homeobox genes have revealed the presence of anterior-like Hox genes beside *Gsx*, so that the ProtoHox cluster must have been duplicated before the cnidarian–bilaterian split. Previous reports[6–13] have proposed that posterior Hox genes, but not central genes or *Hox3*, are also present in cnidarians. We used the publicly available high-coverage genome shotgun sequences to identify the complete set of homeobox genes of two distantly related cnidarians, the freshwater polyp *Hydra magnipapillata* (Hydrozoa) and the sea anemone *Nematostella vectensis* (Anthozoa). *Nematostella* is particularly informative because it is considered to represent the basal group within the Cnidaria[14,15].

Eighteen *Nematostella* candidate homeodomains were allocated to

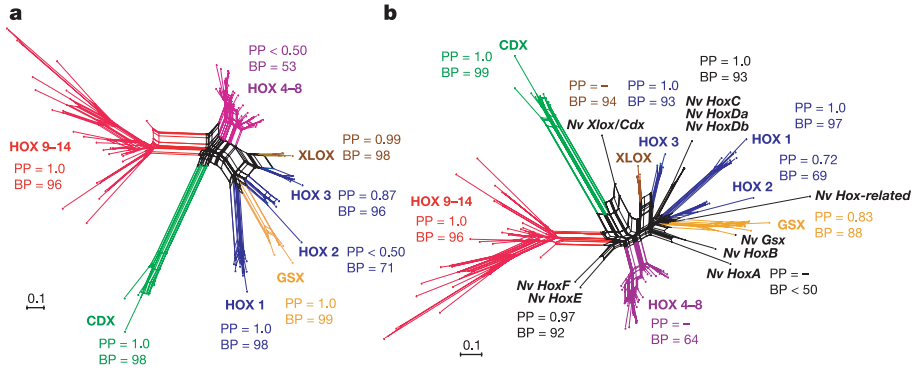

Figure 1 | Phylogenetic analyses of bilaterian and *Nematostella* Hox genes based on their homeodomain. a, Neighbour-net network of 82 bilaterian Hox and ParaHox genes based on maximum-likelihood (ML) distances estimated with a Jones–Taylor–Thornton (JTT) substitution matrix plus an eight-category gamma rate heterogeneity correction (JTT + Γ_8). Note the large differences in evolutionary rates between the fast-evolving posterior Hox and Cdx genes and the slowly evolving central Hox genes. Bayesian posterior probabilities (PP) for the major groups have been mapped on the network. A dash (–) indicates that the corresponding split was not observed in the bayesian analysis. Bootstrap proportions (BP) were inferred from the BioNJ algorithm for the same data set and JTT + Γ_8 ML distance estimates. The scale bar represents the estimated number of changes per site. Note the position of the Cdx group, which is distinct from the posterior Hox group. **b**, Neighbour-net network of 92 Hox and ParaHox genes from bilaterian and *Nematostella* based on JTT + Γ_8 ML distance estimates. Note the position of non-anterior *Nematostella* genes (*Nv HoxE* and *Nv HoxF*). The major groups of bilaterian Hox and ParaHox proteins remain unaffected by the addition of the *Nematostella* data set. Bayesian posterior probabilities (PP) and bootstrap proportions (BP) were calculated as for the bilaterian data set. More detailed figures, including all sequence origins, are available in Supplementary Information.

[1]Sars International Centre for Marine Molecular Biology, University of Bergen, Thormoehlensgt. 55, 5008 Bergen, Norway. [2]Laboratoire de Paléontologie, Phylogénie et Paléobiologie, Institut des Sciences de l'Evolution, UMR5554-CNRS, Université Montpellier II, Place Eugène Bataillon, 34095 Montpellier Cedex 05, France. [3]American Hospital of Paris, 63, boulevard Victor Hugo, BP 109, 92202 Neuilly-sur-Seine Cedex, France. [4]Children's Hospital and Research Center at Oakland, Oakland, California 94609, USA. [5]Department of Biological Chemistry, University of California, Irvine, 240D Medical Sciences I, Irvine, California 92697-1700, USA.

the Antennapedia class/Hox subclass and named on the basis of their relationship with vertebrate and fly sequences. Nine of them were clearly not Hox genes; these included the ParaHox gene *Gsx*, four distinct *Mox* genes, and *Mnx*, *Gbx*, *Rough* and *Evx*. Seven were considered as candidate Hox genes, including five genes already known[13], which we renamed to avoid confusions with bilaterian Hox paralogues: *HoxA* (formerly *anthox6*), *HoxB* (previously unknown), *HoxC* (formerly *anthox7*), *HoxDa* and *HoxDb* (two distinct copies with an identical homeobox of a gene known as *anthox8*), *HoxE* (formerly *anthox1a*) and *HoxF* (formerly *anthox1*). Finally, two previously unknown genes showed higher divergence from Antp/Hox genes of other species (see below). In contrast, a search in the *Hydra* genome revealed only eight Antp/Hox genes, including one *Mox* gene, one *Gsx* gene and six candidate Hox genes.

To clarify the relationships between bilaterian and cnidarian Hox/ParaHox genes, we performed phylogenetic analyses (Supplementary Methods), including the evaluation of alternative hypotheses that receive some support in the data but would not be revealed with traditional tree-building methods. Most reliable trees were obtained with large data sets of homeodomain proteins from putative slowly evolving bilaterian species such as *Nereis virens*, *Cupiennius salei*, *Ptychodera flava* and *Branchiostoma floridae*. First, with the use of bilaterian sequences only, we were able to recognize the major groups of Hox and ParaHox genes (Fig. 1a and Supplementary Information) and confirm that ParaHox genes *Gsx* and *Xlox* cluster with the anterior genes *Hox1/2* and *Hox3*, respectively. Surprisingly, the group of ParaHox *Cdx* genes seemed divergent from all three groups of bilaterian Hox genes and showed only slight phylogenetic affinity (bootstrap proportions = 38) with posterior Hox genes (Fig. 1a and Supplementary Information). The existence of a precursor for Hox posterior genes and *Cdx* in the ProtoHox cluster is therefore questionable.

Neighbour-net analyses clearly showed that within cnidarians *Nematostella* is the slowest evolving species, and the *Nematostella* data set was consequently used for extensive sequence comparisons. *HoxA*, *HoxB*, *HoxC*, *HoxDa*/*Db* and one divergent gene (*Hox-related*) clustered with the anterior genes (*Hox1–3*) of bilaterians (Fig. 1b and Supplementary Information). The precise relationships within the anterior Hox homeodomains could not be statistically resolved despite weak preferential affinities of *HoxA*/*HoxB* with *Hox1/2*, and of *HoxC*/*HoxDa*/*HoxDb* with *Hox3* in several analyses (not shown). By comparison, *HoxE* and *HoxF*, which were previously tentatively classified as posterior genes[13], did not preferentially group with anterior, central or posterior genes (Fig. 1b and Supplementary Information). Instead, they formed an independent and strongly supported branch. Last, one of the two divergent genes showed conflicting affinities in network analyses with *Cdx* and *Xlox* (Fig. 1b) and is therefore most probably a ParaHox gene. This gene, named *Xlox*/*Cdx*, is not a second Gsx gene, according to likelihood-based

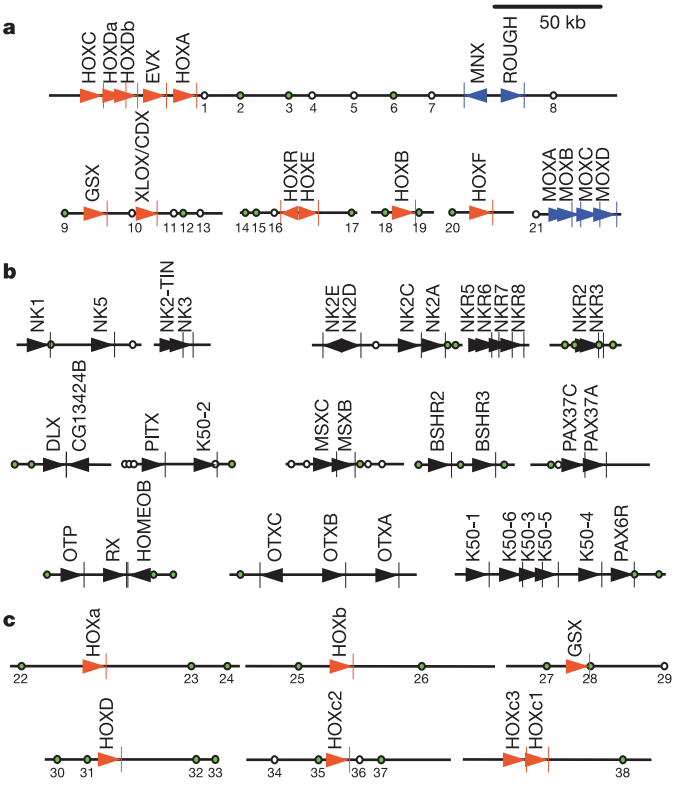

Figure 2 | All physical linkages detected between cnidarian homeobox genes. **a**, Linkages associating *Nematostella* ANTP/Hox-like genes were validated by BAC library screening with individual Hox-like genes, before an *in silico* genome walk using the local assembler of shotgun traces allowed us to reconstruct the cluster DNA sequences. The positions of Hox and Hox-related genes including the ParaHox gene *Gsx* and their respective orientations are represented by red arrows, and other linked homeobox genes by blue arrows. Putative intervening non-homeobox genes are represented by circles, filled with green when expression is confirmed by EST alignments; sequence similarity with known genes was detected by BLASTX (Supplementary Information). **b**, Linkages associating other *Nematostella* homeobox genes were validated with an *in silico* genome walk using the local assembler of shotgun traces. **c**, Linkages associating *Hydra* ANTP/Hox-like genes including *Gsx* were validated with an *in silico* genome walk. Genes found in their environments have been annotated by using BLASTX and alignments with *Hydra* ESTs (Supplementary Information).

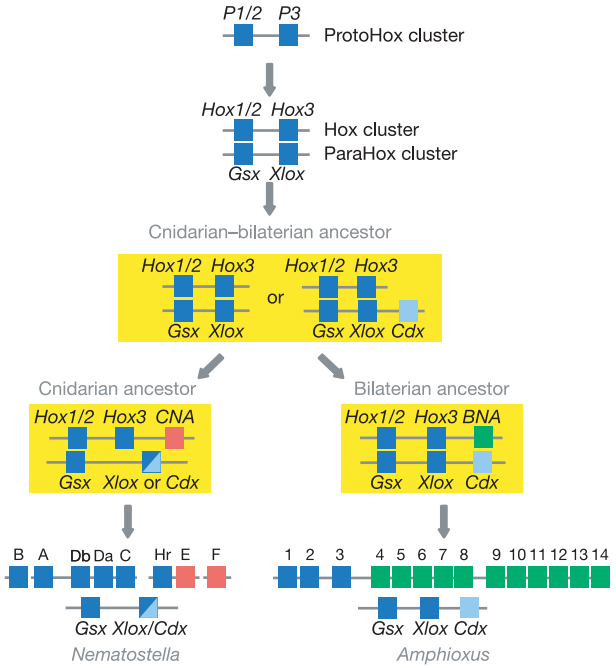

Figure 3 | Parsimonious consensus model for the evolution of Hox/ParaHox clusters of bilaterians. The ProtoHox cluster may have contained only two anterior genes (*P1/2* and *P3*) and its duplication generated two equally simple Hox and ParaHox clusters. The prototypal bilaterian Hox cluster may have contained an extra non-anterior Hox gene (*BNA*), a precursor of future central and posterior genes. An independent duplication in the Hox cluster would have generated the precursor of cnidarian non-anterior Hox genes (*CNA*), which later on became HoxE and HoxF of *Nematostella*. The third ParaHox gene *Cdx* might have appeared through a duplication in the ParaHox cluster, and not in the ProtoHox cluster as generally proposed, either in early bilaterians or before the split of bilaterians and cnidarians. In the latter case, cnidarians had already simplified their ParaHox cluster through the loss of *Xlox*.

statistical tests (Supplementary Information). When Hox genes from other cnidarians were included, the same picture appeared with cnidarian-specific non-anterior Hox genes, although their long branches tended to be attracted towards the fast-evolving posterior genes (Supplementary Information). Overall, cnidarians possess genuine anterior Hox genes, several non-anterior Hox genes not clearly related to central and posterior genes of bilaterians, and *Gsx* and a second candidate ParaHox gene related to *Xlox* and *Cdx*.

Using bacterial artificial chromosome (BAC) library screening and sequence reconstruction from shotgun data, we established linkages between distinct homeobox genes in the *Nematostella* genome. The results (Fig. 2a and Supplementary Information) showed a main cluster of five genes (*HoxC*/*HoxDa*/*HoxDb*/*Evx*/*HoxA*) in a tandem array over about 50 kilobases (kb). Linkage between the non-Hox gene *Evx* and a Hox gene has previously been shown in another anthozoan cnidarian[16]. We extended the cluster sequence and reached two other homeobox genes about 200 kb downstream: one is *Mnx*, a gene also present in the neighbourhood of chordate Hox clusters[17], and the other is the orthologue of the fly gene *Rough*, which is on the same chromosome arm, 3R, as the ANT-C and BX-C complexes (http://www.ncbi.nlm.nih.gov/projects/genome/guide/fly/). Another linkage was established between the anterior gene *Hox-related* and the non-anterior gene *HoxE*, whereas *HoxB* and *HoxF* seemed to be isolated in two distinct genome locations. In sum, linkages in the *Nematostella* genome provide support for the existence of an ancient cnidarian Hox cluster associating *Evx*, several anterior Hox genes and at least one non-anterior Hox gene. In the *Nematostella* lineage, the cluster was expanded through gene duplications. Our analyses robustly support the proposition that two precursor genes have been amplified into *HoxC*/*HoxDa*/*HoxDb* and into *HoxE*/*HoxF* subsets, respectively. A common origin for *HoxA* and *HoxB* is also indicated by the analyses, although with weaker support. The cluster was also affected by a rearrangement placing *Evx* between anterior Hox genes, and by at least two splits. Our linkage studies placed the candidate *Xlox*/*Cdx* gene immediately downstream of *Gsx*, in tandem orientation. Thus, this association can represent a ParaHox cluster. By comparison, genomic sequence reconstruction of the *Hydra* Hox gene complement provided evidence for strong degeneration of the Hox cluster in *Hydra*: only two non-anterior genes, most probably resulting from a recent duplication (*HmHoxc1* and *HmHoxc3*), could be linked (Fig. 2c), and *Gsx* was the only ParaHox gene detected.

As lineage-specific duplications in the Hox cluster are rare within bilaterians, we wanted to know whether or not those found in *Nematostella* are limited to the Hox genes. For this we examined potential linkages between all other homeobox genes. Including the ten Hox-like genes mentioned above, we found a total of 139 homeobox genes, a surprisingly high number in comparison with other invertebrates (Table 1 and Supplementary Information). Phylogenetic analyses allowed the placing of at least 87 of those into 58 known groups of homeobox genes, out of 76 groups known for bilaterians. They also indicated that 42 unclassified genes might have arisen through recent amplifications of at most ten genes. By comparison, *Hydra* has a considerably smaller number of homeobox genes and gene groups. Because all *Hydra* homeobox gene groups have representatives in the *Nematostella* genome, we assume that the homeobox complement has been greatly reduced in the *Hydra* lineage (Table 1). Comparisons between the extended homeobox sequences of *Nematostella* allowed the detection of an additional 13 physical clusters (Fig. 2b and Supplementary Information). Four of these clusters may have a more ancient origin, either because they have been identified in the genomes of bilaterians or because they associate distantly related genes. The other nine clusters are undoubtedly the result of recent tandem duplications. We also did not detect obvious synteny conservation between the environments of well related but unlinked homeobox genes, as might be expected after whole-genome duplication (not shown). Hence the gene duplications observed in the *Nematostella* Hox genes represent a general phenomenon for the homeobox gene complement.

From our data, the most parsimonious model for the evolution of Hox/ParaHox clusters is as follows (Fig. 3). Two ProtoHox genes (*P1/2* and *P3*) gave rise to the Hox cluster consisting of two anterior Hox genes *Hox1/2* and *Hox3*, and to the ParaHox cluster with *Gsx* and *Xlox*. Subsequently, internal duplications expanded the Hox cluster, first by adding a precursor of the non-anterior genes. This gene might have appeared before the cnidarian–bilaterian split or independently in both lineages. Similarly, *Cdx* was added in the ParaHox cluster, either in the bilaterian lineage or before the cnidarian–bilaterian split. After this split, the cnidarian and bilaterian Hox clusters were further expanded through lineage-specific duplications. In this model, cnidarians never had genuine central and posterior Hox genes, and perhaps no *Cdx* gene either. Alternative models can be proposed but they seem to be less parsimonious. For example, the

Table 1 | Classification of homeobox genes from *Nematostella vectensis* and *Hydra magnipapillata* based on phylogenetic analyses

Class/subclass	Superfamily	No. of gene groups (without Hox/ParaHox)			No. of genes (with Hox/ParaHox)	
		Known	*Nematostella*	*Hydra*	*Nematostella*	*Hydra*
Antp/Hox-like	Hox	–	–	–	8	6
	ParaHox	–	–	–	2	1
	Extended-Hox	2	2	1	5	1
	EHG-box	3	3		3	
Antp/NK-like	Placed	20	18	8	35	11
	Unplaced				25	
Prd/Paired-like Q50	Placed	15	9	3	9	5
	Unplaced				8	3
Prd/Paired-like K50	Placed	4	3	3	5	5
	Unplaced				6	1
Prd/Pax S50		4	3	1–2	5	3
Pou		6	4	2	5	2
Lim		6	6	3–4	6	4
Tale		5	5	3	7	4–5
Six		3	3	2	5	2
Prox		1				
Cut		3	1		1	
Zfh		2				
Hnf1		2	1		1	
Unplaced					3	4
Total		76	58	26–28	139	52–53

In *Nematostella*, a significant subset of genes could be allocated to known classes and superfamilies but diverged from known homeobox gene groups, while three genes could not be allocated to known classes (see also Supplementary Information). The *Hydra* genome shows a substantial reduction of the homeobox gene complement, and all identified gene groups are also present in *Nematostella*.

precursor of non-anterior genes could have already existed in the ProtoHox cluster but had then to evolve in multiple directions, leading to central/posterior Hox genes of bilaterians, *HoxE/F* of cnidarians, and *Cdx*. Alternatively, it is possible that there were even more such precursors in the ProtoHox cluster and some were lost in one lineage and/or in one of the two clusters.

Another main angle from which to interpret the existing data is that cnidarians are in fact simplified bilaterians, with a reduced mesoderm as one of the possible regressions[13,18]. A loss of central and posterior genes, or their rapid divergence into current non-anterior Hox genes of cnidarians, could have accompanied the process of simplification, as the Hox cluster breakdowns also would have done. Although this cannot be ruled out completely, extensive expressed sequence tag (EST)[19] and genome surveys in *Nematostella* identified a surprising complexity of genes, including many genes lost in bilaterian lineages. This and the considerable degree of gene sequence conservation of the *Nematostella* lineage argues against a highly reduced and derived organism. In addition, more complex Hox gene complements in more basal metazoan phyla such as Placozoa, which would be indicative of gene loss in Cnidaria, could not be found[20]. We therefore argue that the common ancestor of cnidarians and bilaterians contained rather simple primordial Hox and ParaHox clusters that had distinct fates in Cnidaria and Bilateria. The ProtoHox cluster itself might have consisted of only two anterior genes.

METHODS

Homeodomains were searched within each set of cnidarian genome data with the use of TBLASTN at low stringency (10^{-8}) with representative query sequences from mouse and the other cnidarian and classified into known classes, subclasses and groups. Alignments of shotgun data sets with ESTs had indicated that the probability of a given gene's being represented exceeded 98%. Phylogenetic analyses were based on maximum-likelihood, bayesian and distance-based network analyses of homeodomain proteins. Phylogenetic network analyses allowed us to detect conflicting signals and areas of uncertainty in the data, which appear with homeodomains as a result of reduced alignment lengths. Genomic sequences were extended by using the in-house-developed software MàD (Marche à Droite), which performs walks and jumps using shotgun sequences and their physical links. Linkages between homeobox genes were established through BAC library screening and confirmed with polymerase chain reaction on positive BAC clones, with MàD reconstruction of long sequence contigs. All technical details are provided in Supplementary Information.

Received 11 January; accepted 8 May 2006.

1. Ferrier, D. E. K. & Holland, P. W. Ancient origin of the Hox gene cluster. *Nature Rev. Genet.* **2,** 33–38 (2001).
2. Garcia-Fernandez, J. The genesis and evolution of homeobox gene clusters. *Nature Rev. Genet.* **6,** 881–892 (2005).
3. Brooke, N. M., Garcia-Fernandez, J. & Holland, P. The ParaHox gene cluster is an evolutionary sister of the Hox gene cluster. *Nature* **392,** 920–922 (1998).
4. Collins, A. G. Evaluating multiple alternative hypotheses for the origin of Bilateria: an analysis of 18S molecular evidence. *Proc. Natl Acad. Sci. USA* **95,** 15458–15463 (1998).
5. Philippe, H., Lartillot, N. & Brinkmann, H. Multigene analysis of bilaterian animals corroborate the monophyly of Ecdysozoa, Lophotrochozoa and Protostomia. *Mol. Biol. Evol.* **22,** 1246–1253 (2005).
6. Shenk, M. A., Bode, H. R. & Steele, R. E. Expression of Cnox-2, a HOM/HOX homeobox gene in hydra, is correlated with axial pattern formation. *Development* **117,** 657–667 (1993).
7. Gauchat, D. *et al.* Evolution of Antp-class genes and differential expression of Hydra Hox/paraHox genes in anterior patterning. *Proc. Natl Acad. Sci. USA* **97,** 4493–4498 (2000).
8. Finnerty, J. R. & Martindale, M. Q. Ancient origins of axial patterning genes: Hox genes and ParaHox genes in the Cnidaria. *Evol. Dev.* **1,** 16–23 (1999).
9. Finnerty, J. R. *et al.* Early evolution of a homeobox gene: the parahox gene *Gsx* in the Cnidaria and the Bilateria. *Evol. Dev.* **5,** 331–345 (2003).
10. Kuhn, K., Streit, B. & Schierwater, B. Homeobox genes in the cnidarian *Eleutheria dichotoma*: evolutionary implications for the origin of Antennapedia-class (HOM/Hox) genes. *Mol. Phylogenet. Evol.* **6,** 30–38 (1996).
11. Kuhn, K., Streit, B. & Schierwater, B. Isolation of Hox genes from the scyphozoan *Cassiopeia xamachana*: implications for the early evolution of Hox genes. *J. Exp. Zool.* **285,** 63–75 (1999).
12. Martinez, D. A. *et al.* Cnidarian homeoboxes and the zootype. *Nature* **393,** 748–749 (1998).
13. Finnerty, J. R. *et al.* Origins of bilateral symmetry: *Hox* and *Dpp* expression in a sea anemone. *Science* **304,** 1335–1337 (2004).
14. Collins, A. G. Phylogeny of the Medusozoa and the evolution of the cnidarian life cycles. *J. Evol. Biol.* **15,** 418–432 (2002).
15. Darling, J. A. *et al.* Rising starlet: the starlet sea anemone, *Nematostella vectensis*. *Bioessays* **27,** 211–221 (2005).
16. Miller, D. J. & Miles, A. Homeobox genes and the zootype. *Nature* **365,** 215–216 (1993).
17. Castro, L. F. C. & Holland, P. W. H. Chromosomal mapping of ANTP class homeobox genes in *Amphioxus*: piecing together ancestral genomes. *Evol. Dev.* **5,** 459–465 (2003).
18. Seipel, K. & Schmid, V. Evolution of striated muscle: jellyfish and the origin of triploblasty. *Dev. Biol.* **282,** 14–26 (2005).
19. Technau, U. *et al.* Maintenance of ancestral complexity and non-metazoan genes in two basal cnidarians. *Trends Genet.* **21,** 633–639 (2005).
20. Monteiro, A. S., Schierwater, B., Dellaporta, S. L. & Holland, P. W. H. A low diversity of ANTP class homeobox genes in Placozoa. *Evol. Dev.* **8,** 174–182 (2006).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank D. Rokhsar for his efforts in heading the *Nematostella* Genome project at DoE-JGI, and the J. Craig Venter Institute and the National Human Genome Research Institute for the *Hydra* Genome Project. The initial EST screen in *Nematostella* was funded by the Deutsche Forschungsgemeinschaft, and the construction of the *Nematostella* BAC library was funded by the National Science Foundation (NSF). This is contribution ISEM-2006-37 of the Institut des Sciences de l'Evolution de Montpellier.

Author Contributions The project was conceived and the manuscript written by D.C. and U.T. The bioinformatic identification and genomic assemblies of the homeobox gene complement in the *Hydra* and *Nematostella* genome datasets were conducted by D.C. The software package including MàD was written by P.C., who conceived and trained it with D.C. The phylogenetic analyses were performed by F.D. R.B.E. also participated in the homeobox gene classification and phylogenetic analysis. The cloning of cDNAs and genomic clones of Hox-related genes was performed by U.T., E.R. and M.F.J.; BAC library screening was conducted by U.T. and E.R.; and BAC colony PCR was performed by M.F.J. The *in situ* hybridizations were performed by F.R. R.E.S. and U.T. are members of the *Nematostella* genome consortium, and R.E.S. is a member of the *Hydra* genome consortium. R.E.S., P.deJ. and U.T. were co-principal investigators on the NSF grant for the generation of the *Nematostella* BAC library; U.T. collected the animal material and prepared the DNA; and B.Z. generated the BAC library under the supervision of P.deJ.

Author Information Sequences have been deposited in GenBank with accession numbers DQ500742–DQ500879. Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to D.C. (Daniel.chourrout@sars.uib.no) and U.T. (Ulrich.technau@sars.uib.no).