J Mol Evol (2007) 65:137–153
DOI: 10.1007/s00239-006-0023-0

JOURNAL OF MOLECULAR EVOLUTION

# Comprehensive Analysis of Animal TALE Homeobox Genes: New Conserved Motifs and Cases of Accelerated Evolution

**Krishanu Mukherjee,[1] Thomas R. Bürglin[1,2]**

[1] Department of Biosciences and Nutrition, Karolinska Institutet, and School of Life Sciences, Södertörns Högskola, Huddinge, Sweden
[2] Center for Genomics and Bioinformatics, Karolinska Institutet, Stockholm, Sweden

Received: 30 January 2006 / Accepted: 21 November 2006 [*Reviewing Editor*: Dr. Stuart Newfeld]

**Abstract.** TALE homeodomain proteins are an ancient subgroup within the group of homeodomain transcription factors that play important roles in animal, plant, and fungal development. We have extracted the full complement of TALE superclass homeobox genes from the genome projects of seven protostomes, seven deuterostomes, and Nematostella. This was supplemented with TALE homeobox genes from additional species and phylogenetic analyses were carried out with 276 sequences. We found 20 homeobox genes and 4 pseudogenes in humans, 21 genes in mouse, 8 genes in Drosophila, and 5 genes plus one truncated gene in *Caenorhabditis elegans*. Apart from the previously identified TALE classes MEIS, PBC, IRO, and TGIF, a novel class is identified, termed MOHAWK (MKX). Further, we show that the MEIS class can be divided into two families, PREP and MEIS. Prep genes have previously only been described in vertebrates but are lacking in Drosophila. Here we identify orthologues in other insect taxa as well as in the cnidarian Nematostella. In *C. elegans*, a divergent Prep protein has lost the homeodomain. Full-length multiple sequence alignment of the protostome and deuterostome sequences allowed us to identify several novel conserved motifs within the MKX, TGIF, and MEIS classes. Phylogenetic analyses revealed fast-evolving PBC class genes; in particular, some X-linked PBC genes in nematodes are subject to rapid evolution. In addition, several instances of gene loss were identified. In conclusion, our comprehensive analysis provides a defining framework for the classification of animal TALE homeobox genes and the understanding of their evolution.

**Key words:** Homeobox — TALE — MEIS — PREP — TGIF — PBC — MOHAWK — MKX — IRO

## Introduction

Homeobox genes encode a protein domain, the homeodomain, that is about 60 amino acids long and has been shown to bind DNA (Bürglin 2005; Gehring et al. 1994). Homeodomain proteins are highly conserved in evolution and are found in plants, fungi, and animals. They regulate many embryonic developmental programs including axis formation, limb development, and organogenesis (Duboule 1994). Mutations in homeobox genes have been implicated in a number of diseases, including neuroblastoma, Currarino syndrome, leukemia, and cancer (Belloni et al. 2000; Fu et al. 2003; Geerts et al. 2003; Man et al. 2005; Maulbecker and Gruss 1993; Perri et al. 2005; Soulier et al. 2005; Thorsteinsdottir et al. 1997). Homeodomain proteins are classified into different classes based on the sequence of the homeodomain itself, as well as on the basis of additional conserved domains found in these proteins (Bürglin 2005). The **t**hree-**a**mino acid-**l**oop **e**xtension, or TALE, superclass is characterized by the fact that its members

*Correspondence to:* Thomas R. Bürglin, Alfred Nobels Allé 7, School of Life Sciences, Södertörns Högskola, SE-141 89 Huddinge, Sweden; *email:* thomas.burglin@biosci.ki.se

contain three extra residues between helix 1 and helix 2 of the homeodomain (Bertolino et al. 1995). Some of the TALE homeodomain proteins are important cofactors for the homeodomain proteins of the HOX cluster, which are involved in patterning along the anterior-posterior body axis (Bürglin 1998a; Fognani et al. 2002; Huang et al. 2005; Mann and Affolter 1998).

The TALE superclass has been divided into four classes in animals: PBC, MEIS, TGIF, and IRO (Iroquois) (Bürglin 1997). The PBC class was first defined by the vertebrate Pbx genes and *C. elegans ceh-20* (Bürglin and Ruvkun 1992) and also includes the well-characterized Drosophila gene *extradenticle* (*exd*) (Abu-Shaar et al. 1999; Flegel et al. 1993; Rauskolb et al. 1993). The sequence similarity between the PBC proteins extends downstream of the homeodomain for about 15 amino acids, and upstream of the homeodomain is a 180-amino acid domain termed PBC (Bürglin and Ruvkun 1992). The MEIS class was defined by the vertebrate Meis genes, *C. elegans unc-62* (*ceh-25*) and Drosophila *homothorax* (*hth*). Upstream of the homeodomain is the 130-amino acid-long MEIS domain, also referred to as the HM domain (Bürglin 1997; Rieckhof et al. 1997). The vertebrate Prep genes (some are known by the name Pknox) also encode a MEIS domain upstream of the homeodomain. They have been classified as MEIS class genes (Berthelsen et al. 1998; Bürglin 1998a; Van Auken et al. 2002), because no orthologues exist in Drosophila, and in the absence of genome information from other protostomes it was thought that the Prep genes might be divergent MEIS members. However, there are some differences between the MEIS domains of the Meis and Prep proteins, and the Prep genes have been placed in a separate "subfamily" (Fognani et al. 2002). Here we show that orthologues for the Prep genes are present in insects such as mosquito, honeybee, red flower beetle, silk worm, and wasp, but are lacking in Drosophila. While none of the animal TALE classes directly exists in fungi or plants, some sequence similarity between the MEIS and PBC domains upstream of the homeodomain and the KNOX domain in plant KNOX homeodomain proteins has been detected. This indicates that the MEIS and PBC class genes represent a more ancestral type and are derived from an ancestral gene present in the last common ancestor between plants and animals, which encoded a "MEINOX" domain and a TALE domain (Bürglin 1997, 1998a).

The IRO class of genes was defined by the three genes in the Iroquois complex in Drosophila, *araucan* (*ara*), *caupolican* (*caup*), and *mirror* (*mirr*) and the vertebrate IRX genes. A small protein motif, the IRO box, is found in the C-terminal region of the IRO proteins (Bürglin 1997). The TGIF class was originally defined solely by the vertebrate TGIF proteins, but later the Drosophila members *achinta* (*achi*) and *vismay* (*vis*) were identified (Bürglin 1997; Hyman et al. 2003).

Several genome projects of very high quality are now available, i.e., *Caenorhabditis elegans* (The *C. elegans* Sequencing Consortium 1998), which was finalized and had no gaps remaining in November 2002, *Drosophila melanogaster* (Adams et al. 2000), and *Homo sapiens* (Consortium 2004), which was recently completed (Gregory et al. 2006). These well-completed genomes from distinct phyla prompted us to perform exhausting database searches to identify all TALE homeobox genes in these species. We supplemented these data with TALE homeobox genes from other taxa, whose genomes we thoroughly scanned, to get a complete overview of the evolution of these genes in metazoa. In particular, homeobox genes recovered from recent genome data of the cnidarian *Nematostella vectensis* (Sullivan et al. 2006) provide a valuable basal reference group. We performed multiple sequence alignments of the complete coding regions to identify all conserved motifs within the different groups. Our analyses uncovered a novel TALE class, defined by largely uncharacterized vertebrate genes and an uncharacterized Drosophila gene; a characterization of the mouse gene, named Mohawk, has been published during revision of this study (Anderson et al. 2006). For further analysis, we used maximum likelihood (ML) and other phylogenetic methods to analyze the homeodomain sequences. A ratio of nonsynonymous-to-synonymous substitutions (Ka/Ks) > 1 is an indicator of the positive Darwinian selection (Bielawski and Yang 2003). We have identified diverging TALE homeobox genes that show signs of such positive selection, but in general genes in the TALE class are highly conserved.

## Materials and Methods

### *Sequence Retrieval, Multiple Sequence Alignment, and Phylogenetic Analyses*

*C. elegans* and *Caenorhabditis briggsae* homeobox genes were retrieved from the Sanger center (http://www.sanger.ac.uk), ENSEMBL (http://www.ensembl.org), and NCBI (http://www.ncbi.nlm.nih.gov/BLAST/) using tblastn and blastp (Altschul et al. 1997) with selected TALE homeodomains as query (Bürglin 1997). *Caenorhabditis remanei* sequences were recovered from Wormbase (http://www.wormbase.org). Drosophila and human homeobox genes were recovered from ENSEMBL (Hubbard et al. 2005) using tblastn against the genomic sequence to retrieve homeobox sequences. Likewise, we searched ENSEMBL at the genomic as well as the annotated protein sequence level to retrieve all TALE proteins from *Apis mellifera*, *Anopheles gambiae*, *Canis familiaris*, *Mus musculus, Rattus norvegicus, Gallus gallus*, *Xenopus tropicalis*, and *Danio rerio*. *Tribolium castaneum* sequences were retrieved using tblastn from beetlebase (http://www.bioinformatics.ksu.edu/blast/blast.html) and flybase (http://www.flybase.net/blast/), the latter of which was also used for *Drosophila pseudoobscura*. *Nematostella vectensis* sequences were retrieved from StellaBase (http://www.stellabase.org). Additional sequences from other species were included (see species table) that were retrieved by blastp searches of the nonredundant NCBI database. Insect homologues of Drosophila CG11617 were retrieved using the DroSpeGe database (http://insects.eugenes.org/DroSpeGe/). *Brugia malayi* was searched using tblastn at http://www.tigr.org/. The TIGR sequencing effort is part of the International Brugia Genome Sequencing Project. *Pristionchus pacificus* shotgun assemblies were searched using tblastn at http://www.pristionchus.org/.

Default parameters were used for the blast searches, and results were manually scanned to detect all possible TALE homeobox genes. A local database was created using Filemaker® Pro 6 (Filemaker, Inc.) to store the retrieved sequences, annotations, accession numbers, and other information. In several instances, errors in the automatic predictions were uncovered, for example, there were obvious gaps in the conserved domains of the mosquito and honeybee Prep proteins. In such cases gene prediction was carried out with the genomic sequence using the MIT GENSCAN server (http://genes.mit.edu/GENSCAN.html), as well as the FGENESH+ server at Softberry (http://www.softberry.com); the latter allows the use of an orthologous protein sequence as a guide, which improves prediction accuracy. If necessary, ORFs were corrected manually following three-frame translation and by comparing the genomic sequences to closely related protein sequences using the dotplot program PPCMatrix (Bürglin 1998b). For example, Bombyx Prep was manually predicted from two contigs (contig 60331 and contig 467452). The correct ORF could not be predicted in all instances because of gaps and incomplete sequence information. Such problematic areas can be recognized in the multiple alignments when a sequence displays an uncharacteristic gap in a conserved motif. In regions outside conserved motifs, prediction accuracy is reduced, because sequence similarity does not serve as a guide.

Homeodomain sequences as well as full-length protein sequences from each group were aligned in MUSCLE (Edgar 2004). Alignments were visualized in Clustal_X 1.83 (Thompson et al. 1997) in Mac Os Classic (Apple Computer®), and the color coding of the alignments is according to the default of Clustal_X. ML-based tree building was performed using the program PHYML (Guindon and Gascuel 2003) to infer the phylogenetic relationship among the TALE homeobox genes. To construct the ML tree, the JTT substitution model (Jones et al. 1992) was used with 100 bootstrap replicates. Several typical 60-amino acid homeodomain sequences were taken as outgroups to root the trees. Additionally, we used neighbor joining (NJ) for tree building as implemented in Clustal_X (Thompson et al. 1997). The protein logo was generated using the Java application LogoBar (Pérez-Bercoff et al. 2005).

### *Estimation of Nonsynonymous/Synonymous Substitution Ratios and Tests for Positive Selection for Sequences from Closely Related Species*

Full-length DNA sequences were retrieved for each individual protein and the nucleotide sequence of each homeobox was extracted using PPC Matrix (Bürglin 1998b). The extracted sequences were aligned pairwise using the "needle" application of EMBOSS (Olson 2002) running on Mac OS X as provided in compiled form by Erik Bongcam-Rudloff (http://www.ebioinformatics.org/). The DNA sequence alignment was compared with the protein alignment and corrected manually in a text editor, if necessary. Similarly, DNA alignment for the full-length sequences was performed using Clustal_X or needle followed by either manual adjustment based on the protein alignment or using the recent tool PAL2NAL (Suyama et al. 2006).

For analysis of positive selection we used the codon substitution models for heterogeneous selection pressure of Yang et al. (2000) as implemented in the codeml program of the PAML package (Yang 1997) (http://abacus.gene.ucl.ac.uk/software/paml.html). Models M0, M3, M7, and M8 were used to infer codon sites that were under putative positive selection pressure involving excess of nonsynonymous substitution. The likelihood ratio test (LRT) was used to compare the M0 (single Ka/Ks over the compared sequences) and M3 (three discrete Ka/Ks ratios) and the M7 (does not allow for positively selected sites) and M8 (allows one extra class of sites with positive selection) models, respectively, to determine statistical probabilities; twice the log-likelihood differences, $-2(\ln \lambda)$ column in Fig. 5C and Supplemental Fig. S5D, and is compared with a $\chi^2$ distribution with 4 degrees of freedom for M0 vs. M3 and 2 degrees of freedom for M7 vs. M8 (Yang et al. 2000).

## Results

### *TALE Homeodomain Proteins in Animals*

Members of previously known TALE classes were used as query sequences using blastp and tblastn (Altschul et al. 1997) to identify the full complement of TALE class homeobox genes from *Caenorhabditis elegans*, *Caenorhabditis briggsae*, *Drosophila melanogaster*, and *Homo sapiens*. This was complemented with sequences from the exhaustively searched genomes of *Drosophila pseudoobscura*, *Apis mellifera*, *Anopheles gambiae*, *Tribolium castaneum*, *Canis familiaris*, *Mus musculus, Rattus norvegicus, Gallus gallus, Xenopus tropicalis*, *Danio rerio*, and the cnidarian *Nematostella vectensis* (Table 1). Additional sequences from other species (Table 2) retrieved by blastp searches from NCBI were also included. Overall, we recovered over 270 TALE protein sequences and their homeodomains were extracted and aligned (Fig. 1A, Supplemental Fig. S1, Supplemental Table S1). In the two well-known genetic model systems *D. melanogaster* and *C. elegans*, we have found eight and five (plus a PREP derived gene; see below) TALE homeobox genes, respectively. In mouse we identified 21 genes plus 1 pseudogene, and in humans, 20 genes plus 4 pseudogenes (Table 1, Fig. 1A). However, the number of pseudogenes in a species has to be qualified, since the information is taken from the annotations. This is still subject to change and we have observed changes over time in some instances. While some genes are obvious pseudogenes, because of frameshifts or stop codons, some genes are difficult to classify either way, since a lack of expression data is not a definitive proof, and conversely, pseudogenes can also be expressed. The number of genes identified in the other species that we searched exhaustively is also shown in Table 1. The gene number is usually well conserved within the protostome (about 7) and tetrapod (about 20) branches. Vertebrates have a higher gene number due to the two rounds of genome duplications that are thought to have occurred during early vertebrate

**Table 1.** Classification of TALE superclass homeobox genes from different phyla

| Class | Family | Protostomes | | | | | | Deuterostomes | | | | | | | CA | Cn. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Ce | Dm | Dp | Ag | Am | Tc | Hs | Cf | Mm | Rn | Gg | Xt | Dr | | Nv |
| PBC | | ceh-20<br>ceh-40<br>ceh-60 | exd | 1 | 1 | 1 | 1 | PBX1, PBX2, PBX3, PBX4, EN16602(P) | 4 | 4 | 4 | 4 | 2 | 6 | 1 | 1 |
| MEIS | MEIS | unc-62 | hth | 1 | 1 | 1 | 1 | MEIS1, MEIS2, MEIS3, EN40731(P), GE00912(P) | 3 | 3 | 3 | 2 | 3 | 5 | 1 | 1 |
| | PREP | psa-3 (no HD) | — | — | 1 | 1 | 1 | PKNOX1, PKNOX2 | 2 | 2 | 2 | 2 | 2 | 3 | 1 | 1 |
| | | | | | | | | | | | | | | | | 1 |
| TGIF | | — | achi, vis | 1 | 1 | 1 | 1 | TGIF2LX, TGIF2LY, TGIF, TGIF2, TGIF1P1(P) | 4+ 1P | 5+ 1P | 3+ 2P | 2 | 2 | 2 | 1 | 1 |
| IRO | | irx-1 | ara, caup, mirr | 3 | 2 | 2 | 2 | IRX1, IRX2, IRX3, IRX4, IRX5, IRX6 | 6 | 6 | 6 | 6 | 6 | 12 | 1 | 1 |
| MKX | | — | CG11617 | 1 | 1 | — | 1 | MKX | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 1? |
| Total | | 5.5 | 8 | 7 | 7 | 6 | 7 | 20 + 4P | 20 + 1P | 21 + 1P | 19 + 2P | 17 | 16 | 30 | 6 | 7 |

*Note*. For the three species *C. elegans* (Ce), *Drosophila melanogaster* (Dm), and *Homo sapiens* (Hs), the gene names are given. For the other species, only the retrieved number of genes is indicated. The top row groups the species according to protostomes, deuterostomes, or Cnidaria (Cn.). The common ancestor (CA) column indicates how many sequences are deduced to have been present at the protostome/deuterostome split. One Nematostella MEIS gene could not be placed into either the MEIS or the PREP family and was put in a separate row. P, putative pseudogenes; *irx-1*, C36F7.1; *psa-3*, F39D8.2. For species abbreviations see Table 2.

evolution. In zebrafish, the TALE gene number is even higher (i.e., 30), due to an additional round of genome duplication in teleost fish. From the cnidarian *Nematostella vectensis* we recovered five TALE homeobox genes. Two additional fragments, for which we have not found any corresponding matches in Stellabase, were recently published (Ryan et al. 2006). One caveat regarding the numbers presented in Table 1 is that they depend on the quality of the individual genome projects. We are certain that the *C. elegans* genome is complete and that the *D. melanogaster* and human genomes are of very high quality. But some of the other genome projects, such as, for example, chicken, still have gaps and areas of uncertainty, where a gene might have been missed.

With the aligned homeodomain sequences we created a metazoan TALE specific protein logo (Fig. 1B, Supplemental Table S2) and compared it to the profile of typical homeodomains (Bürglin 1995). While key residues are conserved in both types of homeodomains, some positions in TALE homeodomains display specific differences. For example, position 19 is primarily occupied by a tyrosine or tryptophan residue, position 20 is frequently a leucine instead of a phenylalanine, and position 23 is predominantly histidine. The loop region between helix 1 and helix 2 (positions 23 to 27; Fig. 1) very frequently contains the sequence "PYP," although in the TGIF class homeodomains the sequence "AYP" is found. In helix 2 at position 31 a highly conserved lysine is present. Helix 3 displays a pattern of conservation similar to typical homeodomains, but position 47 is primarily either an asparagine or a tyrosine, not a small hydrophobic residue, and position 50, which is a key position for DNA binding in the major groove, is a small nonpolar residue, i.e., isoleucine, alanine, or glycine.

### *Phylogenetic Analyses, Revised Classification, and a New Class*

We performed phylogenetic analyses of the homeodomain sequences for classification purposes (Fig. 2). However, classification was not based solely on the phylogenetic tree, but also took into consideration the additional conserved sequence motifs flanking the homeodomain (see below). For classification purposes we use "family" to denote a group of genes that are derived from a single ancestral gene at the split of deuterostomes and protostomes. A class consists of one or several families that share common features, such as conserved motifs flanking the homeodomain, that distinguish them from other classes. Within vertebrates we refer to the duplicated gene families as paralogue groups (see also Bürglin 2005). We have primarily used ML analysis as implemented by Guindon et al. (2003) but obtained

**Table 2.** Species abbreviations used in this paper

| Species code | Species names |
|---|---|
| Aae | *Aedes aegypti* (yellow fever mosquito) |
| Ag | *Anopheles gambiae* (malaria mosquito) |
| Am | *Apis mellifera* (honey bee) |
| Bm | *Brugia malayi* (nematode, agent of lymphatic filariasis) |
| Bmo | *Bombyx mori* (silk worm) |
| Bt | *Bos taurus* (cow) |
| Cb | *Caenorhabditis briggsae* |
| Ce | *Caenorhabditis elegans* |
| Cf | *Canis familiaris* (dog) |
| Ci | *Ciona intestinalis* (tunicate) |
| Cr | *Caenorhabditis remanei* |
| Cs | *Cupiennius salei* (spider) |
| Dm | *Drosophila melanogaster* (fruitfly) |
| Dp | *Drosophila pseudoobscura* |
| Dr | *Danio rerio* (zebrafish) |
| Fr | *Fugu rubripes* (pufferfish) |
| Gg | *Gallus gallus* (chicken) |
| Ggo | *Gorilla gorilla* |
| Hm | *Hydra magnipapillata* |
| Hs | *Homo sapiens* (human) |
| Mf | *Macaca fascicularis* (crab-eating macaque) |
| Mmu | *Macaca mulatta* (rhesus monkey) |
| Mm | *Mus musculus* (mouse) |
| Nv | *Nematostella vectensis* (starlet sea anemone) |
| Ph | *Papio hamadryas* (hamadryas baboon) |
| Pp | *Pongo pygmaeus* (orangutan) |
| Ppa | *Pristionchus pacificus* (nematode) |
| Pt | *Pan troglodytes* (chimp) |
| Rn | *Rattus norvegicus* (rat) |
| Sd | *Suberites domuncula* (sponge) |
| Sj | *Schistosoma japonicum* (flatworm, trematode) |
| Sk | *Saccoglossus kowalevskii* (Hemichordate) |
| Sm | *Schmidtea mediterranea* (flatworm, Planaria) |
| Sp | *Strongylocentrotus purpuratus* (purple sea urchin) |
| Tc | *Tribolium castenium* (red flour beetle) |
| Tn | *Tetraodon nigroviridis* (fish) |
| Vc | *Venturia canescens* (wasp) |
| Xl | *Xenopus laevis* (African clawed frog) |
| Xt | *Xenopus tropicalis* (western clawed frog) |

similar results using NJ (Saitou and Nei 1987). We had found that ML gave more consistent results for highly divergent sequences when analyzing plant homeodomain sequences (K.M. and T.R.B., in preparation).

The previously recognized classes form distinct clades as expected: i.e., PBC, TGIF, MEIS, and IRO (Fig. 2, Supplemental Fig. S2). Within the MEIS class we find two distinct clades comprised of the Meis and Prep genes. Both *D. melanogaster* and *C. elegans* have only one MEIS class gene each, *hth* and *unc-62 (ceh-25)*, respectively, which has led to the implicit supposition that the Prep genes may represent a chordate specific diversification within the MEIS class, because no Prep orthologues from *D. melanogaster* were known at the time (Berthelsen et al. 1998; Bürglin 1998a). Surprisingly, however, our searches found several predicted genes in other insect species (mosquito, honey bee, wasp, and beetle) that share significant similarity with the vertebrate Prep genes (Fig. 2, Supplemental Fig. S2). Thus, the MEIS class can be subdivided into two distinct families, MEIS and PREP, each arising from an ancestral gene that already existed before the divergence of protostomes and deuterostomes. In fact, both families are also present in Nematostella, revealing that the two families are very ancient. In Nematostella a third MEIS class gene (NvHD143) is present that falls into neither family and may be a divergent member of one of these two families, or may possibility represent a family lost in Bilateria (Supplemental Fig. S2B). At present only the homeodomain sequence is known, thus flanking sequences are not available to help resolve this issue.

The TGIF class is also already present in Nematostella. Previous phylogenetic analysis showed that the TGIF class is most similar to the MEIS class (Bürglin 1997). In the present analysis we have found that the TGIF class clusters with the MEIS class with good bootstrap values (78%; Fig. 2). Since the MEIS class proteins share similarity with the plant KNOX class proteins in both the homeodomain and the respective upstream MEIS and KNOX domains, the MEIS class protein structure clearly represents the more ancestral state. Therefore, we conclude that the TGIF genes are derived from a MEIS class gene in early metazoan evolution before the split of Bilateria and Cnidaria.

The IRO class is also of an ancient nature, because single IRO class genes have been found in hydra and Nematostella (Fig. 2, Supplemental Fig. S2). Blast searches in *D. melanogaster* with an IRO protein as query revealed a divergent sequence, CG11617 (~45% identity), that clusters with the IRO sequences. We identified orthologues of CG11617 in many other species, including vertebrates. These sequences form a distinct clade (Fig. 2). However, this clade is not well supported in the ML analysis of the homeodomain sequences (30%, Fig. 2). Phylogenetic trees built using alignments that include conserved flanking sequences provide better support with ML analysis (71%; Fig. 3), and NJ analysis of the latter alignments yields bootstrap values with very high support (99%; Supplemental Fig. S4A). Independently of the phylogenetic analyses, this clade of proteins contains additional conserved motifs that are unique to this group (see below). Therefore, we consider this group of genes its own distinct class. During the revision of this paper an independent study also identified some of these genes and analyzed the mouse gene, named Mohawk (Mkx) (Anderson et al. 2006). Hence, we propose to adopt MOHAWK (MKX) as the name for this class. While the flanking motifs provide clear evidence for the relatedness of the MKX class genes, the protostome and deuterostome branches have significantly diverged, e.g., the homeodomain sequences of mosquito and human

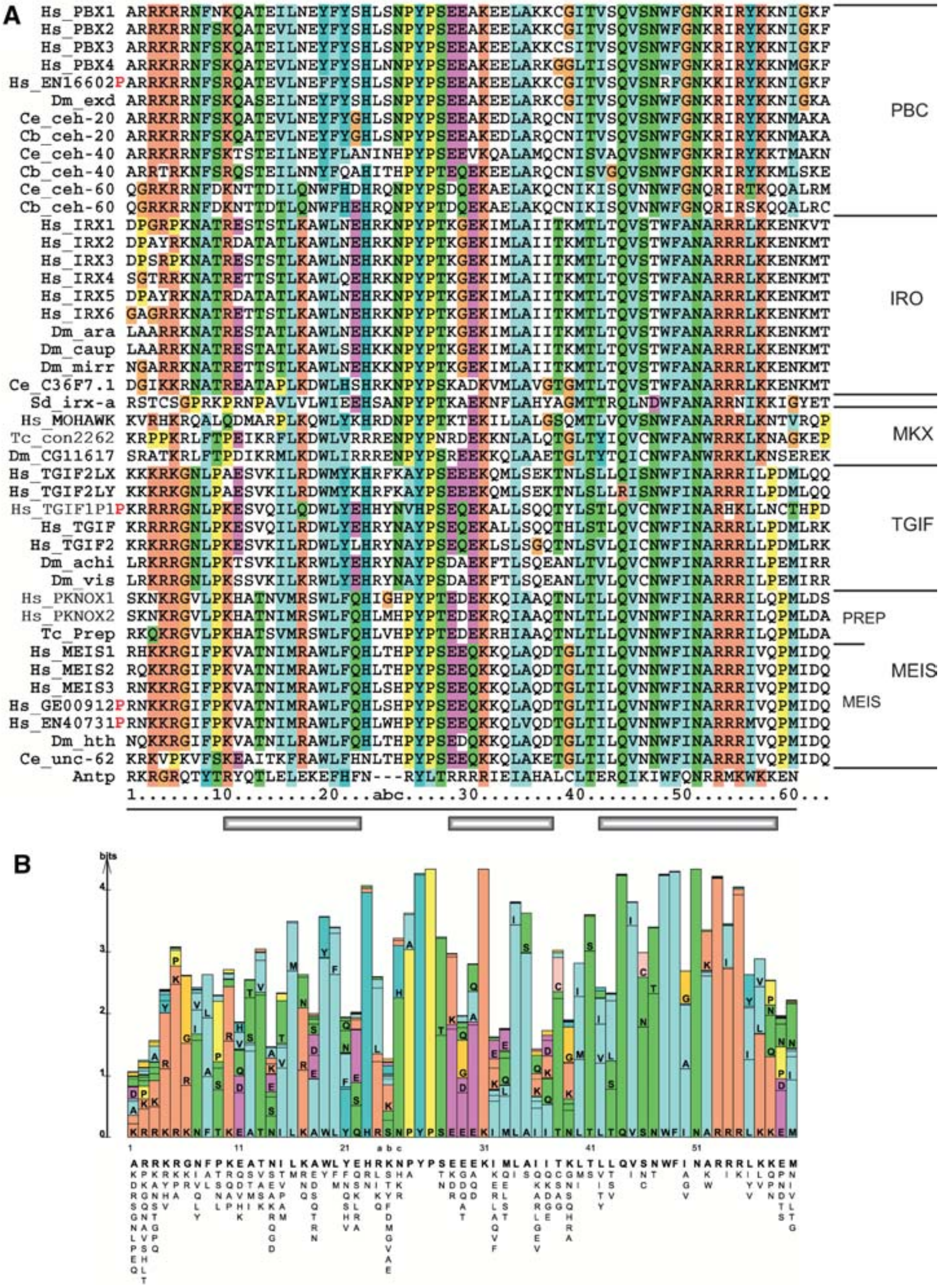


**Fig. 1.** **A** Aligned TALE homeodomain sequences from human, Drosophila, and *C. elegans*, with some additional sequences. At the bottom is the sequence of *Drosophila* Antennapedia (Antp), representing a typical homeodomain. Numbering of the residues is according to Bürglin (1997) and Li et al. (1995); abc marks the three extra residues found in TALE homeodomains. The position of the helixes are marked underneath; classes and families are indicated on the right. Putative pseudogenes are marked with a red P in this and all subsequent figures. For species abbreviations see Table 2. **B** Protein logo generated from 276 aligned TALE homeodomain protein sequences. Under the logo is a consensus sequence under which residues found at a position are listed in decreasing order of frequency. Residues occurring in less than 2% of the sequences have been omitted.

MKX are only 45% identical. While in Nematostella a clear IRO class gene is present, an obvious MKX gene is not present. However, a gene fragment, Nv_TALE-L (Chourrout et al. 2006; Ryan et al. 2006), that groups with MKX genes in phylogenetic trees has been found, although the bootstrap values are low (Supplemental Fig. S2B). A gene from sponges, called IRX-a, branches off before the IRO/MKX split, although not with significant bootstrap values (Fig. 2). Its homeodomain sequence is 34% identical to mosquito MKX, and 42% identical to mosquito IRO, but does not share particular residues with either IRO or MKX. The flanking sequences of IRX-a show no motifs that are conserved with either IRO or MKX, although this could be a secondary loss. The presence of an IRO/MKX-like gene in sponges that cannot be assigned to either class could indicate that IRO and MKX diverged after sponges separated from the branch leading to eumetazoans. The IRO/MKX clade of sequences may be more closely related to the MEIS/PREP/TGIF clade than to the PBC class (Fig. 2).

## *Associated Domains and Motifs*

Homeodomain proteins often contain associated conserved motifs besides the homeodomain that are distinct for the different classes. These motifs provide

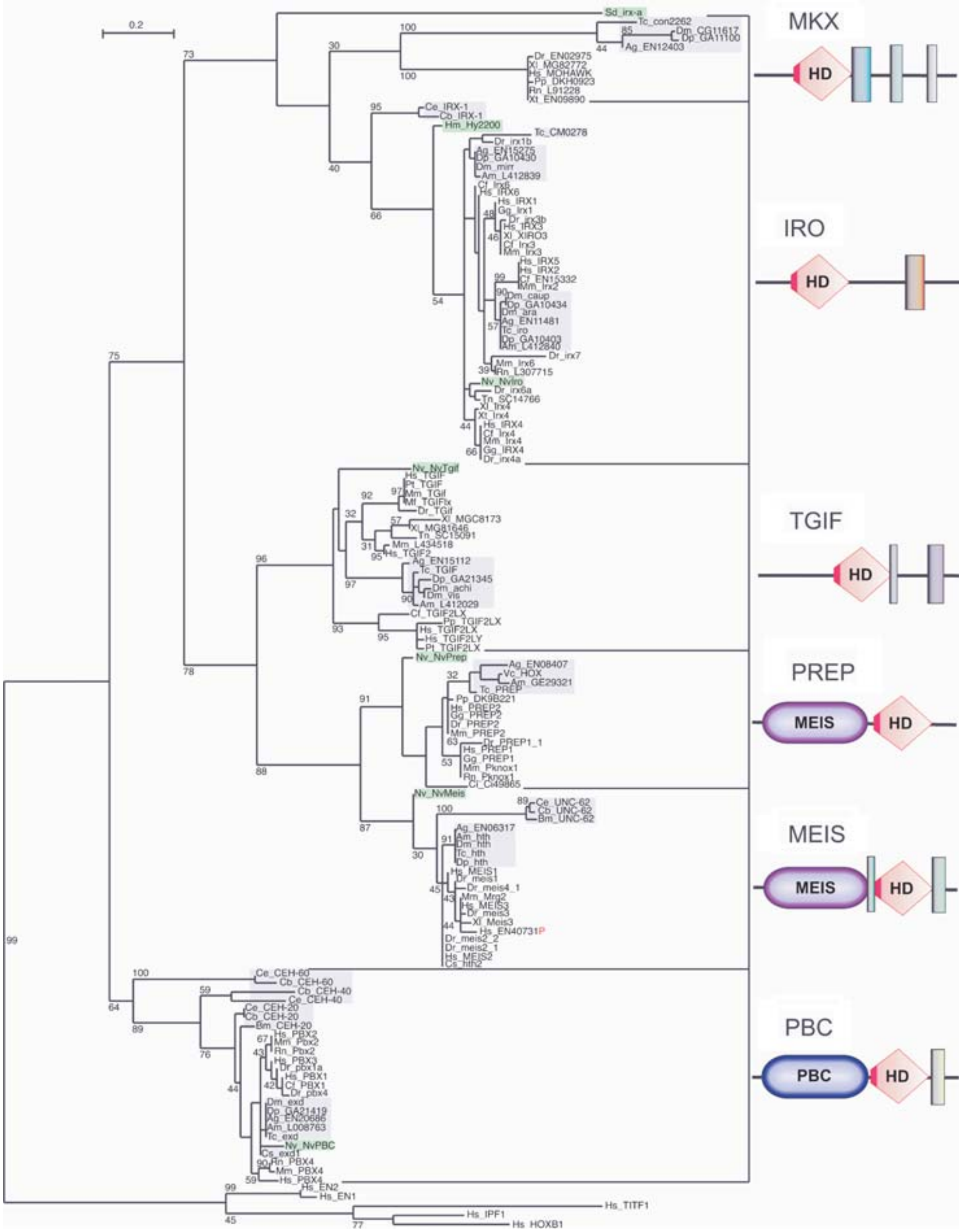


**Fig. 2.** ML tree of selected TALE homeodomain sequences. The tree was built using the PHYML tree building program with the JTT substitution model. One hundred bootstrap trials were run to obtain relative support of the branches. Bootstrap values are indicated and values lower than 30 are removed from the tree for clarity. The typical homeodomains of TITF1, IPF1, and HOXB1 were used as outgroup. In this and subsequent figures members of the protostome group are highlighted in light blue; sponge and cnidarian members, in light green. The domain architecture for the proteins of the different TALE classes is shown on the right. A red "P" marks putative pseudogenes. Species abbreviations in Table 2.

additional independent support for the classification derived from the phylogenetic analyses of the homeodomain sequences. We generated multiple sequence alignments of the full-length protein sequences for the different classes to identify and characterize domains and motifs conserved between protostomes and deuterostomes (Supplemental Fig. S3).

**IRO class**. Immediately downstream of the homeodomain is a well-conserved region of about 13 residues that we refer to as IRO A, which is followed by an acidic stretch (Supplemental Fig. S3A). The IRO box motif of about 15 residues (Supplemental Fig. S3A) has been described previously (Bürglin 1997).

**MKX class**. We identified three conserved motifs downstream of the homeodomain, named MKX A, B, and C (Fig. 4B, Supplemental Fig. S3B). MKX A is about 30 amino acids in length and is a C-terminal extension of the homeodomain. MKX B and MKX C are located farther downstream and are 15 and 12 residues in length, respectively. Possibly these motifs, in particular, MKX B, which is rich in hydrophobic residues, might be important for protein-protein interaction.

**MEIS class**. The MEIS class is comprised of two families, MEIS and PREP, both of which have a bipartite MEIS domain (A and B), which has been

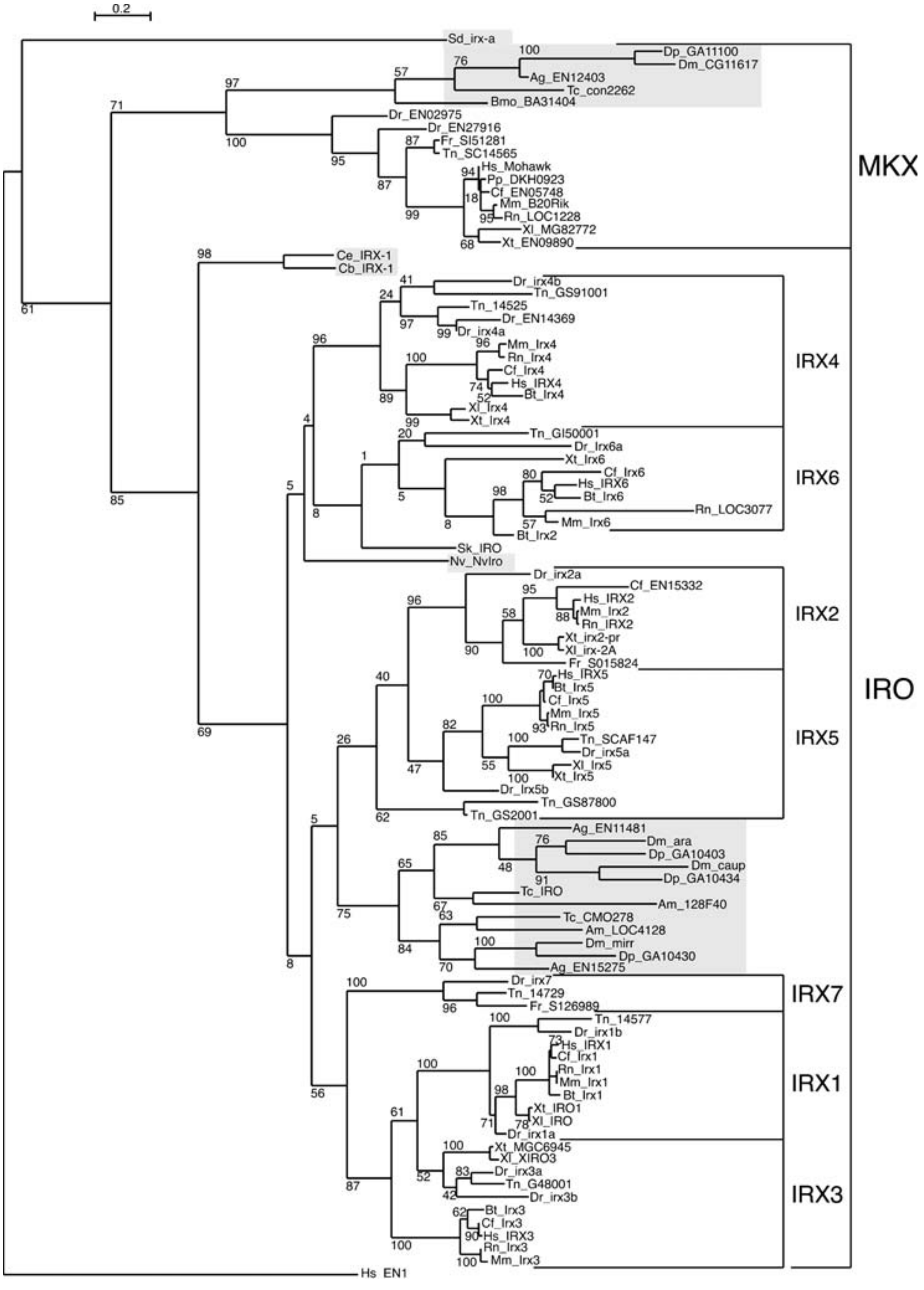


**Fig. 3.** ML tree of the MKX and IRO class proteins using the homeodomain as well as flanking conserved regions. PHYML was used with 100 bootstrap trials, and the tree was built using the JTT substitution model. The vertebrate specific paralogue groups of the IRO class are labeled with their respective names. Species abbreviations in Table 2.

described previously (Bürglin 1997). There are some distinctions in the MEIS domain between the two families, primarily a shorter MEIS A domain in the PREP family (Supplemental Fig. S3C). A few residues downstream of the homeodomain are also conserved between the MEIS and PREP families (Supplemental Figs. S3D and S3F). By aligning the two families separately, we identified new motifs immediately upstream and downstream of the homeodomain that are found only in the MEIS family, termed MEIS C and D (Supplemental Fig. S3D). These motifs are about 20 residues long. Inspection of the N-terminus of the MEIS family proteins revealed conservation between some of the protostome and deuterostome sequences. We removed sequences from the MEIS alignment that did not match well in this region and were able to define a region of about 30 residues at the N-terminus that shows conservation among Tribolium, flies, and vertebrates, which we name MEIS N (Supplemental Fig. S3E). In Caenorhabditis UNC-62 this region is not conserved. Blast searches with the MEIS domain revealed two genes encoding MEIS domains in *C. elegans*: one is the MEIS family gene *unc-62*, and

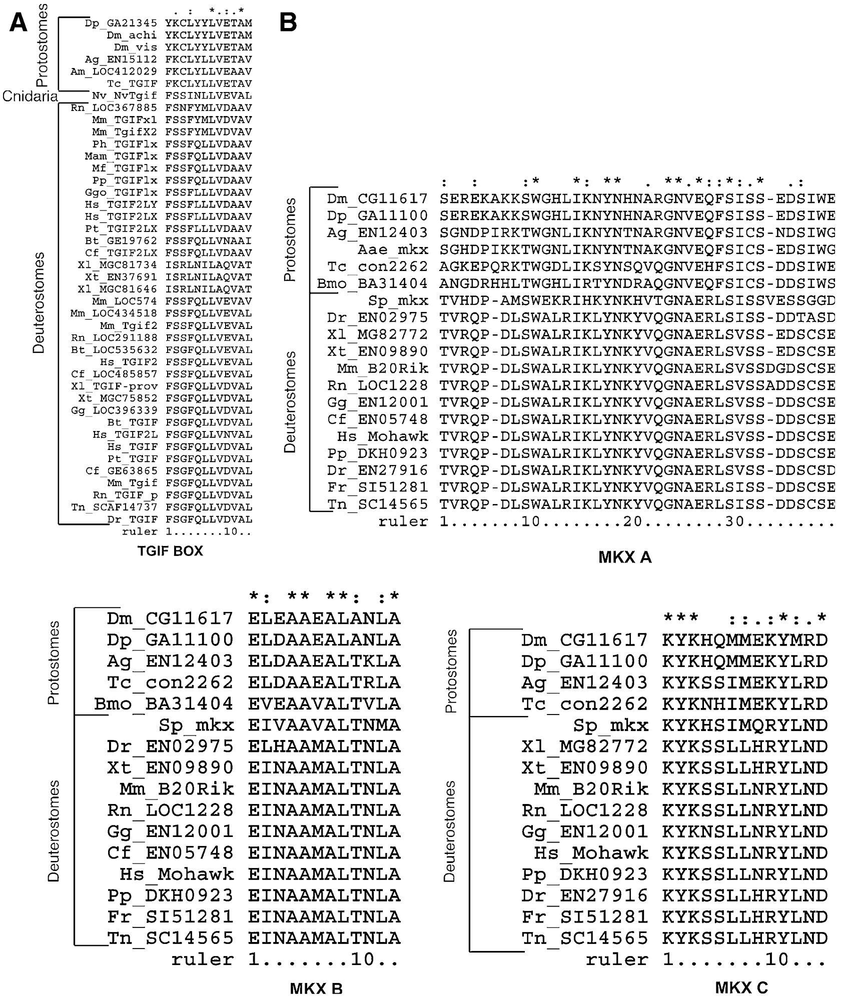
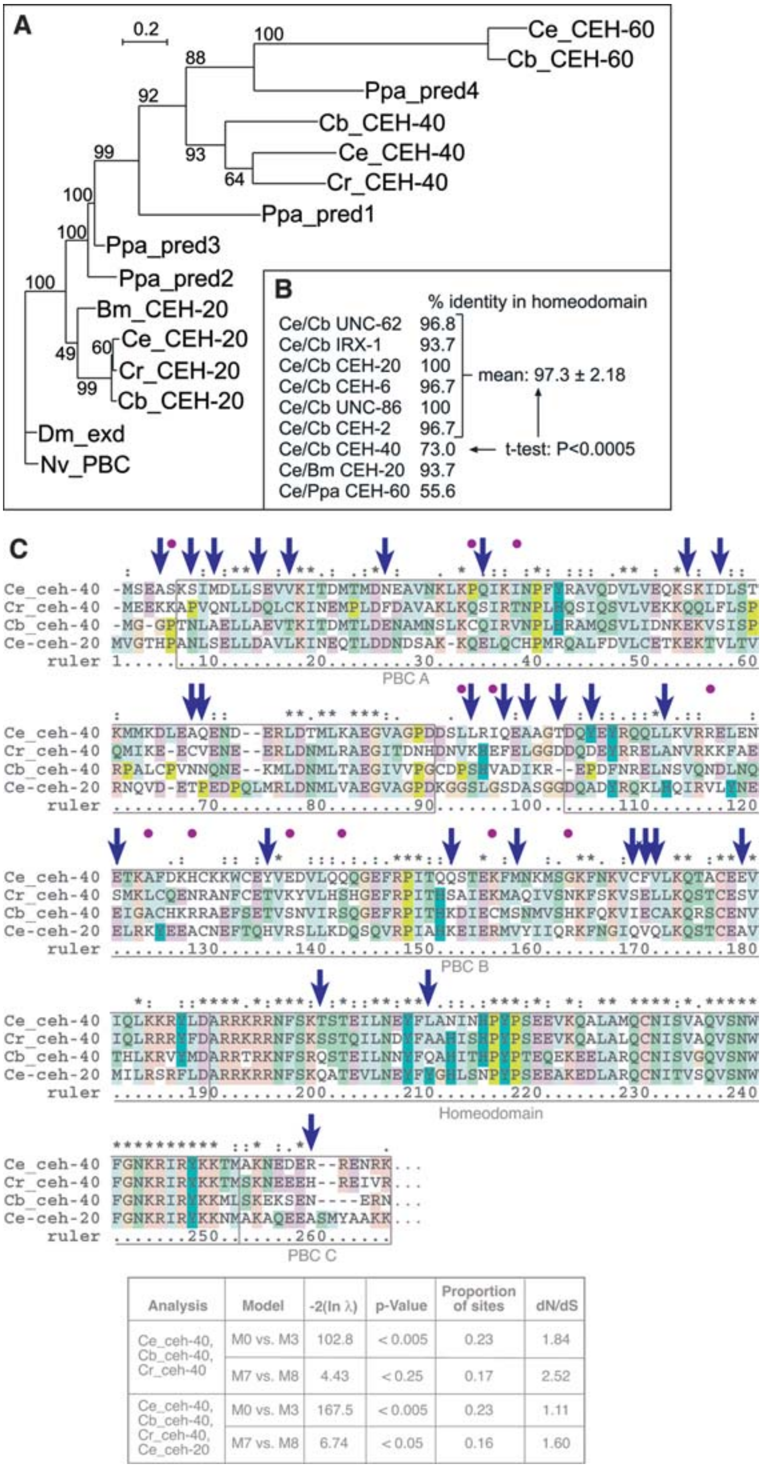


**Fig. 4.** New conserved motifs detected in MKX and TGIF class homeodomain proteins. Species abbreviations in Table 2. **A** The short TGIF Box found in the C-terminal region of TGIF homeodomain proteins. **B** The three short conserved motifs found in MKX homeodomain proteins, named MKX A, MKX B, and MKX C.

the other is *psa-3* (F39D8.2), which encodes only a MEIS domain. *psa-3* is rather divergent, but the size of the MEIS domain, in particular, the shorter MEIS A domain, indicates that *psa-3* is a derived member of the PREP family that must have lost the homeodomain (Supplemental Figs. S3C and S3F). We would like to note that *unc-62* (*ceh-25*) undergoes alternative splicing and can encode two different homeodomain forms that are distinct in the first 15 residues of the homeodomain (Bürglin 1997; Van Auken et al. 2002), although we present only the more conserved splice form in this study. This extra complexity of UNC-62 may perhaps compensate for the loss of the homeodomain in *psa-3*. The PREP family proteins do not contain any well-conserved motifs specific to PREP. However, downstream of the homeodomain is a region where Tribolium shares a number of residues with the vertebrate homologues (Supplemental Fig. S3F) that may be the last remnant of a specific motif.

**TGIF class.** A purported sequence conservation has been reported in the C-terminus of TGIF proteins among Drosophila, mouse, and human based on Clustal W analysis of few sequences (Wang and Mann 2003). However, using MUSCLE and more arthropod sequences, we have obtained a different multiple sequence alignment with more conserved residues in a block (Supplemental Fig. S3G). This new motif of about 12 residues is located in the C-terminus of the TGIF proteins, termed the TGIF box (Fig. 4A). Because of the overrepresentation of hydrophobic residues in the TGIF box, we suspect that this motif may play a role in protein-protein interaction. Another short motif, TGIF A, which is 17 residues long, is present immediately downstream of the homeodomain. TGIF A is conserved between arthropods and the vertebrate TGIF and TGIF2 proteins but has diverged in the TGIFLX proteins (Supplemental Fig. S3G).

**PBC class.** The PBC class genes tend to be highly conserved, with a large, bipartite PBC domain upstream of the homeodomain and a short region of about 13 residues downstream of the homeodomain that we refer to as PBC C here (Bürglin and Ruvkun 1992) (Supplemental Fig. S3H). In Caenorhabditis though, we find three PBC class genes, two of which,


A
0.2
100
88
92
99
93
64
100
100
100
49
60
99
Ce_CEH-60
Cb_CEH-60
Ppa_pred4
Cb_CEH-40
Ce_CEH-40
Cr_CEH-40
Ppa_pred1
Ppa_pred3
Ppa_pred2
Bm_CEH-20
Ce_CEH-20
Cr_CEH-20
Cb_CEH-20
Dm_exd
Nv_PBC
B
mean: 97.3 ± 2.18
t-test: P<0.0005
C
Ce_ceh-40 -MSEASKSIMDLLSEVVKITDMTMDNEAVNKLKPQIKINPFYRAVQDVLVEQKSKIDLST
Cr_ceh-40 -MEEKKAPVQNLLDQLCKINEMPLDFDAVAKLKQSIRTNPLHQSIQSVLVEKKQQLFLSP
Cb_ceh-40 -MG-GPTNLAELLAEVTKITDMTLDENAMNSLKCQIRVNPLHRAMQSVLIDNKEKVSISP
Ce-ceh-20 MVGTHPANLSELLDAVLKINEQTLDDNDSAK-KQELQCHPMRQALFDVLCETKEKTVLTV
ruler 1.......10.........20.........30.........40.........50.........60
PBC A
Ce_ceh-40 KMMKDLEAQEND--ERLDTMLKAEGVAGPDDSLLRIQEAAGTDQYEYRQQLLKVRRELEN
Cr_ceh-40 QMIKE-ECVENE--ERLDNMLRAEGITDNHDNVKHEFELGGDDQDEYRRELANVRKKFAE
Cb_ceh-40 RPALCPVNNQNE--KMLDNMLTAEGIVVPGCDPSHVADIKR--EPDFNRELNSVQNDLNQ
Ce-ceh-20 RNQVD-ETPEDPQLMRLDNMLVAEGVAGPDKGGSLGSDASGGDQADYRQKLHQIRVLYNE
ruler .........70.........80.........90.........100.........110.........120
Ce_ceh-40 ETKAFDKHCKKWCEYVEDVLQQQGEFRPITQQSTEKFMNKMSGKFNKVCFVLKQTACEEV
Cr_ceh-40 SMKLCQENRANFCETVKYVLHSHGEFRPITHSAIEKMAQIVSNKFSKVSELLKQSTCESV
Cb_ceh-40 EIGACHKRRAEFSETVSNVIRSQGEFRPITHKDIECMSNMVSHKFQKVIECAKQRSCENV
Ce-ceh-20 ELRKYEEACNEFTQHVRSLLKDQSQVRPIAHKEIERMVYIIQRKFNGIQVQLKQSTCEAV
ruler .........130.........140.........150.........160.........170.........180
PBC B
Ce_ceh-40 IQLKKRYLDARRKRRNFSKTSTEILNEYFLANINHPYPSEEVKQALAMQCNISVAQVSNW
Cr_ceh-40 IQLRRRYFDARRKRRNFSKSSTQILNDYFAAHISHPYPSEEVKQALALQCNISVAQVSNW
Cb_ceh-40 THLKRVYMDARRTRKNFSRQSTEILNNYFQAHITHPYPTEQEKEELARQCNISVGQVSNW
Ce-ceh-20 MILRSRFLDARRKRRNFSKQATEVLNEYFYGHLSNPYPSEEAKEDLARQCNITVSQVSNW
ruler .........190.........200.........210.........220.........230.........240
Homeodomain
Ce_ceh-40 FGNKRIRYKKTMAKNEDER--RENRK...
Cr_ceh-40 FGNKRIRYKKTMSKNEEEH--REIVR...
Cb_ceh-40 FGNKRIRYKKMLSKEKSEN----ERN...
Ce-ceh-20 FGNKRIRYKKNMAKAQEEASMYAAKK...
ruler .........250.........260.......
PBC C

| | % identity in homeodomain |
|---|---|
| Ce/Cb UNC-62 | 96.8 |
| Ce/Cb IRX-1 | 93.7 |
| Ce/Cb CEH-20 | 100 |
| Ce/Cb CEH-6 | 96.7 |
| Ce/Cb UNC-86 | 100 |
| Ce/Cb CEH-2 | 96.7 |
| Ce/Cb CEH-40 | 73.0 |
| Ce/Bm CEH-20 | 93.7 |
| Ce/Ppa CEH-60 | 55.6 |

| Analysis | Model | -2(ln λ) | p-Value | Proportion of sites | dN/dS |
|---|---|---|---|---|---|
| Ce_ceh-40, Cb_ceh-40, Cr_ceh-40 | M0 vs. M3 | 102.8 | < 0.005 | 0.23 | 1.84 |
| | M7 vs. M8 | 4.43 | < 0.25 | 0.17 | 2.52 |
| Ce_ceh-40, Cb_ceh-40, Cr_ceh-40, Ce_ceh-20 | M0 vs. M3 | 167.5 | < 0.005 | 0.23 | 1.11 |
| | M7 vs. M8 | 6.74 | < 0.05 | 0.16 | 1.60 |

**Fig. 5.** Accelerated evolution in nematode PBC proteins. **A** ML phylogenetic tree of nematode PBC proteins. One hundred bootstrap trials are shown. Nv_PBC is used as outgroup. **B** Percentage identities between selected orthologous homeodomains of different nematode species. The mean and standard deviation of the percentage identities of six homeodomains conserved between *C. elegans* and *C. briggsae* were taken and compared with those of *ceh-40* and statistically evaluated using a one-tailed *t*-test. **C** Multiple sequence alignment of Caenorhabditis CEH-40 and *C. elegans* CEH-20 proteins. *ceh-40* sequences were analyzed either with or without including *ceh-20* using codeml. Models used for log ratio tests are indicated and the twofold log ratio difference between models and resulting *p*-values are shown. The proportion of sites under positive selection is also indicated. Sites that were identified with both models on both data sets are shown with blue arrows on the multiple sequence alignment, and sites that were identified with the M3 model on both data sets plus one M8 model are shown with pink dots.

*ceh-40* and *ceh-60*, are highly divergent. Very little sequence conservation is found in the PBC domain of CEH-60.

### Evolution Within Classes

To obtain a finer resolution of the evolution of genes within each class and to trace their evolution in different taxa, class-specific phylogenetic analyses were carried out. The full-length protein sequence alignments that we had generated for each class were trimmed to the conserved regions and then subjected to ML as well as NJ analysis (Fig 3, Supplemental Fig. S4; data not shown). The phylogenetic trees of individual classes reveal distinct vertebrate specific subfamilies within each class that we refer to as paralogue groups (PGs). Usually, each PG is represented by a single gene, although cases of gene expansion in particular taxa and PGs are seen. In tetrapods there are four PGs in the PBC class, three

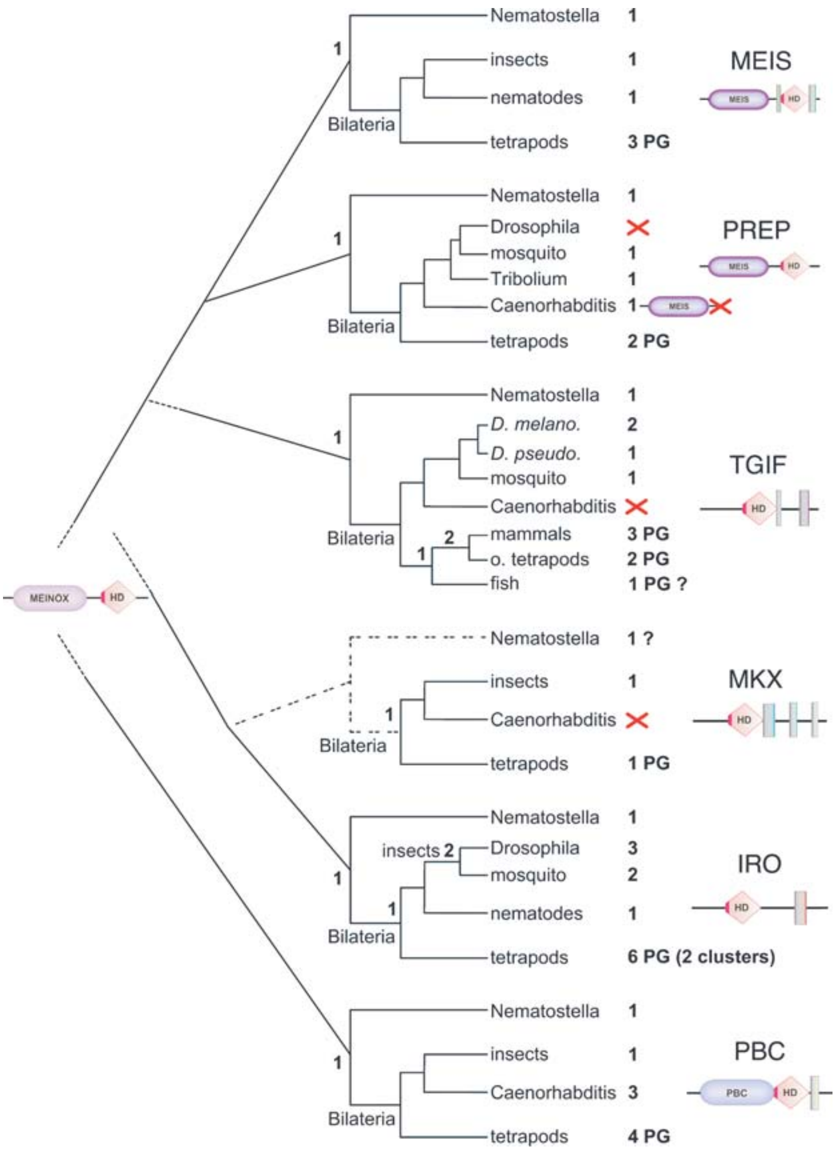


**Fig. 6.** Summary of the evolution of TALE homeodomain proteins in animals. The putative evolution from an ancestral MEINOX-TALE homeodomain protein into the six present data families. Numbers indicate the number of homeobox genes at a particular time point in evolution; present-day numbers are on the right. A red X indicates a loss. For further discussion see text.

in the MEIS family, two in the PREP family, two in the TGIF class (plus one mammalian specific one), and six in the IRO class but only one in the MKX class (Table 1). In protostomes we generally find only a single gene for each family and class.

**IRO class**. The six tetrapod IRO class genes are located on two paralogous clusters comprised of the genes IRX1/IRX2/IRX4 and IRX3/IRX5/IRX6. The original three genes in the cluster formed by tandem duplication, while a subsequent duplication gave rise to two paralogous clusters (Dildrop and Rüther 2004; Feijóo et al. 2004). Our phylogenetic analyses support the latter duplication event, since the paralogous genes IRX1 and IRX3, IRX2 and IRX5, and IRX4 and IRX6, respectively, each form separate clades (Fig. 3, Supplemental Fig. S4A). The cluster duplication is most likely the result of the same genome duplication during early vertebrate evolution that also gave rise to the paralogous Hox clusters, although only two IRX clusters survived. In zebrafish a fish-specific group of genes, IRX7, could not be reliably placed (Feijóo et al. 2004). Our data indicate that IRX7 is derived from IRX1 or IRX3. The IRX7 PG may thus represent the only remaining member of an additional IRX cluster that formed during the extra round of genome duplication thought to have occurred in telost fish. In *D. melanogaster* we find three IRO genes. The gene duplication of the IRO complex in arthropods occurred in two steps. The split of *mirror* from an ancestral *ara*/*caup* must have happened before Coleoptera and Diptera separated, since we find two genes in species such as Tribolium and mosquito (Fig. 3, Supplemental Fig. S4A). The duplication into *ara* and *caup* is more recent but happened before the diversification of drosophilids, because only a single gene is found in mosquito, while all 12

*Drosophila* species present today in flybase have both (data not shown). In *C. elegans* only a single IRO gene is present, *irx-1* (C36F7.1).

**MKX class**. In all species where we found MKX genes, only a single copy is present, apart from zebrafish, which has two divergent copies (Fig. 3). In fugu and Tetraodon we recovered only a single gene. In Caenorhabditis this gene has been lost.

**MEIS class**. The genes in the MEIS family are well conserved, with a single copy in protostomes (*hth*) and three PGs in tetrapods (Supplemental Fig. S4B). In spider, a taxon specific duplication event gave rise to two paralogous MEIS genes (Supplemental Fig. S4B). By contrast, the PREP family was subject to evolutionary changes in protostomes. It was lost at some point during early evolution of drosophilids but is still present in mosquito (Table 1, Fig. 2). In the nematode branch, only the homeodomain was lost, and the remaining MEIS domain was subject to substantial sequence divergence. While the bootstrap values that place Caenorhabditis PSA-3 within the PREP family are not significant (30%; Supplemental Fig. S4B), this grouping is supported by the fact that PSA-3 has the same short linker region between MEIS A and MEIS B as other PREP proteins.

**TGIF class**. Two of the three TGIF PGs are not represented throughout the entire vertebrate lineage. We identified members of the TGIFLX PG only in mammals, suggesting that it is a mammalian specific adaptation that arose through a duplication in early mammals. Members of the TGIF2 PG were identified in chicken, *Xenopus laevis*, and *Xenopus tropicalis*, but we failed to find any in zebrafish and Tetraodon (Table 1, Supplemental Fig. S4C). Thus, TGIF2 may have been lost in fish or may have arisen only in tetrapods. These data indicate that TGIF genes are evolving more rapidly, which has already been noticed for the TGIFLX genes (Wang and Zhang 2004). Further support for this is provided by the number of putative pseudogenes in mammals (Table 1). There are also duplications within taxa, for example, in mouse we find that there are two recently duplicated genes residing in tandem in the TGIFLX PG, and there are three TGIF2 genes, one of which appears to be a pseudogene, because it lacks a residue within the homeodomain and no corresponding ESTs have been identified (Supplemental Fig. S4C). *D. melanogaster* has two TGIF genes, *vis* and *achi*. This is a recent duplication event, since we can find two genes in *D. sechellia* and *D. virilis* (data not shown), but in the more distantly related *D. pseudoobscura* only a single gene is present and the phylogenetic analysis places it as an outgroup to the vis/achi split (Supplemental Fig. S4C). In Caenorhabditis the TGIF gene has been lost.

**PBC class**. The PBC class has four paralogue groups in tetrapods and a single copy in insects. However, in Caenorhabditis we find three PBC genes (Fig. 2, Supplemental Fig. S4D). The CEH-40 and CEH-60 proteins are highly divergent, and both the homeodomain and the PBC domain have numerous substitutions, even when orthologous pairs are compared between *C. briggsae* and *C. elegans*. The long branch lengths indicate that these two genes are fast-evolving and diverging rapidly. By contrast, the Caenorhabditis protein CEH-20 is much better conserved compared to the other PBC proteins. Further evidence from other nematode genes (see below) supports the notion that *ceh-40* and *ceh-60* are nematode specific diversifications. Four PBC PGs are present in vertebrates. ML did not resolve the fish PBX2 and PBX4 PGs into the respective tetrapod PBX2 and PBX4 PGs (Supplemental Fig. S4D), but NJ placed them within PBX2 and PBX4, respectively. In particular, zebrafish pbxy and its Tetraodon orthologue can be identified as members of the PBX2 PG (Supplemental Fig. S4E).

### *Some TALE Superclass Genes Show Elevated Rates of Evolution and Nonsynonymous Substitutions*

The phylogenetic analyses of the individual classes revealed sequences that have longer branch lengths and appear to undergo faster evolution, for example, the *C. elegans* PBC class genes *ceh-40* and *ceh-60* and the mammalian PBX4 PG genes (Fig. 2, Supplemental Fig. S4D, E). Rapid evolution and positive Darwinian selection have already been reported for the TGIF genes in the TGIFLX PG (Wang and Zhang 2004). Therefore, we examined all TALE classes for cases of accelerated evolution and positive selection.

To investigate the evolution of the PBC genes in nematodes, we retrieved additional sequences from the ongoing sequencing projects of *C. remanei*, *B. malayi*, and *P. pacificus*. *C. elegans*, *C. briggsae, and C. remanei* are closely related species, while the parasitic nematode *B. malayi* is separated from Caenorhabditis by about 350 MY and *P. pacificus* by about 200–300 MY (De Ley 2006). *B. malayi* has a *ceh-20* gene whose homeodomain is 93.7% identical to *C. elegans ceh-20*, which suggests a strong evolutionary constraint, but appears to lack both *ceh-40* and *ceh-60*, although the genome sequence is not yet complete (Figs. 5A and B, Supplemental Fig. S5A). In *P. pacificus* we recovered four PBC gene fragments; two, Ppa_pred1 and Ppa_pred2, are located within about 1 kb of each other. Ppa_pred2 is most likely the orthologue of Caenorhabditis *ceh-20*, while Ppa_pred1 is divergent (Fig. 5A, Supplemental Fig. S5B). Ppa_pred4 is placed in a clade with *ceh-60*, but the homeodomain is only 55.6% identical to that of *C. elegans ceh-60* (Fig. 5B), indicating either that this gene is an independent duplication in *P. pacificus* and not a homologue of *ceh-60* or that the *ceh-60* genes have substantially diverged. A fourth gene fragment, Ppa_pred3, is highly similar to Ppa_pred2 even at the nucleotide level (Supplemental Fig. S5C). Eight nonsynonymous and four synonymous substitutions differentiate Ppa_pred3 and Ppa_pred2, but presently the contig on which Ppa_pred3 resides is comprised of only two sequence reads. Therefore, Ppa_pred3 may be a recently duplicated gene or may be merged with Ppa_pred2 in the future. *C. elegans* and *C. briggsae* CEH-60 are not that divergent from each other, but they have substantially diverged from the other PBC genes, in particular, in the PBC domain (Supplemental Fig. S5A).

The Caenorhabditis CEH-40 proteins are unusually divergent for such closely related species. Normally, the identity of a homeodomain between *C. elegans* and *C. briggsae* ranges between 90% and 100%, but in the case of CEH-40 it is only 73%. A *t*-test comparison of the later value with an average of six homeodomains (97.3% ± 2.18%) shows that this difference is significant ($P < 0.0005$) (Fig. 5B). To estimate whether the cause of divergence is a result of relaxed constraint or due to positive selection, we tested the Caenorhabditis *ceh-40* genes for positive selection using codeml in PAML (Yang et al. 2000). We compared the *ceh-40* sequences either among themselves or together with *C. elegans ceh-20*. The M3 and M8 codon models were compared to M0 and M7, respectively (Fig. 5C). Sites that were identified as being under positive selection are indicated in the multiple sequence alignment (Fig. 5C). About 20% of the sites are estimated to be under positive selection, many of them in the PBC domain at its N-terminal end and the linker region between PBC A and PBC B. In the homeodomain two sites were identified, both in helix 1 of the homeodomain, that are on the outside surface and can interact with other molecules (Fig. 5C). The PBC domain is important for protein-protein interactions (e.g., Mann and Affolter 1998), thus changes in that domain will also lead to adaptations and modifications of interactions. The mammalian PBX4 genes are also diverging at a rate faster than the other PBX PGs (Supplemental Fig. S4E). Analysis with codeml found two sites with positive Darwinian selection between mouse and rat Pbx4, which are also detected when human, mouse, and rat Pbx4 are compared. One of the sites is located in the conserved PBC A domain (Supplemental Fig. S5D). However, to analyze this further, additional data from other rodent taxa would be helpful. We have investigated positive selection also in all other classes but did not find any other obvious instances with the sequences used in this study.

## Discussion

We have retrieved the full complement of TALE homeobox genes from *C. elegans*, Drosophila, and humans, and we have performed exhaustive genome searches of numerous other species (Table 1). Our study has two limitations. First, some of the genome projects are not yet fully completed and thus the number of TALE homeobox genes in some of the species is still subject to change. Second, even when one can be confident to have the complete genome sequence, such as with humans, it is not trivial to distinguish between "real" genes and pseudogenes. While some genes are clearly pseudogenes, because of frame shifts, others seem to have an intact ORF and may potentially not be pseudogenes. Despite these caveats, we think that our exhaustive analysis of animal homeobox genes provides a definitive framework for the classification of these genes in metazoans. In addition to the known classes, we identify here a new class of TALE homeobox genes, MKX, as well as a new conserved MEIS family, i.e., PREP. Given that we sampled many species, including a Cnidaria, it is unlikely that we missed a major TALE family, even if one or the other TALE homeobox gene has been missed, because of incomplete genome sequences. There may of course be phylum specific diversifications and adaptations within particular families. Hints of such diversifications stem from the *C. elegans* protein PSA-3, which lost its homeodomain, or CEH-60, whose PBC domain is very degenerate.

In this paper we have unearthed several new conserved sequence motifs. We have concentrated on identifying those motifs that have been conserved across the protostome-deuterostome chasm. Within the protein sequences of particular vertebrate PGs it is not uncommon to see high sequence conservation throughout the length of the proteins, since the sequences are more closely related. Nevertheless, our multiple sequence alignments can be used to identify specific regions that are conserved within vertebrate PGs (Supplemental Fig. S3).

### *Evolution of TALE Homeobox Genes*

Based on our analysis the following scenario for the evolution of the animal TALE homeobox genes can be proposed (Fig. 6). Though both insects and vertebrates have clusters of IRO class genes, only a single gene was present at the time of the protostome and deuterostome IRO split. This is supported by the fact that in *C. elegans*, cnidarians, and the hemichordate *Saccoglossus kowalevskii*, only a single IRO gene has been found. In insects, tandem duplication events gave rise to a cluster of IRO genes. In Apis, Anopheles, and Tribolium two genes are present, while in Drosophila an extra duplication resulted in three genes. Functional redundancy can still be found between the Drosophila genes, for example, *ara* and *caup* have similar expression patterns, function in the wing, and can functionally replace each other (Gómez-Skarmeta et al. 1996). In the vertebrate lineage several duplication events gave rise to a cluster of three tandem duplicated genes that, in an additional duplication event, gave rise to two paralogous clusters with the genes Irx1/Irx2/Irx4 and Irx3/Irx5/Irx6. The large number of IRX genes in vertebrates suggests that numerous adaptations and modifications could have been possible. While there are sequence motifs specific for particular PGs, we have found little evidence for faster evolution.

The MKX class of genes is most similar to the IRO class. Yet this class is clearly distinguished from IRO by the different domain structures outside of the homeodomain. Most interestingly, an IRO-like sequence (irx-a) was discovered in sponges that lacks IRO and MKX motifs. Our phylogenetic analyses place this gene outside the IRO/MKX clade. This suggests that only a single gene may have existed in early metazoans, and the IRO/MKX duplication and divergence may have occurred in early eumetazoan evolution. At present, the fragmentary MKX-like homeobox from Nematostella does not allow us to be 100% confident that the MKX class exists in cnidarians. Although we favor this hypothesis, it is still possible that the MKX class arose by duplication from an IRO gene after the split of Cnidaria and Bilateria, and that the Nematostella TALE-L homeobox arose independently through duplication and divergence in Cnidaria. Even vertebrate species that usually tend to have several paralogous genes have only a single MKX class gene with the exception of fish. This suggests that there was selective pressure against having multiple copies of this gene in vertebrates. We also note that the protostome and deuterostome MKX proteins are more divergent from each other than is usual in other TALE classes. It is formally possible that there were two paralogous genes present at the protostome/deuterostome split, and one was lost in the protostome branch and one in the deuterostome branch. But at present we favor the hypothesis that there was just a single gene, since the protostome branch is evolving quite fast, which would explain the sequence divergence, and *C. elegans* appears to have lost this gene completely. Little is known about the function of the MKX genes. In mouse, Mkx is expressed in somite-derived cell lineages that, among other cell types, give rise to skeletal muscle (Anderson et al. 2006). Similarly the *D. melanogaster* orthologue is expressed in muscle during larval development, as revealed in a large-scale in situ hybridization screen (Tomancak et al. 2002). This suggests that the MKX genes play a role in the mesoderm.

The MEIS class can be split into two ancient families, MEIS and PREP, that were already present before Cnidaria and Bilateria separated, which is rather surprising, given the similarity between the two families (70% identity between human PREP1 and MEIS1 in the homeodomain). While Drosophila has lost the PREP gene, other insects have retained it (Fig. 6). In Caenorhabditis, yet a different adaptation occurred. There the homeodomain was lost, and only the MEIS domain was retained. Only a single PREP gene is present in Ciona, while in vertebrates this has expanded into two well-conserved PGs. Thus, multiple independent events have tinkered with the PREP family during evolution. We have also identified structural differences between the MEIS and the PREP families. The MEIS family has regions of conservation upstream and downstream of the homeodomain, MEIS C and MEIS D, as well as a lesser-conserved region in the N-terminus, MEIS N. These regions have no counterparts in the PREP proteins. We expect that these two motifs will contribute to the functional divergence between the two families.

The TGIF class is most likely derived from an ancestral MEIS class gene. In the process of diversification, the MEIS domain must have been lost, and additional motifs, such as the TGIF, were gained. In vertebrates three TGIF PGs can be defined, but TGIF2 seems to be restricted to tetrapods, while TGIFLX is restricted to mammals. Wang and Zhang (2004) have recently analyzed the X-linked TGIFLX gene from 16 primate groups and found that this gene is subject to rapid evolution and is expressed in the testis. The fly TGIF genes *achi* and *vis*, although not located on the X chromosome, have been shown to be involved in spermatogenesis (Wang and Mann 2003). Other fast-evolving homeobox genes that are X-linked have been found in Drosophila, e.g., *OdsH* (Ting et al. 1998), and in mouse, for example, Rhox5 (Maclean et al. 2005), both of which are expressed in the reproductive system. Thus, there seems to be some correlation among expression in the reproductive system, being X-linked, and fast evolution.

In this study we have identified interesting cases of accelerated evolution in the PBC class of homeobox genes. The vertebrate PG PBX4 is evolving more rapidly than its sister PGs. The accelerated rate of evolution can be explained by functional changes. In mammals, PBX4 functions during spermatogenesis (Wagner et al. 2001), while in zebrafish PBX4 is involved in hindbrain development (Waskiewicz et al. 2002). Another instance of high rates of evolution is found in the PBC class genes in nematodes. One gene, *ceh-20*, is highly conserved, while *ceh-40* and *ceh-60* are evolving much more rapidly. In *P. pacificus* four PBC gene fragments are found, two of which could be homologous to *ceh-40* and *ceh-60*, respectively, but have diverged substantially. *ceh-40* is presently still evolving faster than other homeobox genes, as shown by the differences in this gene between the closely related Caenorhabditis species.

*ceh-20* and *ceh-40* play—at least partially—functionally redundant roles during embryogenesis of *C. elegans*. *ceh-40* mutations alone have no obvious mutant phenotypes, but in conjunction with *ceh-20*, embryonic lethality is increased, and in *unc-62*, *ceh-20*, *ceh-40* triple mutant combinations, embryonic lethality is substantially increased (Van Auken et al. 2002). The differences observed between the Caenorhabditis CEH-40 proteins suggest that CEH-40 is undergoing adaptations in residues on the surface that have the potential to interact with other proteins; likely partners could be other PBC, MEIS, and Hox cluster proteins. Interestingly, both *ceh-40* and *ceh-60* are located on the X chromosome in *C. elegans*, like the fast-evolving X-linked gene TGIFLX (Wang and Zhang 2004). A cluster of three onecut homeobox genes (*ceh-21*, *ceh-39*, and *ceh-41*) is also located on the X chromosome and shows fast divergence and evolution (Bürglin and Cassata 2002), since this cluster is not conserved in *C. briggsae* (K.M. and T.R.B., unpublished). Both *ceh-39* and *ceh-40* are ubiquitously expressed during early embryogenesis, and *ceh-39* has been shown to be involved in sex determination (B. Meyer, personal communication; Van Auken et al. 2002). Hence it may be possible that the fast-evolving, sex-linked homeobox genes *ceh-40* and *ceh-60* are also involved in processes related to specifying sexual differences or contributing to species barriers, either by playing roles in sex determination, by dosage compensation, or by being differentially expressed in the reproductive system.

Vertebrates are thought to have undergone two rounds of genome duplication during early vertebrate evolution. The vertebrate specific expansion of multiple PGs in different classes (four PBC, three MEIS, two PREP, two IRO clusters with three genes each) is consistent with this. But only in the case of the PBC class did all four genes survive the duplication events. In the case of TGIF and MKX, only a single PG seems to have been present in early vertebrate evolution, and the other duplicated copies were presumably eliminated. In the case of the TGIF class, extra genes arose during vertebrate evolution later.

The evolutionary position of nematodes is still debated. While we favor the newer model of nematodes belonging to a clade of ecdysozoa (Aguinaldo et al. 1997; Copley et al. 2004; Mallatt and Winchell 2002), or at least being within the protostomes, the possibility of nematodes forming an outgroup to the coelomates has not been refuted completely (Wolf et al. 2004). Because all TALE classes, perhaps with the exception of MKX, exist also in Cnidaria, we know that the PREP homeodomain and the TGIF gene were lost in Caenorhabditis. In the case of the MKX gene it is formally still possible that if nematodes are an outgroup, the MKX gene arose after the separation of nematodes and does not represent a loss in nematodes. However, we consider this unlikely, since the loss of homeobox genes in *C. elegans* is not unusual. Genes within the Hox cluster have also been shown to have been lost during nematode evolution (Aboobaker and Blaxter 2003).

In conclusion, we can confidently state that the minimal number of TALE homeobox genes at the divergence of protostomes and deuterostomes was six. While it is formally possible that this number could be higher, and that independent gene loss occurred in both protostomes and deuterostomes, we consider this alternative less likely. The evolution of the IRO/MKX, MEIS/TGIF, and PBC clades in early metazoan evolution is presently uncertain, and further full-genome data from, for example, sponges and Placozoa will hopefully provide further insights. Ultimately, these genes all arose from a single ancestral TALE homeodomain protein with a MEINOX domain.

## Note Added in Proof

The present official nomenclature for the human and mouse PREP genes is PKNOX1 and PKNOX2, and Pknox1 and Pknox2, respectively, which in full stands for "Pbx/knotted 1 homeobox." This naming is unfortunate, as the Pbx genes are very distinct, and furthermore in plants a distinct TALE class called "KNOX" is present. Hence, we suggest to retain PREP as the family name here. Note that the list of pseudogenes is not complete.

*Acknowledgments.* We thank Barbara Meyer for discussion and unpublished information. We thank Christoph Dieterich, Ralf Sommer, and Ulrich Technau for comments on the Pristionchus and Nematostella sequence data. Preliminary sequence data for *B. malayi* are deposited regularly into the GSS division of GenBank. The sequencing effort is part of the International Brugia Genome Sequencing Project and is supported by an award from the National Institute of Allergy and Infectious Diseases, National Institutes of Health. This research was supported by the Swedish Foundation for Strategic Research and the Swedish Research Council.

## References

Aboobaker AA, Blaxter ML (2003) Hox gene loss during dynamic evolution of the nematode cluster. Curr Biol 13:37–40

Abu-Shaar M, Ryoo HD, Mann RS (1999) Control of the nuclear localization of Extradenticle by competing nuclear import and export signals. Genes Dev 13:935–945

Adams MD, Celniker SE, Holt RA, et al. (2000) The genome sequence of Drosophila melanogaster. Science 287:2185–95

Aguinaldo AMA, Turbeville JM, Linford LS, Rivera MC, Garey JR, Raff RA, Lake JA (1997) Evidence for a clade of nematodes, arthropods and other moulting animals. Nature 387:489–493

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

Anderson DM, Arredondo J, Hahn K, Valente G, Martin JF, Wilson-Rawls J, Rawls A (2006) Mohawk is a novel homeobox gene expressed in the developing mouse embryo. Dev Dyn 235:792–801

Belloni E, Martucciello G, Verderio D, Ponti E, Seri M, Jasonni V, Torre M, Ferrari M, Tsui LC, Scherer SW (2000) Involvement of the HLXB9 homeobox gene in Currarino syndrome. Am J Hum Genet 66:312–319

Berthelsen J, Zappavigna V, Mavilio F, Blasi F (1998) Prep1, a novel functional partner of Pbx proteins. EMBO J 17:1423–1433

Bertolino E, Reimund B, Wildt-Perinic D, Clerc RG (1995) A novel homeobox protein which recognizes a TGT core and functionally interferes with a retinoid-responsive motif. J Biol Chem 270:31178–31188

Bielawski JP, Yang Z (2003) Maximum likelihood methods for detecting adaptive evolution after gene duplication. J Struct Funct Genomics 3:201–212

Bürglin TR (1995) The evolution of homeobox genes. In: Arai R, Kato M, Doi Y (eds) Biodiversity and evolution. National Science Museum Foundation, Tokyo, pp 291–336

Bürglin TR (1997) Analysis of TALE superclass homeobox genes (MEIS, PBC, KNOX, Iroquois, TGIF) reveals a novel domain conserved between plants and animals. Nucleic Acids Res 25:4173–4180

Bürglin TR (1998a) The PBC domain contains a MEINOX domain: Coevolution of Hox and TALE homeobox genes? Dev Genes Evol 208:113–116

Bürglin TR (1998b) PPCMatrix: a PowerPC dotmatrix program to compare large genomic sequences against protein sequences. Bioinformatics 14:751–752

Bürglin TR (2005) Homeodomain proteins. In: Meyers RA (ed) Encyclopedia of molecular cell biology and molecular medicine. Wiley-VCH Verlag, Weinheim, pp 179–222

Bürglin TR, Cassata G (2002) Loss and gain of domains during evolution of cut superclass homeobox genes. Int J Dev Biol 46:115–123

Bürglin TR, Ruvkun G (1992) New motif in PBX genes. Nature Genet 1:319–320

Chourrout D, Delsuc F, Chourrout P, Edvardsen RB, Rentzsch F, Renfer E, Jensen MF, Zhu B, de Jong P , Steele RE, Technau U (2006) Minimal ProtoHox cluster inferred from bilaterian and cnidarian Hox complements. Nature 442:684–687

Consortium IHGS(2004) Finishing the euchromatic sequence of the human genome. Nature 431:931–945Consortium IHGS (2004) Finishing the euchromatic sequence of the human genome. Nature 431:931–945

Copley RR, Aloy P, Russell RB, Telford MJ (2004) Systematic searches for molecular synapomorphies in model metazoan genomes give some support for Ecdysozoa after accounting for the idiosyncrasies of Caenorhabditis elegans. Evol Dev 6:164–169

De Ley P (2006) A quick tour of nematode diversity and the backbone of nematode phylogeny. In: Community TCeR (ed) WormBook; http://www.wormbook.org

Dildrop R, Rçther U (2004) Organization of Iroquois genes in fish. Dev Genes Evol 214:267–276

Duboule D (1994) Guidebook to the homeobox genes. Oxford University Press, Oxford

Edgar RC (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5:113

Feijóo CG, Manzanares M, de la Calle-Mustienes E , Gómez-Skarmeta JL, Allende ML (2004) The Irx gene family in zebrafish: genomic structure, evolution and initial characterization of irx5b. Dev Genes Evol 214:277–284

Flegel WA, Singson AW, Margolis JS, Bang AG, Posakony JW, Murre C (1993) *Dpbx*, a new homeobox gene closely related to the human proto-oncogene *pbx1*. Molecular structure and developmental expression. Mech Dev 41:155–161

Fognani C, Kilstrup-Nielsen C, Berthelsen J, Ferretti E, Zappavigna V, Blasi F (2002) Characterization of PREP2, a paralog of PREP1, which defines a novel sub-family of the MEINOX TALE homeodomain transcription factors. Nucleic Acids Res 30:2043–2051

Fu SW, Schwartz A, Stevenson H, Pinzone JJ, Davenport GJ, Orenstein JM, Gutierrez P, Simmens SJ, Abraham J, Poola I, Stephan DA, Berg PE (2003) Correlation of expression of BP1, a homeobox gene, with estrogen receptor status in breast cancer. Breast Cancer Res 5:R82–R87

Geerts D, Schilderink N, Jorritsma G, Versteeg R (2003) The role of the MEIS homeobox genes in neuroblastoma. Cancer Lett 197:87–92

Gehring WJ, Affolter M, Bçrglin TR (1994) Homeodomain proteins. Annu Rev Biochem 63:487–526

Gómez-Skarmeta J-L, Diez del Corral R, de la Calle-Mustienes E, Ferrès-Marcû D, Modolell J (1996) *araucan* and *caupolican*, two members of the novel Iroquois complex, encode homeoproteins that control proneural and vein-forming genes. Cell 85:95–105

Gregory SG, Barlow KF, McLay KE, et al. (2006) The DNA sequence and biological annotation of human chromosome 1. Nature 441:315–321

Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol 52:696–704

Huang H, Rastegar M, Bodner C, Goh SL, Rambaldi I, Featherstone M (2005) MEIS C termini harbor transcriptional activation domains that respond to cell signaling. J Biol Chem 280:10119–10127

Hubbard T, Andrews D, Caccamo M, Cameron G, Chen Y, Clamp M, Clarke L, Coates G, Cox T, Cunningham F, Curwen V, Cutts T, Down T, Durbin R, Fernandez-Suarez XM, Gilbert J, Hammond M, Herrero J, Hotz H, Howe K, Iyer V, Jekosch K, Kahari A, Kasprzyk A, Keefe D, Keenan S, Kokocinsci F, London D, Longden I, McVicker G, Melsopp C, Meidl P, Potter S, Proctor G, Rae M, Rios D, Schuster M, Searle S, Severin J, Slater G, Smedley D, Smith J, Spooner W, Stabenau A, Stalker J, Storey R, Trevanion S, Ureta-Vidal A, Vogel J, White S, Woodwark C, Birney E (2005) Ensembl 2005. Nucleic Acids Res 33:D447–D453

Hyman CA, Bartholin L, Newfeld SJ, Wotton D (2003) Drosophila TGIF proteins are transcriptional activators. Mol Cell Biol 23:9262–9274

Jones DT, Taylor WR, Thornton JM (1992) The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci 8:275–282

Li T, Stark MR, Johnson AD, Wolberger C (1995) Crystal structure of the MATa1/MATα2 homeodomain heterodimer bound to DNA. Science 270:262–269

Maclean JA 2nd, Chen MA, Wayne CM, Bruce SR, Rao M, Meistrich ML, Macleod C, Wilkinson MF (2005) Rhox: a new homeobox gene cluster. Cell 120:369–382

Mallatt J, Winchell CJ (2002) Testing the new animal phylogeny: first use of combined large-subunit and small-subunit rRNA gene sequences to classify the protostomes. Mol Biol Evol 19:289–301

Man YG, Fu SW, Schwartz A, Pinzone JJ, Simmens SJ, Berg PE (2005) Expression of BP1, a novel homeobox gene, correlates with breast cancer progression and invasion. Breast Cancer Res Treat 90:241–247

Mann RS, Affolter M (1998) Hox proteins meet more partners. Curr Opin Genet Dev 8:423–429

Maulbecker CC, Gruss P (1993) The oncogenic potential of deregulated homeobox genes. Cell Growth Differ 4:431–441

Olson SA (2002) EMBOSS opens up sequence analysis. European Molecular Biology Open Software Suite. Brief Bioinform 3:87–91

Pérez-Bercoff Å, Koch J, Bürglin TR (2006) LogoBar: a Java application to visualize protein logos with gaps. Bioinformatics 22:112–114

Perri P, Bachetti T, Longo L, Matera I, Seri M, Tonini GP, Ceccherini I (2005) PHOX2B mutations and genetic predisposition to neuroblastoma. Oncogene 24:3050–3053

Rauskolb C, Peifer M, Wieschaus E (1993) *extradenticle*, a regulator of homeotic gene activity, is a homolog of the homeobox-containing human proto-oncogene *pbx1*. Cell 74:1101–1112

Rieckhof GE, Casares F, Ryoo HD, Abu-Shaar M, Mann RS (1997) Nuclear Translocation of Extradenticle Requires *homothorax*, which Encodes an Extradenticle-Related Homeodomain Protein. Cell 91:171–183

Ryan JF, Burton PM, Mazza ME, Kwong GK, Mullikin JC, Finnerty JR (2006) The cnidarian-bilaterian ancestor possessed at least 56 homeoboxes. Evidence from the starlet sea anemone, Nematostella vectensis. Genome Biol 7:R64

Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4:406–425

Soulier J, Clappier E, Cayuela JM, Regnault A, Garcia-Pedro M, Dombret H, Baruchel A, Toribio ML, Sigaux F (2005) HOXA genes are included in genetic and biological networks defining human acute T-cell leukemia (T-ALL). Blood 106:274–286

Sullivan JC, Ryan JF, Watson JA, Webb J, Mullikin JC, Rokhsar D, Finnerty JR (2006) StellaBase: the Nematostella vectensis Genomics Database. Nucleic Acids Res 34:D495–D499

Suyama M, Torrents D, Bork P (2006) PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res 34:W609–W612

The *C. elegans* Sequencing Consortium (1998) Genome sequence of the nematode *C. elegans*: a platform for investigating biology. Science 282:2012–2018

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res 25:4876–4882

Thorsteinsdottir U, Sauvageau G, Hough MR, Dragowska W, Lansdorp PM, Lawrence HJ, Largman C, Humphries RK (1997) Overexpression of HOXA10 in murine hematopoietic cells perturbs both myeloid and lymphoid differentiation and leads to acute myeloid leukemia. Mol Cell Biol 17:495–505

Ting CT, Tsaur SC, Wu ML, Wu CI (1998) A rapidly evolving homeobox at the site of a hybrid sterility gene. Science 282:1501–1504

Tomancak P, Beaton A, Weiszmann R, Kwan E, Shu S, Lewis SE, Richards S, Ashburner M, Hartenstein V, Celniker SE, Rubin GM (2002) Systematic determination of patterns of gene expression during Drosophila embryogenesis. Genome Biol 3:RESEARCH0088

Van Auken K, Weaver D, Robertson B, Sundaram M, Saldi T, Edgar L, Elling U, Lee M, Boese Q, Wood WB (2002) Roles of the Homothorax/Meis/Prep homolog UNC-62 and the Exd/Pbx homologs CEH-20 and CEH-40 in C. elegans embryogenesis. Development 129:5255–5268

Wagner K, Mincheva A, Korn B, Lichter P, Popperl H (2001) Pbx4, a new Pbx family member on mouse chromosome 8, is expressed during spermatogenesis. Mech Dev 103:127–131

Wang X, Zhang J (2004) Rapid evolution of mammalian X-linked testis-expressed homeobox genes. Genetics 167:879–888

Wang Z, Mann RS (2003) Requirement for two nearly identical TGIF-related homeobox genes in Drosophila spermatogenesis. Development 130:2853–2865

Waskiewicz AJ, Rikhof HA, Moens CB (2002) Eliminating zebrafish pbx proteins reveals a hindbrain ground state. Dev Cell 3:723–733

Wolf YI, Rogozin IB, Koonin EV (2004) Coelomata and not Ecdysozoa: evidence from genome-wide phylogenetic analysis. Genome Res 14:29–36

Yang Z (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 13:555–556

Yang Z, Nielsen R, Goldman N, Pedersen AM (2000) Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics 155:431–449