Published in final edited form as:
Mol Biol Evol. 2005 December ; 22(12): 2386–2394.

Evolutionary Change of the Numbers of Homeobox Genes in Bilateral Animals

Jongmin Nam and **Masatoshi Nei**
Institute of Molecular Evolutionary Genetics and Department of Biology, Pennsylvania State University

Abstract

It has been known that the conservation or diversity of homeobox genes is responsible for the similarity and variability of some of the morphological or physiological characters among different organisms. To gain some insights into the evolutionary pattern of homeobox genes in bilateral animals, we studied the change of the numbers of these genes during the evolution of bilateral animals. We analyzed 2,031 homeodomain sequences compiled from 11 species of bilateral animals ranging from *Caenorhabditis elegans* to humans. Our phylogenetic analysis using a modified reconciled-tree method suggested that there were at least about 88 homeobox genes in the common ancestor of bilateral animals. About 50–60 genes of them have left at least one descendant gene in each of the 11 species studied, suggesting that about 30–40 genes were lost in a lineage-specific manner. Although similar numbers of ancestral genes have survived in each species, vertebrate lineages gained many more genes by duplication than invertebrate lineages, resulting in more than 200 homeobox genes in vertebrates and about 100 in invertebrates. After these gene duplications, a substantial number of old duplicate genes have also been lost in each lineage. Because many old duplicate genes were lost, it is likely that lost genes had already been differentiated from other groups of genes at the time of gene loss. We conclude that both gain and loss of homeobox genes were important for the evolutionary change of phenotypic characters in bilateral animals.

Keywords

homeobox genes; molecular evolution; gene duplication; gene loss; evolutionary developmental biology

Introduction

Homeobox genes that regulate morphogenesis were first discovered by Garber, Kuroiwa, and Gehring (1983) and Scott et al. (1983) in *Drosophila melanogaster* (fruit-fly). Subsequent studies of homeobox genes in fruitflies, frogs, and humans revealed a highly conserved motif of about 180 bp called the homeobox (McGinnis et al. 1984; Scott and Weiner 1984). The homeobox encodes a DNA-binding domain called the homeodomain. In the genomes of animals and plants, homeobox genes form a large transcription factor gene family, with more than 200 genes in humans and about 80 genes in *Arabidopsis*.

Animal homeobox genes were previously classified into about 30 different groups or families based on their sequence similarity and protein domain structure (Burglin 1994). Additional groups of homeobox genes were identified later (e.g., PBC, MEIS, PKNOX/PREP, TGIF, and

E-mail: jnam@caltech.edu..
William Martin, Associate Editor

IRO; Burglin 1997), and now the homeobox genes in animals can be classified into at least 49 different gene families (see Burglin 2005, for a more detailed classification). Member genes of the same family are often functionally related, and different families of homeobox genes are concerned with different aspects of development (Burglin 1994). For example, the genes of the HOX, CDX, and EVX families and their cognate genes play important roles in different steps of body pattern formation during early embryogenesis of animals. The PAX6, SIX, VAX, and EMX gene families are concerned with the development of eyes, whereas the LIM and HMX gene families are important in the development of neurons (reviewed in Duboule 1994). Because of their important roles in development, homeobox genes have been studied extensively by both developmental and evolutionary biologists.

Homeobox genes are generally highly conserved and control similar phenotypic characters among distantly related organisms (reviewed in De Robertis 1994). However, they are also responsible for controlling different phenotypic characters among relatively closely related species (e.g., Galant and Carroll 2002; Ronshaugen, N. McGinnis, and W. McGinnis 2002). The formation of similar phenotypic characters can be explained by the conservation of shared homeobox genes. By contrast, different phenotypic characters are believed to be generated by duplication of homeobox genes and their functional differentiation. It has also been hypothesized that the loss of some homeobox genes are responsible for morphological differentiation (Ruddle et al. 1994). Therefore, it is interesting to study the pattern of duplication and loss of homeobox genes to have some insights into the evolutionary change of phenotypic characters. It is likely that the number of homeobox genes is related to the complexity of organisms.

Although the patterns of gain and loss of homeobox genes belonging to some families have been studied (e.g., Zhang and Nei 1996; Aparicio et al. 1997; Kappen 2000; Wada et al. 2003; Amores et al. 2004; Edvardsen et al. 2005), no one appears to have studied the gain and loss of the entire set of homeobox genes covering diverse bilateral animals. We have therefore decided to study the evolutionary change of the homeobox gene superfamily examining 11 completely or nearly completely sequenced genomes from bilateral animals. For this purpose, we used a modified version of the reconciled-tree method (Goodman et al. 1979; Page and Charleston 1997) taking into account the ambiguity of gene tree. Although our estimates are crude and conservative, we can still obtain a rough picture of the gain and loss of homeobox genes and their significance for morphological evolution. The method used and the results obtained will be presented in this paper.

Materials and Methods

Identification of Homeodomain-Containing Proteins

To find homeodomain-containing proteins, we performed homology search using the tool Psi-Blast (Altschul et al. 1997) for the entire set of annotated proteins of *Caenorhabditis elegans*, *Caenorhabditis briggsae*, mosquito (*Anopheles gambiae*), fruitfly (*D. melanogaster*), tunicate (*Ciona intestinalis*), zebrafish (*Danio rerio*), pufferfish (*Fugu rubripes*), frog (*Xenopus tropicalis*), rat, mouse, and humans. All sequence data except for the tunicate were downloaded from the ENSEMBL (ftp://ftp. ensembl.org) as of February 21, 2005. The tunicate data set (version 1) was downloaded from the Joint Genome Institute (http://genome.jgi-psf.org/). We used 207 homeodomain sequences from animals, plants, and fungi as queries, with an E value $\leq 10^{-5}$ (see Supplementary Material online). We also searched for homeobox genes from the expressed sequence tag (EST) database of the tunicate from the DNA Data Bank of Japan (http://www.ddbj.nig.ac.jp/) because Wada et al. (2003) reported several unannotated homeobox genes from the EST database of this organism.

Phylogenetic Analysis

Because the homeodomain is the only alignable region between different groups of homeodomain-containing proteins, we used only this domain (≈60 aa) for phylogenetic analysis. The homeodomain sequences were aligned against the alignment of 207 query sequences (seed alignments) using the profile alignment of the ClustalX program (Thompson et al. 1997). We then constructed a neighbor-joining tree (Saitou and Nei 1987) using the computer program NJBOOT (Takezaki, Rzhetsky, and Nei 1995) with the pairwise deletion option, proportional amino acid differences (*p*-distances), and 1,000 bootstrap resamplings (Nei and Kumar 2000). Because of the large number of sequences used, other tree construction methods such as maximum-parsimony and maximum-likelihood methods were notused.

Each homeobox gene was assigned to one of the 49 previously defined groups according to the sequence similarity and protein domain structure. Domain structure was examined by using the computer program HMMER (Eddy 2001) for each protein domain profile downloaded from the Pfam (http://pfam.wustl.edu/). A phylogenetic tree for the 49 families of genes was constructed to find their evolutionary relationships.

Estimation of the Number of Genes in the Ancestral Species

When there is a rooted tree of m species, the tree has m − 1 ancestral nodes or species (Nei and Kumar 2000). We are interested in estimating the number of homeobox genes in each of the ancestral species and how the number has changed in the evolutionary process. This can be studied by comparing the species tree with the gene tree for a given set of genes and constructing a reconciled tree (Goodman et al. 1979; Page and Charleston 1997). In this paper, we use a modified version of this reconciled-tree method in which multifurcating branching patterns are taken into account.

For simplicity, let us consider the tree of three species α, β, and γ in figure 1*A* and assume that species α, β, and γ have three, two, and two genes, respectively. Suppose that the gene tree inferred for the seven genes from the three species is given in figure 1*B* and that this tree represents the true gene tree. This gene tree can be decomposed into three groups of genes (I, II, and III), in which genes a, b, and c come from species α, β, and γ, respectively. Group I genes do not include any c gene but contain two pairs of genes a and b. Therefore, to reconcile this portion of the gene tree with the species tree under the principle of parsimony, we must assume that one deletion of gene c and one duplication of the ancestral gene of genes a and b occurred (fig. 1*C*). Similarly, to reconcile the group II genes with the species tree, we will have to assume that gene b was deleted. In the case of group III genes, we have to consider the deletion of genes a and b. In this case, we assume that one event of deletion occurred in the ancestral lineage of genes a and b. The reconciled tree in figure 1*C* therefore suggests that the ancestral species δ had three genes (ancestral genes of the three groups of genes).

A similar inference indicates that the ancestral species ε also contained three genes, i.e., two genes in group I, one gene in group II, and zero gene in group III (fig. 1*C*). Figure 1*A* now shows the change of the number of genes in the evolutionary lineages for α, β, and γ.

In the above estimation of the number of genes, we assumed that the gene tree is correct. In practice, however, some interior branches are often weakly supported in terms of bootstrap values. For example, one interior branch may have a low bootstrap value (<50% in the present study). In this case, the existence of this branch is questionable, so that the length of this branch is reduced to zero, and a condensed tree (Nei and Kumar 2000) is constructed (fig. 1*D*). The real gene tree in this case will be one of the three possible trees given in figure 1*E*–*G*. Tree E is identical with tree B, and we already estimated the number of genes in the ancestral species δ and ε. In the case of tree F, the reconciled tree is given by figure 1*H*, and the numbers of genes in species δ and ε are given in figure 1*I*. This tree is more parsimonious than tree *A* with respect to the change of gene number. Tree *G* is different from tree *F* in that group II genes are closer to gene c, but the number of genes estimated in the ancestral organisms becomes the same as those for tree *B*. Therefore, we assume that tree *I* gives the actual numbers of ancestral species. Note that if low bootstrap support was observed between two clades representing different family of genes, we did not apply this multifurcation rule.

When there are several branches with low bootstrap values, the numbers of genes in ancestral species are estimated by the same procedure as the above under the principle of parsimony. Therefore, one can estimate the number of genes for any number of ancestral species. Obviously, the number of genes estimated would be minimal, but because homeobox genes evolve very slowly, the present method appears to give reasonably good estimates (see below). When m is large, the computation can be quite complicated, and we have developed a computer program (available by request to J.N.).

Results

Number of Homeobox Genes in the Genome

Table 1 shows the numbers of nonredundant homeobox genes obtained from the annotated gene sets of 11 species. The majority of the homeobox genes encode only one homeodomain (single-homeobox genes), but some encode more than one domain (multihomeobox genes). The number of homeodomains encoded by a multihomeobox gene was less than 10 with some exceptions. All vertebrate species studied (pufferfish, zebrafish, frog, mouse, rat, and humans) had about 200 or more homeobox genes, and all invertebrate species (*C. elegans*, *C. briggsae*, fruitfly, mosquito, and tunicate) had about 100 or fewer homeobox genes. All the sequences used are presented as Supplementary Material online (see file 1).

Evolutionary Relationships of Different Families of Homeobox Genes

The majority of the homeobox genes were assigned into the 49 previously defined groups or families. The remaining homeobox genes were either highly divergent or multihomeobox genes. The list of the genes in each of the 49 groups and the multihomeobox genes is available from the Supplementary Material online. Figure 2 shows the evolutionary relationships of the 49 groups of genes. The homeobox gene superfamily was initially classified into two groups, the typical and atypical homeobox gene groups (Burglin 1994). The homeobox of typical genes encodes a 60-aa-long homeodomain composed of three helical regions, and the homeobox of atypical genes encodes additional amino acids either between helices 1 and 2 or between helices 2 and 3 (Burglin 1994). The typical and atypical groups of genes are presented in figure 2. One group of atypical homeobox genes encodes three additional amino acids between helices 1 and 2 and are called TALE (three amino acids loop elongation) class genes. Bharathan et al. (1997) and Burglin (1997) proposed that the typical and TALE homeobox genes diverged before the animal-plant split. Our tree shows the separation of most typical homeobox genes (except SIX group genes) and TALE homeobox genes and supports their notion. However, because we do not know the exact position of the root in the tree, it is difficult to know whether the early evolved groups of typical homeobox genes (e.g., SIX) are evolutionarily closer to TALE genes than to other typical genes or not.

At least 13 groups of homeobox genes (gene groups with orange boxes in fig. 2) encode other evolutionarily conserved domains in addition to the homeodomain according to the hidden Markov model searches for conserved domains using the Pfam database (see Burglin 2005 for the detailed domain structures of homeodomain proteins). Interestingly, a majority of them appear to be early evolved groups of homeobox genes. Therefore, our results suggest that there were already many homeobox genes in the most recent common ancestor (MRCA) of the 11 species, and their domain structures were already quite complex. Because we did not include the multihomeodomain proteins and the proteins that are not assigned to any of the 49 groups, the evolutionary history of the entire set of homeobox genes should be more complex than that shown in figure 2.

Figure 2 also shows the numbers of homeobox genes for each of the 49 families as well as for those of multihomeobox genes and unassigned genes. In most groups, vertebrates have about two to four times more genes than invertebrates. However, there are several gene groups that do not show this pattern. For example, the NOT gene, which is important for the development of notochords in zebrafish (Talbot et al. 1995), has been found as a single-copy gene in three invertebrate species, fruitflies, mosquitoes, and tunicates, and three vertebrate species, zebrafish, pufferfish, and frog, but no gene has been found in three mammalian species, mouse, rat, and humans. Recently, it has been claimed that a mouse ortholog of the NOT gene was found (Abdelkhalek et al. 2004;Plouhinec et al. 2004), but the phylogenetic analysis presented actually suggests that it is a paralog of the NOT genes. There are also other families of genes where vertebrates have fewer or no more genes than those of invertebrates. Because the genome sequencing or annotation of several species has not been completed, the absence of some families of genes should be reexamined, though it may not change the general pattern of evolution. It is also possible that the genes have not been lost completely but only their homeoboxes are missing, and other regions of the gene are still functional as in the case of some PAX genes (Chi and Epstein 2002).

Evolutionary Change of the Number of Homeobox Genes in Bilateral Animals

Knowing that there were already many homeobox genes in the MRCA of all the 11 species (archi-MRCA), we estimated the numbers of homeobox genes in all ancestral organisms and their increase and decrease in different stages of the evolution of bilateral animals. We constructed a phylogenetic tree of 2,031 homeodomain sequences compiled from single- and multihomeodomain–containing proteins in relation to the species tree (fig. 3). The species tree is based on the observation that insects, tunicates, and vertebrates are coelomates but nematodes are pseudocoelomates (Coelomata hypothesis) as well as phylogenetic analyses using more than 100 nuclear proteins from several species (e.g., Blair et al. 2002;Wolf, Rogozin, and Koonin 2004;Philip, Creevey, and McInerney 2005). (The Ecdysozoa hypothesis in which insects and nematodes are sister groups will be considered later.) Because the protein sequences used are generally closely related, we used a bootstrap cutoff point of 50% for generating a multifurcating node of gene trees. As mentioned above, the number of homeoboxes in multihomeobox genes varies with the gene, and therefore it is difficult to estimate the real number of ancestral genes for these genes. We therefore decided to regard each homeobox as one gene. However, this will not affect our results significantly because the number of multihomeobox genes included was small.

To check the reliability of our estimates, we first analyzed HOX family genes. The numbers of ancestral HOX genes at several evolutionary time points have already been estimated by several researchers (e.g., Holland and Garcia-Fernandez 1996; Zhang and Nei 1996; Stellwag 1999; Wada et al. 2003). In the case of HOX genes, estimation of the numbers of ancestral genes is relatively easy because information about the conserved genomic locations of HOX genes can also be used for reconstructing the ancestral states. We compared our estimates of the numbers of ancestral genes with the previous estimates, assuming that the previous estimates are correct (fig. 3*A*). The previous estimates for ancestral species α, β, γ, δ, and ζ in figure 3*A* were 5, 6, 9, 43, and 39, respectively, whereas our estimates were 4, 6, 9, 24, and 26 in this order. This result suggests that our estimates of the numbers of genes in the ancestral species tend to be smaller than the previous ones. This has happened mainly because some of the paralogous homeo-domains from different species had identical or nearly identical sequences, and therefore the interior branches involved sometimes had zero substitutions and showed low bootstrap values. In this case, the numbers of genes in ancestral organisms are expected to be underestimated. In particular, the number of genes in the MRCAs δ and ζ was underestimated for this reason. However, we note that the number of genes gained at the exterior branch of each lineage is likely to have been overestimated because of underestimation of the numbers of ancestral genes.

Increase of Homeobox Genes in the Evolutionary Process

Keeping in mind this possibility of underestimation, we estimated the numbers of ancestral genes and the numbers of genes lost and gained for the entire homeobox gene superfamily. Figure 3*B* shows that there were at least 88 ancestral homeobox genes in the archi-MRCA. This archi-MRCA already had several genes in some homeobox families (e.g., six genes in PAX group) (fig. 2). However, some gene families (LBX, VAX/NOT, MEOX, and zinc-finger [ZF]) were not found in nematodes (fig. 2), and separation of these gene families from other families was not always clear-cut in the tree of 2,031 sequences (data not shown). Therefore, these gene families were not considered in the estimation of genes in the archi-MRCA. However, figure 2 shows that these groups of genes already diverged from other groups of genes before the nematode and mammalian split. Therefore, it is possible that the genes from these four families actually existed in the archi-MRCA. If this is the case, the number of homeobox genes in the archi-MRCA would increase to about 92.

After the divergence of coelomates and pseudocoelomates, the number of homeobox genes increased almost threefold in the vertebrate lineages. In invertebrates, however, the increase was small or moderate, and our results suggest that the number of homeobox genes did not merely increase during the evolutionary process, but the number sometimes decreased. For example, the MRCA of insects and vertebrates had at least 118 homeobox genes, but fruit-flies have 102 at present. Tunicates also have fewer homeobox genes than the MRCA of tunicates and vertebrates. In the case of vertebrate lineages, the number of genes increased primarily in two time periods, that is, the early stages of coelomate evolution (between nodes α and β in fig. 3*B*) and the early stages of vertebrate evolution (between nodes γ and δ in fig. 3*B*). The major increase of gene number in these two time periods are consistent with that of the total number of genes in the genome by X. Gu, Wang, and J. Gu (2002). Note that if a gene is duplicated and one of the two duplicate genes is lost during the same time interval, our approach cannot detect them, as mentioned earlier. Therefore, the gene losses estimated here are losses of fairly old duplicate genes.

The Ecdysozoa hypothesis (e.g., Aguinaldo et al. 1997; H. Dopazo and J. Dopazo 2005) proposes that insects are more closely related to nematodes than to vertebrates. We therefore examined the numbers of ancestral genes for all the MRCAs of the species tree (fig. 4). A conspicuous difference between the two hypotheses was observed in the archi-MRCA (node α), the numbers for the former and the latter hypotheses being 88 and 145, respectively. In other MRCAs, the number was nearly the same for the two hypotheses.

The Coelomata and the Ecdysozoa hypotheses are quite controversial now. However, the studies based on a large number of nuclear protein sequences (e.g., Blair et al. 2002; Wolf, Rogozin, and Koonin 2004; Philip, Creevey, and McInerney 2005) generally support the Coelomata tree. It should also be noted that in our data set the number of gains and losses of genes in the entire evolutionary process is considerably smaller (more parsimonious) for the Coelomata tree than for the Ecdysozoa tree (figs. 3 and 4). This result lends support to the former tree. For these reasons, we will consider primarily the results obtained from the Coelomata tree subsequently.

Retention and Loss of Ancestral Homeobox Genes in Each Species

It is interesting to know how many gene families of the archi-MRCA have left descendent genes in the 11 species and how many gene families have been lost during this evolutionary period. Figure 2 shows that in nematodes 12 of the 49 gene families in the archi-MRCA have been lost, whereas 32 of them have been retained, the remaining 5 gene families existing in neither archi-MRCA nor nematodes. In vertebrates, however, all the archi-MRCA genes were found. In addition, some new gene families originated in insects, tunicates, and vertebrates. The HOX gene family has the largest number of member genes in vertebrates, and the archi-MRCA also had a relatively high number of member genes (four genes). The highest number of archi-MRCA genes was observed in the LIM family, but the number of genes did not increase very much compared with the HOX gene family. Some gene families such as the BSH and VSX families had small numbers of genes in all species.

We also studied the numbers of ancestral homeobox genes lost during the time period from the archi-MRCA to the present species (table 2). Let us consider figure 1*C* to illustrate how we counted the numbers of genes lost. This figure shows three groups of genes (I, II, and III) that were derived from the three genes in the ancestral species δ. In group I, genes a and b are retained, but gene c was lost. In group II, genes a and c are retained, but gene b was lost. Similarly, in group III genes, gene c is retained, but genes a and b were lost. Therefore, in species α, two ancestral genes are retained and one gene was lost. Similarly, the number of genes lost is two in species β and one in species γ. The total number of genes lost in each extant species is given by the sum of these losses for all gene families. Note that this number of gene losses can be computed for each MRCA.

Table 2 shows that the invertebrate lineages lost about 30–38 genes of the 88 ancestral genes in the archi-MRCA, whereas the vertebrate lineages lost about 25–28 genes during the same period. This suggests that invertebrates lost somewhat more genes than vertebrates. However, the difference is much smaller than that observed with full genome analysis, which suggests that about two-fold or more gene losses occurred in the lineage leading to *C. elegans* than the lineage leading to human (Hughes and Friedman 2004;Koonin et al. 2004;Ogura, Ikeo, and Gojobori 2005). Similarly, the numbers of genes lost from the ancestor β to insects and tunicates were somewhat higher than those in the vertebrate lineages (see column 3 in table 2). In vertebrates, the numbers of lost genes are more or less the same for each MRCA. However, because the major increase of gene number occurred in the early stage of vertebrate evolution (between γ and δ), fishes lost somewhat smaller numbers of genes than other vertebrates. These results suggest that the degree of gene loss varies significantly among different families of homeobox genes, but it is not so different among different species.

When the Ecdysozoa tree was used, the number of genes lost from the archi-MRCA is more than two times greater than that for the Coelomata tree (table 3). This indicates that the Coelomata tree is much more parsimonious than the Ecdysozoa tree. Therefore, our data support the former tree, as mentioned earlier.

Discussion

In this study, we showed that there were at least 88 homeobox genes in the archi-MRCA of bilateral animals when the Coelomata tree was used. Previously, we mentioned that our statistical method would give minimum estimates of the numbers of ancestral genes. However, our estimate of the total number of genes in the archi-MRCA is close to the current number of genes in nematodes and insects. This is also true with the number in each gene family. These observations suggest that our estimates may not be too far off from the true numbers. Furthermore, the similarity of the estimates for the archi-MRCA and those for nematodes and insects suggest that the archi-MRCA had the same degree of phenotypic complexity as that of current nematodes or insects. Because vertebrates gained more homeobox genes than invertebrates, it appears that this increase in the number of homeobox genes is responsible for the formation of more complex characters in vertebrates than in invertebrates.

We have also seen that many homeobox genes have been lost in the process of evolution of phenotypic characters. This loss of homeobox genes might have been either inactivation of redundant genes after gene duplication or loss of functionally differentiated genes (Ruddle et al. 1994; Wagner, Amemiya, and Ruddle 2003). The genes lost in our study are losses of fairly old duplicate genes, and therefore it is likely that the genes lost were already functionally differentiated from their paralogous genes at the time of gene loss. This raises the question of why genes could be lost so often. There are at least three possible explanations. First, without closely related paralogous genes, functional redundancy can be achieved by something called distributed robustness (reviewed in Wagner 2005). In other words, loss (or mutation) of a homeobox gene can be buffered by the rewiring of functionally different parts of the regulatory network. If so, it is possible that losses of homeobox genes might not have caused any noticeable changes of phenotypes. Second, it is also possible that gene loss occasionally has had beneficial effects. For example, loss of genes may be related to the reduction of unused characters. Third, the phenotypic changes caused by the loss of homeobox genes might have been more or less neutral with respect to fitness. In the case of multifunctional genes, this is possible if the critical functions are shared by duplicate genes.

The gain and loss of homeobox genes are probably initially opportunistic, but these events may change the evolutionary courses of different organisms. However, the possible causes of gene loss mentioned above are speculative, and more detailed studies are needed to identify the real reason.

Supplementary Material

Supplementary files are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

1. A file containing 207 query sequences.
2. A file containing 2,031 homeodomain sequences.
3. A file containing list of genes for each of 49 families and multihomeodomain proteins.

Acknowledgements

We thank David Geiser, Claudede Pamphilis, Hong Ma, Jan Klein, Junyul Kim, Li Hao, Yoshi Suzuki, and Kerstin Kaufmann for valuable comments on an earlier version of the manuscript. We also thank Thomas Bürglin for sending his up-to-date classification of homeobox genes. This study was supported by the National Institutes of Health Grant GM20293 to M.N.

References

Abdelkhalek HB, Beckers A, Schuster-Gossler K, et al. The mouse homeobox gene Not is required for caudal notochord development and affected by the truncate mutation. Genes Dev 2004;18:1725–1736. [PubMed: 15231714](14 co-authors).

Aguinaldo AM, Turbeville JM, Linford LS, Rivera MC, Garey JR, Raff RA, Lake JA. Evidence for a clade of nematodes, arthropods and other moulting animals. Nature 1997;387:489–493. [PubMed: 9168109]

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–3402. [PubMed: 9254694]

Amores A, Suzuki T, Yan YL, Pomeroy J, Singer A, Amemiya C, Postlethwait JH. Developmental roles of pufferfish Hox clusters and genome evolution in ray-fin fish. Genome Res 2004;14:1–10. [PubMed: 14707165]

Aparicio S, Hawker K, Cottage A, Mikawa Y, Zuo L, Venkatesh B, Chen E, Krumlauf R, Brenner S. Organization of the Fugu rubripes Hox clusters: evidence for continuing evolution of vertebrate Hox complexes. Nat Genet 1997;16:79–83. [PubMed: 9140399]

Bharathan G, Janssen BJ, Kellogg EA, Sinha N. Did homeodomain proteins duplicate before the origin of angiosperms, fungi, and metazoa? Proc Natl Acad Sci USA 1997;94:13749–13753. [PubMed: 9391098]

Blair JE, Ikeo K, Gojobori T, Hedges SB. The evolutionary position of nematodes. BMC Evol Biol 2002;2:1–7. [PubMed: 11801181]

Burglin, T. R. 1994. A comprehensive classification of homeobox genes. Pp. 25–72 in D. Duboule, ed. Guidebook to the homeobox genes. Oxford University Press, New York.

———. 1997Analysis of TALE superclass homeobox genes (MEIS, PBC, KNOX, Iroquois, TGIF) reveals a novel domain conserved between plants and animals. Nucleic Acids Res2541734180 [PubMed: 9336443]

———.2005. Homeodomain proteins. Pp. 179–222 in R. A. Meyers, ed. Encyclopedia of molecular cell biology and molecular medicine. Wiley-VCH Verlag GmbH & Co., Weinheim, Germany.

Chi N, Epstein JA. Getting your Pax straight: Pax proteins in development and disease. Trends Genet 2002;18:41–47. [PubMed: 11750700]

De Robertis, E. M. 1994. The homeobox in cell differentiation and evolution. Pp. 13–23 in D. Duboule, ed. Guidebook to the homeobox genes. Oxford University Press, New York.

Dopazo, H., and J. Dopazo. 2005. Genome-scale evidence of the nematode-arthropod clade. Genome Biol. (in press).

Duboule, D. 1994. Guidebook to the homeobox genes. Oxford University Press, New York.

Eddy, S. R. 2001. HMMER: profile hidden Markov models for biological sequence analysis. (http://hmmer.wustl.edu/).

Edvardsen RB, Seo HC, Jensen MF, et al. Remodelling of the homeobox gene complement in the tunicate Oikopleura dioica. Curr Biol 2005;15:R12–R13. [PubMed: 15649342](11 co-authors).

Galant R, Carroll SB. Evolution of a transcriptional repression domain in an insect Hox protein. Nature 2002;415:910–913. [PubMed: 11859369]

Garber RL, Kuroiwa A, Gehring WJ. Genomic and cDNA clones of the homeotic locus Antennapedia in Drosophila. EMBO J 1983;2:2027–2036. [PubMed: 6416827]

Goodman M, Czelusniak J, Moore GW, Romero-Herrera AE, Matsuda G. Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by cladograms constructed from globin sequences. Syst Zool 1979;28:132–163.

Gu X, Wang Y, Gu J. Age distribution of human gene families shows significant roles of both large- and small-scale duplications in vertebrate evolution. Nat Genet 2002;31:205–209. [PubMed: 12032571]

Holland PW, Garcia-Fernandez J. Hox genes and chordate evolution. Dev Biol 1996;173:382–395. [PubMed: 8605999]

Hughes AL, Friedman R. Shedding genomic ballast: extensive parallel loss of ancestral gene families in animals. J Mol Evol 2004;59:827–833. [PubMed: 15599514]

Kappen C. Analysis of a complete homeobox gene repertoire: implications for the evolution of diversity. Proc Natl Acad Sci USA 2000;97:4481–4486. [PubMed: 10781048]

Koonin EV, Fedorova ND, Jackson JD, et al. A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. Genome Biol 2004;5:R7. [PubMed: 14759257](18 co-authors).

Kumar S, Hedges SB. A molecular timescale for vertebrate evolution. Nature 1998;392:917–920. [PubMed: 9582070]

McGinnis W, Levine MS, Hafen E, Kuroiwa A, Gehring WJ. A conserved DNA sequence in homoeotic genes of the Drosophila Antennapedia and bithorax complexes. Nature 1984;308:428–433. [PubMed: 6323992]

Nei, M., and S. Kumar. 2000. Molecular evolution and phylogenetics. Oxford Press, New York.

Nei M, Xu P, Glazko G. Estimation of divergence times from multiprotein sequences for a few mammalian species and several distantly related organisms. Proc Natl Acad Sci USA 2001;98:2497–2502. [PubMed: 11226267]

Ogura A, Ikeo K, Gojobori T. Estimation of ancestral gene set of bilaterian animals and its implications to dynamic change of gene content in bilaterian evolution. Gene 2005;345:65–71. [PubMed: 15716111]

Page RD, Charleston MA. From gene to organismal phylogeny: reconciled trees and the gene tree/species tree problem. Mol Phylogenet Evol 1997;8:349–362. [PubMed: 9417893]

Philip GK, Creevey CJ, McInerney JO. The Opisthokonta and the Ecdysozoa may not be clades: stronger support for the grouping of plant and animal than for animal and fungi and stronger support for the Coelomata than Ecdysozoa. Mol Biol Evol 2005;22:1175–1184. [PubMed: 15703245]

Plouhinec JL, Granier C, Le Mentec C, Lawson KA, Saberan-Djoneidi D, Aghion J, Shi DL, Collignon J, Mazan S. Identification of the mammalian Not gene via a phylogenomic approach. Gene Expr Patterns 2004;5:11–22. [PubMed: 15533813]

Ronshaugen M, McGinnis N, McGinnis W. Hox protein mutation and macroevolution of the insect body plan. Nature 2002;415:914–917. [PubMed: 11859370]

Ruddle FH, Bentley KL, Murtha MT, Risch N. Gene loss and gain in the evolution of the vertebrates. Dev Suppl 1994:155–161. [PubMed: 7579516]

Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 1987;4:406–425. [PubMed: 3447015]

Scott MP, Weiner AJ. Structural relationships among genes that control development: sequence homology between the Antennapedia, Ultrabithorax, and fushi tarazu loci of Drosophila. Proc Natl Acad Sci USA 1984;81:4115–4119. [PubMed: 6330741]

Scott MP, Weiner AJ, Hazelrigg TI, Polisky BA, Pirrotta V, Scalenghe F, Kaufman TC. The molecular organization of the Antennapedia locus of Drosophila. Cell 1983;35:763–776. [PubMed: 6418389]

Shu DG, Morris SC, Han J, Chen L, Zhang XL, Zhang ZF, Liu HQ, Li Y, Liu JN. Primitive deuterostomes from the Chengjiang Lagerstatte (Lower Cambrian, China). Nature 2001;414:419–424. [PubMed: 11719797]

Stellwag EJ. Hox gene duplication in fish. Semin Cell Dev Biol 1999;10:531–540. [PubMed: 10597637]

Takezaki N, Rzhetsky A, Nei M. Phylogenetic test of the molecular clock and linearized trees. Mol Biol Evol 1995;12:823–833. [PubMed: 7476128]

Talbot WS, Trevarrow B, Halpern ME, Melby AE, Farr G, Postlethwait JH, Jowett T, Kimmel CB, Kimelman D. A homeobox gene essential for zebrafish notochord development. Nature 1995;378:150–157. [PubMed: 7477317]

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG. The ClustalX windows interface, flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res 1997;24:4876–4882. [PubMed: 9396791]

Wada S, Tokuoka M, Shoguchi E, et al. A genomewide survey of developmentally relevant genes in Ciona intestinalis. II. Genes for homeobox transcription factors. Dev Genes Evol 2003;213:222–234. [PubMed: 12736825](13 co-authors).

Wagner A. Distributed robustness versus redundancy as causes of mutational robustness. Bioessays 2005;27:176–188. [PubMed: 15666345]

Wagner GP, Amemiya C, Ruddle F. Hox cluster duplications and the opportunity for evolutionary novelties. Proc Natl Acad Sci USA 2003;100:14603–14606. [PubMed: 14638945]

Wolf YI I, Rogozin B, Koonin EV. Coelomata and not Ecdysozoa: evidence from genome-wide phylogenetic analysis. Genome Res 2004;14:29–36. [PubMed: 14707168]

Zhang J, Nei M. Evolution of Antennapedia-class homeobox genes. Genetics 1996;142:295–303. [PubMed: 8770606]

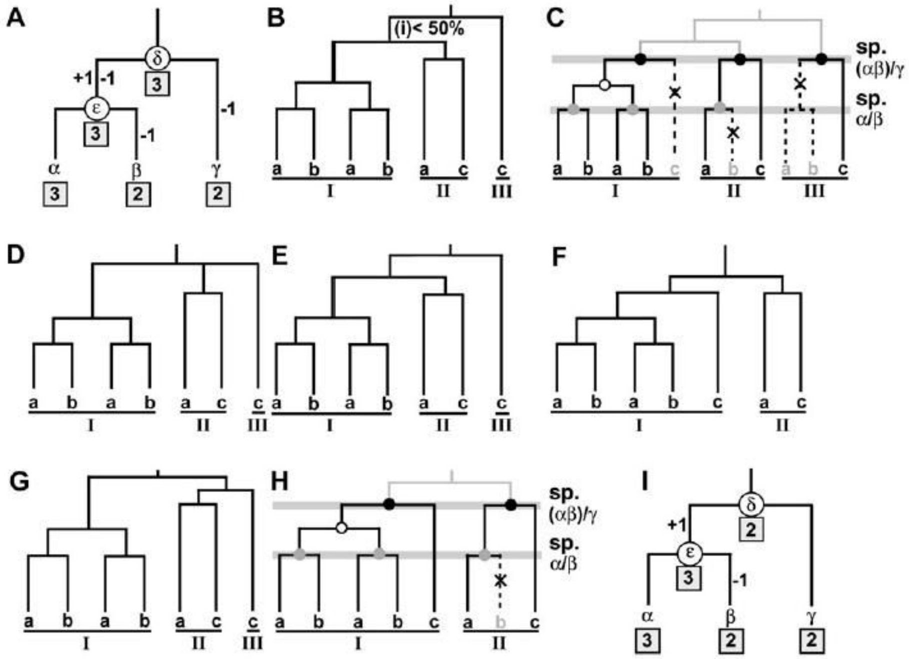
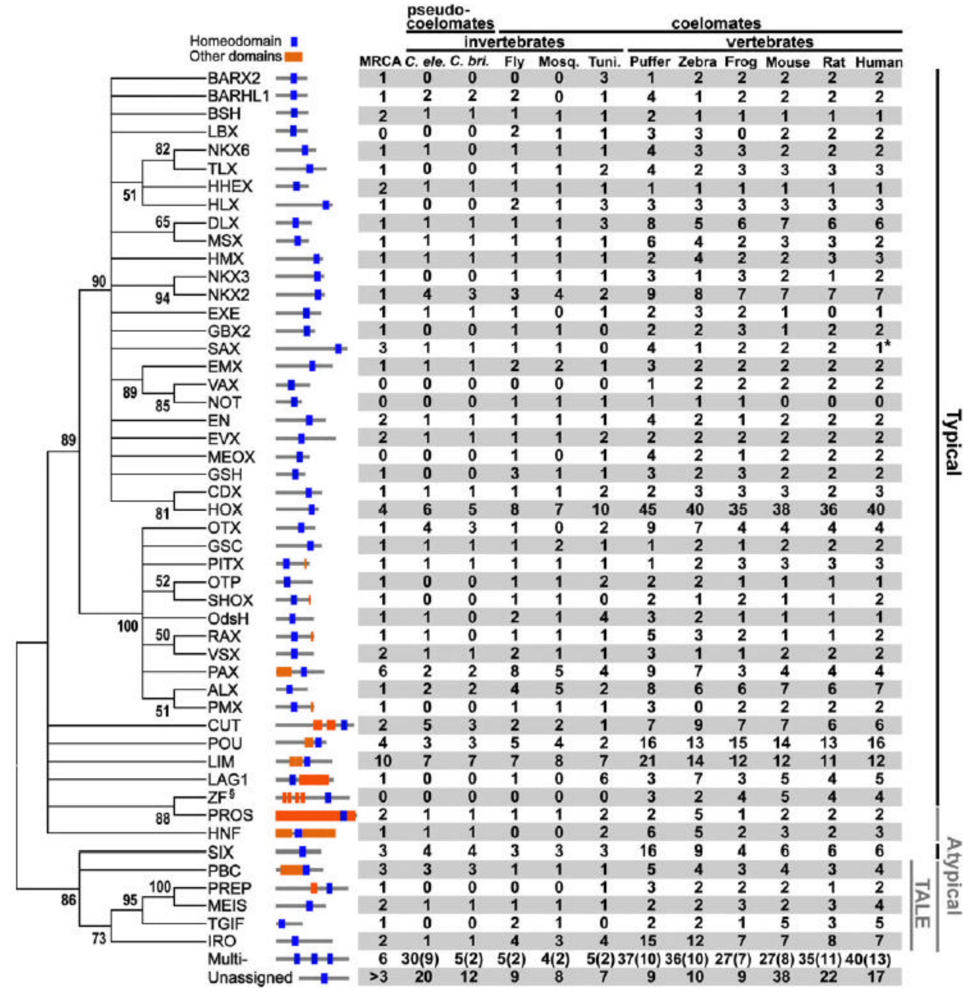


Fig. 1.

A simple example illustrating the method for estimating the numbers of ancestral, gained, and lost genes. We assume that there are three species (species α, β, and γ), and species α, β, and γ have three, two, and two genes, respectively. (*A*) The species tree and the numbers of ancestral, gained, and lost genes. The MRCA of species α, β, and γ and the MRCA of species α and β are labeled δ and ε, respectively. The numbers within square boxes are the numbers of genes in extant species (species α, β, and γ) or ancestral species (species δ and ε). The numbers of genes gained and lost in each ancestral branch are shown on the right and left sides of each branch, respectively. (*B*) The gene tree of the seven genes. (*C*) The reconciled tree of (*A*) and (*B*). Black and gray dots stand for speciation events (sp.), empty black circles for gene duplication events, and crosses for gene losses. (*D*) The condensed tree of (*B*). (*E*)–(*G*) Three possible gene trees that can be inferred from the condensed tree in (*D*). (*H*) The simplest reconciled tree of (*D*). (*I*) The species tree and the numbers of ancestral, gained, and lost genes counted from (*H*).



		pseudo-coelomates		coelomates									
		invertebrates					vertebrates						
	MRCA	*C. ele.*	*C. bri.*	Fly	Mosq.	Tuni.	Puffer	Zebra	Frog	Mouse	Rat	Human
BARX2	1	0	0	0	0	3	1	2	2	2	2	2
BARHL1	1	2	2	2	0	1	4	1	2	2	2	2
BSH	2	1	1	1	1	1	2	1	1	1	1	1
LBX	0	0	0	2	1	1	3	3	0	2	2	2
NKX6	1	1	0	1	1	1	4	3	3	2	2	2
TLX	1	0	0	1	1	2	4	2	3	3	3	3
HHEX	2	1	1	1	1	1	1	1	1	1	1	1
HLX	1	0	0	2	1	3	3	3	3	3	3	3
DLX	1	1	1	1	1	3	8	5	6	7	6	6
MSX	1	1	1	1	1	1	6	4	2	3	3	2
HMX	1	1	1	1	1	1	2	4	2	2	3	3
NKX3	1	0	0	1	1	1	3	1	3	2	1	2
NKX2	1	4	3	3	4	2	9	8	7	7	7	7
EXE	1	1	1	1	0	1	2	3	2	1	0	1
GBX2	1	0	0	1	1	0	2	2	3	1	2	2
SAX	3	1	1	1	1	0	4	1	2	2	2	1*
EMX	1	1	1	2	2	1	3	2	2	2	2	2
VAX	0	0	0	0	0	0	1	2	2	2	2	2
NOT	0	0	0	1	1	1	1	1	1	0	0	0
EN	2	1	1	1	1	1	4	2	1	2	2	2
EVX	2	1	1	1	1	2	2	2	2	2	2	2
MEOX	0	0	0	1	0	1	4	2	1	2	2	2
GSH	1	0	0	3	1	1	3	2	3	2	2	2
CDX	1	1	1	1	1	2	2	3	3	3	2	3
HOX	4	6	5	8	7	10	45	40	35	38	36	40
OTX	1	4	3	1	0	2	9	7	4	4	4	4
GSC	1	1	1	1	2	1	1	2	1	2	2	2
PITX	1	1	1	1	1	1	1	2	3	3	3	3
OTP	1	0	0	1	1	2	2	2	1	1	1	1
SHOX	1	0	0	1	1	0	2	1	2	1	1	2
OdsH	1	1	0	2	1	4	3	2	1	1	1	1
RAX	1	1	0	1	1	1	5	3	2	1	1	2
VSX	2	1	1	2	1	1	3	1	1	2	2	2
PAX	6	2	2	8	5	4	9	7	3	4	4	4
ALX	1	2	2	4	5	2	8	6	6	7	6	7
PMX	1	0	0	1	1	1	3	0	2	2	2	2
CUT	2	5	3	2	2	1	7	9	7	7	6	6
POU	4	3	3	5	4	2	16	13	15	14	13	16
LIM	10	7	7	7	8	7	21	14	12	12	11	12
LAG1	1	0	0	1	0	6	3	7	3	5	4	5
ZF§	0	0	0	0	0	0	3	2	4	5	4	4
PROS	2	1	1	1	1	2	2	5	1	2	2	2
HNF	1	1	1	0	0	2	6	5	2	3	2	3
SIX	3	4	4	3	3	3	16	9	4	6	6	6
PBC	3	3	3	1	1	1	5	4	3	4	3	4
PREP	1	0	0	0	0	1	3	2	2	2	1	2
MEIS	2	1	1	1	1	1	2	2	3	2	3	4
TGIF	1	0	0	2	1	0	2	2	1	5	3	5
IRO	2	1	1	4	3	4	15	12	7	7	8	7
Multi-	6	30(9)	5(2)	5(2)	4(2)	5(2)	37(10)	36(10)	27(7)	27(8)	35(11)	40(13)
Unassigned	>3	20	12	9	8	7	9	10	9	38	22	17

Fig. 2.

Evolutionary relationships of 49 different families of homeobox genes and their phylogenetic distribution in the 11 species of bilateral animals. The tree is constructed by the neighbor-joining method using average *p*-distances between 49 groups and is a 50% bootstrap consensus tree (100 bootstrap replications). Bootstrap values higher than 50% are shown. Representative domain organization is shown on the right-hand side of each family name. Black vertical lines indicate the typical homeobox gene family and gray vertical lines the atypical homeobox gene family. Each blue square indicates a homeodomain, and orange squares indicate the conserved family-specific domains. Gray horizontal lines indicate full-length proteins. Domains of *E* value < 0.01 in the HMM search are shown. The numbers of homeobox genes for each family in each species is also shown. Numbers in parentheses are the numbers of homeoboxes from multihomeobox genes. Numbers under ‘‘MRCA’’ are the estimated numbers of homeobox genes in the MRCA of the 11 species using species trees based on the Coelomata hypothesis. No SAX family gene was found in the annotated data set of human genes from the ENSEMBL. However, the annotation data set of human genes from the GenBank contains one copy of SAX group gene. We therefore included this gene in this tabulation (number with ‘‘*’’ mark). We did not show any ZF homeobox genes in invertebrates, though there are ZF homeobox genes in these animals. This is because the ZF homeobox genes (subfamily with ‘‘§’’ mark) in invertebrates are either ZF multihomeobox genes that have their own vertebrate multihomeobox orthologs or unclassified genes that appear to have been derived from multihomeobox genes (data not shown). Note that the SIX family genes are typical homeobox genes and that the gene numbers for the HOX gene family are slightly higher than those of the genes in the HOX cluster. This is because other closely related genes (e.g., IPF) were also included in this family. We used the same notations as Burglin's (2005) to represent homeobox gene families.

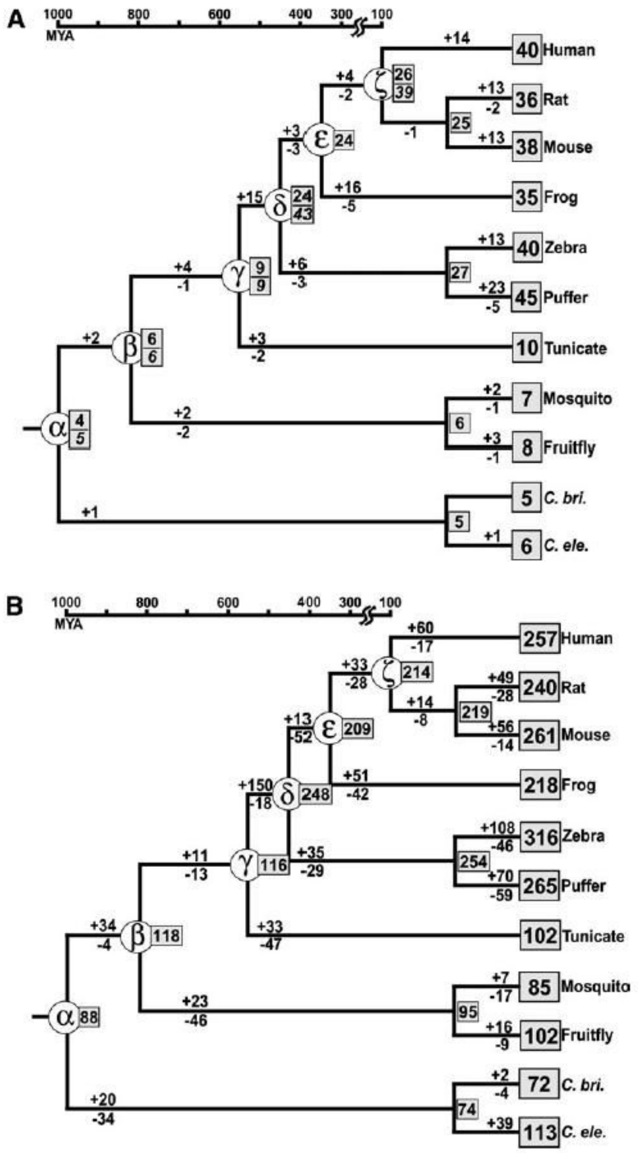


Fig. 3.

Estimated numbers of ancestral, gained, and lost genes during the evolution of bilateral animals when the Coelomata tree is used. Species name is given on the right-hand side of each external node. Ancestral species of our interest are labeled by α to ζ. The number within a square box is the number of genes in each extant species or ancestral species. The numbers above and below each branch are the numbers of gained and lost genes, respectively. The divergence times for ancestral nodes α, β, δ, ε, and ζ are based on the molecular clock (Kumar and Hedges 1998; Nei, Xu, and Glazko 2001) and that for node γ is based on the fossil record (Shu et al. 2001). The remaining ancestral nodes are not on the time-scale. (*A*) Evolution of the HOX family genes. For ancestral nodes α, β, γ, δ, and ζ, the numbers in italic are the numbers of ancestral HOX genes estimated by other studies, and those above the italic are the estimated numbers in this study. Other studies are as follows: node α, Zhang and Nei (1996); node β, Holland and Garcia-Fernandez (1996); node γ, Wada et al. (2003); and nodes δ and ζ, Stellwag (1999). (*B*) Evolution of the entire homeobox gene superfamily. The numbers of gained and lost genes for the exterior branches are not so reliable (see *Results*). The notations used are the same as those of (*A*).

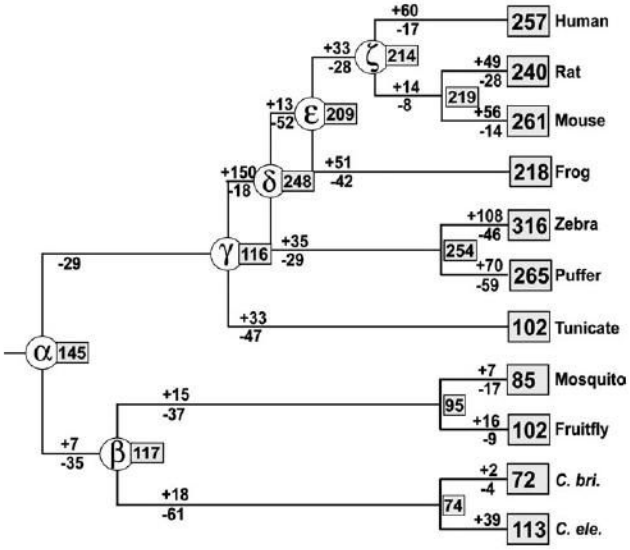


Fig. 4.

Estimated numbers of ancestral, gained, and lost genes during the evolution of bilateral animals when the Ecdysozoa tree is used. The notations used are the same as those in figure 3.

Table 1

Estimates of the Numbers of Homeobox Genes in 11 Species and the MRCA of All 11 Species (Archi-MRCA)

		Nematodes		Insects			Teleosts			Rodents		
	Archi-MRCA	***Caenorhabditis elegans***	***Caenorhabditis briggsae***	**Fruitfly**	**Mosquito**	**Tunicate**	**Pufferfish**	**Zebrafish**	**Frog**	**Mouse**	**Rat**	**Humans**
Single-homeodomain proteins	82	83	68	97	81	97	229	279	191	234	205	217
Multihomeodomain proteins	(6)	9 (30)	2 (4)	2 (5)	2 (4)	2 (5)	10 (36)	10 (37)	7 (27)	8 (27)	11 (35)	13 (40)
Total number of proteins (homeodomains)	(88)	92 (113)	70 (72)	99 (102)	83 (85)	99 (102)	239 (265)	289 (316)	198 (218)	242 (261)	216 (240)	230 (257)

NOTE—The estimates of the numbers of homeobox genes are shown outside parentheses, and those of homeodomains are shown in parentheses. These two numbers are different because of multihomeodomain proteins.

Table 2

Estimates of the Numbers of Genes Lost from Each MRCA to Extant Species When the Coelomata Tree Was Used

	Ancestral Nodes in Figure 3					
	α	**β**	**γ**	**δ**	**ε**	**ζ**
Species	**88**	**118**	**116**	**248**	**209**	**214**
Caenorhabditis elegans	34					
Caenorhabditis briggsae	38					
Fruitfly	30	53				
Mosquito	35	59				
Tunicate	35	57	47			
Zebrafish	25	40	29	71		
Pufferfish	27	43	31	78		
Frog	26	43	34	90	42	
Mouse	26	43	34	91	41	26
Rat	28	46	38	97	47	36
Humans	26	42	33	86	34	17

Note.—Estimates of the numbers of genes lost in each species from the same MRCA are shown in the same column.

Table 3

Estimates of the Numbers of Genes Lost from Each MRCA to Extant Species When the Ecdysozoa Tree Was Used

	Ancestral Nodes in Figure 4					
	α	**β**	**γ**	**δ**	**ε**	**ζ**
Species	**145**	**117**	**116**	**248**	**209**	**214**
Caenorhabditis elegans	96	61				
Caenorhabditis briggsae	100	65				
Fruitfly	76	46				
Mosquito	83	52				
Tunicate	79		47			
Zebrafish	60		29	71		
Pufferfish	63		31	78		
Frog	64		34	90	42	
Mouse	63		34	91	41	26
Rat	66		38	97	47	36
Humans	62		33	86	34	17

Note.—Estimates of the numbers of genes lost in each species from the same MRCA are shown in the same column.