Int. J. Dev. Biol. 46: 115-123 (2002)

Loss and gain of domains during evolution of *cut* superclass homeobox genes

THOMAS R. BÜRGLIN*[1] and GIUSEPPE CASSATA[2]

Division of Cell Biology, Biozentrum, University of Basel, Switzerland

ABSTRACT The *cut* superclass of homeobox genes has been divided into three classes: *CUX, ONECUT* and *SATB*. Given the various completed genomes, we have now made a comprehensive survey. We find that there are only two *cut* domain containing genes in *Drosophila*, one *CUX* and one *ONECUT* type. *Caenorhabditis elegans* has undergone an expansion of the *ONECUT* subclass genes and has a gene cluster with three *ONECUT* class genes, one of which has lost the *cut* domain. Two of these genes contain a conserved sequence motif, termed OCAM, which also occurs in another gene in *C. elegans*; this motif seems to be nematode specific. A recently uncovered *C. elegans CUX* gene has sequence conservation in its amino-terminus with vertebrate CUX proteins. Further, the 5' end of this gene containing the conserved region can undergo alternative splicing to give rise to a protein with a different carboxy-terminus lacking the cut- and homeodomain. This protein is conserved in its entirety with vertebrate genes termed *CASP* - which are also alternative splice products of the *CUX* genes - and with plant and fungal genes. The highly divergent *SATB* genes share a conserved amino-terminal domain, COMPASS, with the *Drosophila defective proventriculus* gene and a *C. elegans* ORF. These two *"COMPASS"* family genes encode two highly divergent homeodomains, may be homologues of the *SATB* genes and thus probably belong to the cut superclass, too.

KEY WORDS: *homeobox gene, cut, ONECUT, CUX, evolution*

Introduction

Homeobox genes encode transcription factors that have been found to play fundamental roles in the development of animals, and also in plants and fungi they control developmental decisions (see for example, Duboule, 1994). Homeobox genes have been classified into many different classes and families based on their degree of sequence similarity within the homeodomain (Bürglin, 1994; Bürglin, 1995; Ruddle *et al.*, 1994). One group was named after its founding member, *cut*, a homeobox gene that is involved in external sensory organ development in *Drosophila melanogaster* (Blochlinger *et al.*, 1990). The Cut protein contains three copies of the cut repeat upstream of the homeodomain. This gene structure, with three cut domains and one homeodomain was subsequently found to be conserved in vertebrate homologues (Andres *et al.*, 1992; Neufeld *et al.*, 1992). The cut domain has been shown to be a DNA-binding domain (for example, Harada *et al.*, 1995; Lannoy *et al.*, 1998).

We have previously surveyed and classified all of the then known cut superclass homeobox genes into three classes (Lannoy *et al.*, 1998): 1)The CUX class genes with the founding member *cut*. These genes encode three cut domains upstream of a homeodomain. 2) The ONECUT class genes, which encode only a single cut domain upstream of a homeodomain, represented by genes such as HNF-6 (Lemaigre *et al.*, 1996). 3) The SATB class genes, which are highly divergent and encode two cut domains upstream of a homeodomain. In initial classifications of homeobox genes cut genes were defined as a class (Bürglin, 1994). Lannoy *et al.*, (1998) has given the cut genes the status of "superclass", because of the high degree of sequence divergence between these genes. For instance, the SATB genes were not even recognised initially as cut homeobox genes (Nakagomi *et al.*, 1994). Nevertheless, these genes are considered to be monophyletic in origin, based on structural (cut domain) and phylogenetic grounds (see below).

Abbreviations used in this paper: C. elegans, Caenorhabditis elegans; CMP, COMPASS; dve, defective proventriculus; ORF, open reading frame.

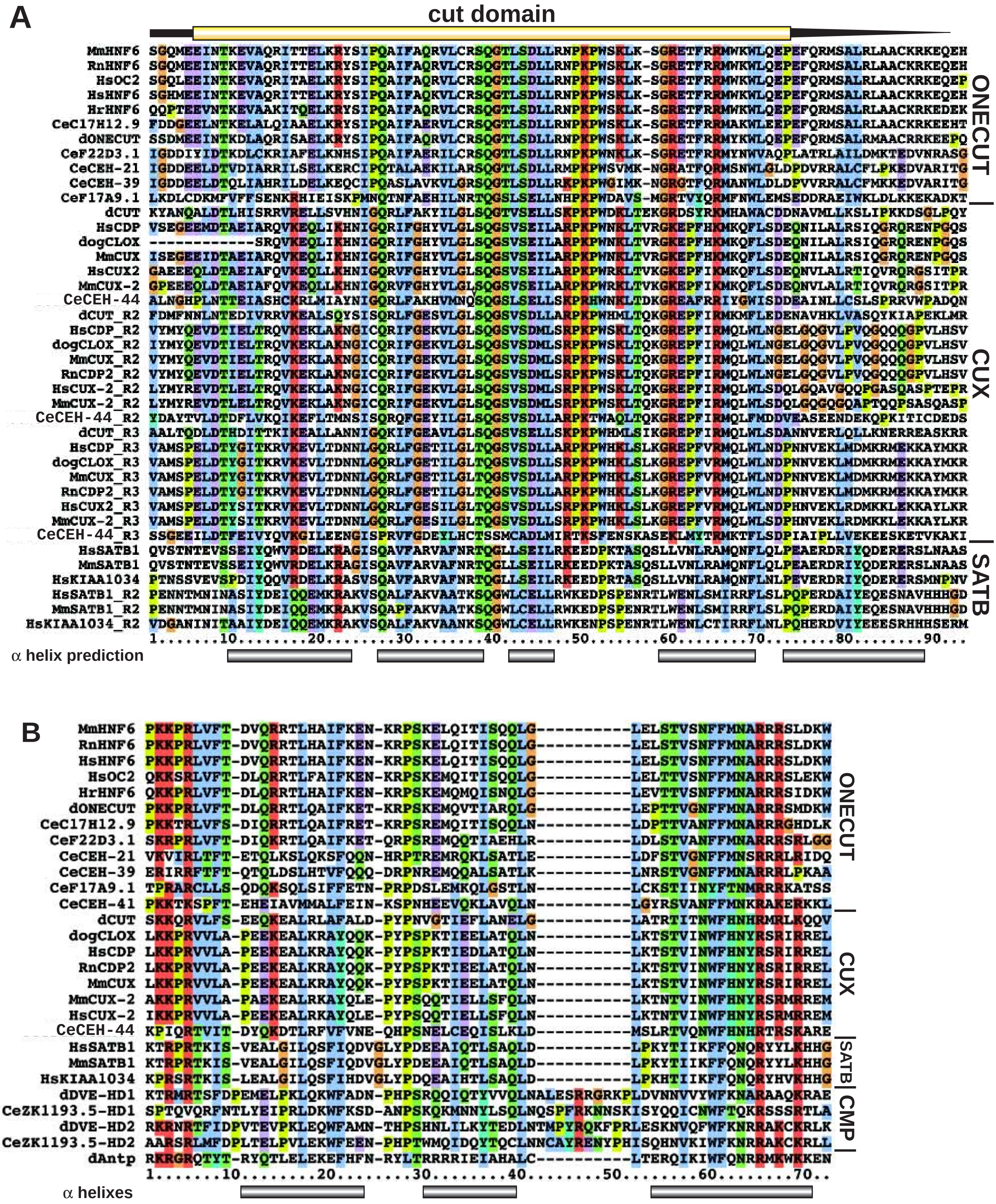
[1] **Present address:** Department of Bioscienes at Novum, and Center for Genomics and Bioinformatics. Karolinska Institutet. Södertörns Högskola, Box 4101, Alfred Nobels Alle 10, SE-141 04 Huddinge, Sweden.

[2] **Present address:** Adolf Butenandt-Institut. Molecular Neurogenetics, Schillerstr. 44, D-80336 Munich, Germany.

***Address correspondence to:** Thomas Bürglin. Södertörns Högskola, Box 4101, SE-141 04 Huddinge, Sweden. Fax: +46-8585-88510.
e-mail: thomas.burglin@biosci.ki.se

0214-6282/2002/$25.00

Printed in Spain
www.ijdb.ehu.es



```
A
MmHNF6         SGQMEEINTKEVAQRITTELKRYSIPQAIFAQRVLCRSQGTLSDLLRNPKPWSKLK-SGRETFRRMWKWLQEPEFQRMSALRLAACKRKEQEH
RnHNF6         SGQMEEINTKEVAQRITTELKRYSIPQAIFAQRVLCRSQGTLSDLLRNPKPWSKLK-SGRETFRRMWKWLQEPEFQRMSALRLAACKRKEQEH
HsOC2          SGQLEEINTKEVAQRITAELKRYSIPQAIFAQRVLCRSQGTLSDLLRNPKPWSKLK-SGRETFRRMWKWLQEPEFQRMSALRLAACKRKEQEP
HsHNF6         SGHMEEINTKEVAQRITTELKRYSIPQAIFAQKVLCRSQGTLSDLLRNPKPWSKLK-SGRETFRRMWKWLQEPEFQRMSALRLAACKRKEQEH
HrHNF6         QQPTEEVNTKEVAAKITQELKRYSIPQAIFAQRVLCRSQGTLSDLLRNPKPWSKLK-SGRETFRRMWKWLQEPEFQRMSALRLAACKRKEDEK
CeC17H12.9     FDDGEELNTKELALQIAAELKRYSIPQAIFAERVLCRSQGTLSDLLRNPKPWSKLK-SGRETFRRMAKWLEEPEFQRMSALRLAACKRKEEHT
dONECUT        SSDMEEINTKDLAQRISAELKRYSIPQAIFAQRVLCRSQGTLSDLLRNPKPWSKLK-SGRETFRRMYKWLQEPEFQRMSALRMAACRRKEEPQ
CeF22D3.1      IGDDIYIDTKDLCKRIAFELKNHSIPQAIFAERILCRSQGTLSDLLRNPKPWNKLK-SGRETFRRMYNWVAQPLATRLAILDMKTEDVNRASG
CeCEH-21       IGDDEELDTVDIARRILSELKERCIPQTALAEKILARSQGTLSDLLRMPKPWSVMK-NGRATFQRMSNWLGLDPDVRRALCFLPKEDVARITG
CeCEH-39       IGDDEELDTQLIAHRILDELKEQCIPQASLAVKVLGRSQGTLSDLLRKPKPWGIMK-NGRGTFQRMANWLDLDPVVRRALCFMKKEDVARITG
CeF17A9.1      LKDLCDKMFVFFSENKRHIEISKPMNQTNFAEHILNRTQGSLSELLNHPKPWDAVS-MGRTVYQRMFNWLEMSEDDRAEIWKLDLKKEKADKT
dCUT           KYANQALDTLHISRRVRELLSVHNIGQRLFAKYILGLSQGTVSELLSKPKPWDKLTEKGRDSYRKMHAWACDDNAVMLLKSLIPKKDSGLPQY
HsCDP          VSEGEEMDTAEIARQVKEQLIKHNIGQRIFGHYVLGLSQGSVSEILARPKPWNKLTVRGKEPFHKMKQFLSDEQNILALRSIQGRQRENPGQS
dogCLOX        ------------SRQVKEQLIKHNIGQRIFGHYVLGLSQGSVSEILARPKPWNKLTVRGKEPFHKMKQFLSDEQNILALRSIQGRQRENPGQS
MmCUX          ISEGEEIDTAEIARQVKEQLIKHNIGQRIFGHYVLGLSQGSVSEILARPKPWNKLTVRGKEPFHKMKQFLSDEQNILALRSIQGRQRENPGQS
HsCUX2         GAEEEQLDTAEIAFQVKEQLLKHNIGQRVFGHYVLGLSQGSVSEILARPKPWRKLTVKGKEPFIKMKQFLSDEQNVLALRTIQVRQRGSITPR
MmCUX-2        GPEEEQLDTAEIAFQVKEQLLKHNIGQRVFGHYVLGLSQGSVSEILARPKPWRKLTVKGKEPFIKMKQFLSDEQNVLALRTIQVRQRGSITPR
CeCEH-44       ALNGHPLNTTEIASHCKRLMIAYNIGQRLFAKHVMNQSQGSLSELLSKPRHWNKLTDKGREAFRRIYGWISDDEAINLLCSLSPRRVWPADQN
dCUT_R2        FDMFNNLNTEDIVRRVKEALSQYSISQRLFGESVLGLSQGSVSDLLARPKPWHMLTQKGREPFIRMKMFLEDENAVHKLVASQYKIAPEKLMR
HsCDP_R2       VYMYQEVDTIELTRQVKEKLAKNGICQRIFGEKVLGLSQGSVSDMLSRPKPWSKLTQKGREPFIRMQLWLNGELGQGVLPVQGQQQGPVLHSV
dogCLOX_R2     IYMYQEVDTIELTRQVKEKLAKNGICQRIFGEKVLGLSQGSVSDMLSRPKPWSKLTQKGREPFIRMQLWLNGELGQGVLPVQGQQQGPVLHSV
MmCUX_R2       VYMYQEVDTIELTRQVKEKLAKNGICQRIFGEKVLGLSQGSVSDMLSRPKPWSKLTQKGREPFIRMQLWLNGELGQGVLPVQGQQQGPVLHSV
RnCDP2_R2      VYMYQEVDTIELTRQVKEKLAKNGICQRIFGEKVLGLSQGSVSDMLSRPKPWSKLTQKGREPFIRMQLWLNGELGQGVLPVQGQQQGPVLHSV
HsCUX-2_R2     LYMYREVDTLELTRQVKEKLAKNGICQRIFGEKVLGLSQGSVSDMLSRPKPWSKLTQKGREPFIRMQLWLSDQLGQAVGQQPGASQASPTEPR
MmCUX-2_R2     LYMYREVDTLELTRQVKEKLAKNGICQRIFGEKVLGLSQGSVSDMLSRPKPWSKLTQKGREPFIRMQLWLSDQLGQGQGQAPTQQPSASQASP
CeCEH-44_R2    YDAYTVLDTDFLVKQIKEFLTMNSISQRQFGEYILGLSQGSVSDLLARPKTWAQLTQKGREPFIRMQLFMDDVEASEENDEKQPKITICDEDS
dCUT_R3        AALTQDLDTHDITTKIKEALLANNIGQKIFGEAVLGLSQGSVSELLSKPKPWHMLSIKGREPFIRMQLWLSDANNVERLQLLKNERREASKRR
HsCDP_R3       VAMSPELDTYGITKRVKEVLTDNNLGQRLFGETILGLTQGSVSDLLARPKPWHKLSLKGREPFVRMQLWLNDPNNVEKLMDMKRMEKKAYMKR
dogCLOX_R3     VAMSPELDTYGITKRVKEVLTDNNLGQRLFGETILGLTQGSVSDLLSRPKPWHKLSLKGREPFVRMQLWLNDPNNVEKLMDMKRMEKKAYMKR
MmCUX_R3       VAMSPELDTYGITKRVKEVLTDNNLGQRLFGETILGLTQGSVSDLLARPKPWHKLSLKGREPFVRMQLWLNDPNNVEKLMDMKRMEKKAYMKR
RnCDP2_R3      VAMSPELDTYGITKRVKEVLTDNNLGQRLFGETILGLTQGSVSDLLARPKPWHKLSLKGREPFVRMQLWLNDPNNVEKLMDMKRMEKKAYMKR
HsCUX2_R3      VAMSPELDTYSITKRVKEVLTDNNLGQRLFGESILGLTQGSVSDLLSRPKPWHKLSLKGREPFVRMQLWLNDPHNVEKLRDMKKLEKKAYLKR
MmCUX-2_R3     VAMSPELDTYSITKRVKEVLTDNNLGQRLFGESILGLTQGSVSDLLSRPKPWHKLSLKGREPFVRMQLWLSDPHNVEKLRDMKKLEKKAYLKR
CeCEH-44_R3    SSGEEILDTFEIVYQVKGILEENGISPRVFGDEYLHCTSSMCADLMIRTKSFENSKASEKLMYTRMKTFLSDPIAIPLLVEKEESKETVKAKI
HsSATB1        QVSTNTEVSSEIYQWVRDELKRAGISQAVFARVAFNRTQGLLSEILRKEEDPKTASQSLLVNLRAMQNFLQLPEAERDRIYQDERERSLNAAS
MmSATB1        QVSTNTEVSSEIYQWVRDELKRAGISQAVFARVAFNRTQGLLSEILRKEEDPKTASQSLLVNLRAMQNFLQLPEAERDRIYQDERERSLNAAS
HsKIAA1034     PTNSSVEVSPDIYQQVRDELKRASVSQAVFARVAFNRTQGLLSEILRKEEDPRTASQSLLVNLRAMQNFLNLPEVERDRIYQDERERSMNPNV
HsSATB1_R2     PENNTMNINASIYDEIQQEMKRAKVSQALFAKVAATKSQGWLCELLRWKEDPSPENRTLWENLSMIRRFLSLPQPERDAIYEQESNAVHHHGD
MmSATB1_R2     PENNTMNINASIYDEIQQEMKRAKVSQAPFAKVAATKSQGWLCELLRWKEDPSPENRTLWENLSMIRRFLSLPQPERDAIYEQESNAVHHHGD
HsKIAA1034_R2  VDGANINITAAIYDEIQQEMKRAKVSQALFAKVAANKSQGWLCELLRWKENPSPENRTLWENLCTIRRFLNLPQHERDVIYEEESRHHHSERM
               1.......10........20........30........40........50........60........70........80........90...
α helix prediction

B
MmHNF6          PKKPRLVFT-DVQRRTLHAIFKEN-KRPSKELQITISQQLG----------LELSTVSNFFMNARRRSLDKW
RnHNF6          PKKPRLVFT-DVQRRTLHAIFKEN-KRPSKELQITISQQLG----------LELSTVSNFFMNARRRSLDKW
HsHNF6          PKKPRLVFT-DVQRRTLHAIFKEN-KRPSKELQITISQQLG----------LELSTVSNFFMNARRRSLDKW
HsOC2           QKKSRLVFT-DLQRRTLFAIFKEN-KRPSKEMQITISQQLG----------LELTTVSNFFMNARRRSLEKW
HrHNF6          QKKPRLVFT-DLQRRTLHAIFKEN-KRPSKEMQMQISNQLG----------LEVTTVSNFFMNARRRSIDKW
dONECUT         PKKPRLVFT-DLQRRTLQAIFKET-KRPSKEMQVTIARQLG----------LEPTTVGNFFMNARRRSMDKW
CeC17H12.9      PKKTRLVFS-DIQRRTLQAIFRET-KRPSREMQITISQQLN----------LDPTTVANFFMNARRRGHDLK
CeF22D3.1       SKRPRLVFT-DIQKRTLQAIFKET-QRPSREMQQTIAEHLR----------LDLSTVANFFMNARRRSRLGG
CeCEH-21        VKVIRLTFT-ETQLKSLQKSFQQN-HRPTREMRQKLSATLE----------LDFSTVGNFFMNSRRRLRIDQ
CeCEH-39        ERIRRFTFT-QTQLDSLHTVFQQQ-DRPNREMQQALSATLK----------LNRSTVGNFFMNARRRLPKAA
CeF17A9.1       TPRARCLLS-QDQKSQLSIFFETN-PRPDSLEMKQLGSTLN----------LCKSTIINYFTNMRRRKATSS
CeCEH-41        PKKTKSPFT-EHEIAVMMALFEIN-KSPNHEEVQKLAVQLN----------LGYRSVANFFMNKRAKERKKL
dCUT            SKKQRVLFS-EEQKEALRLAFALD-PYPNVGTIEFLANELG----------LATRTITNWFHNHRMRLKQQV
dogCLOX         LKKPRVVLA-PEEKEALKRAYQQK-PYPSPKTIEELATQLN----------LKTSTVINWFHNYRSRIRREL
HsCDP           LKKPRVVLA-PEEKEALKRAYQQK-PYPSPKTIEDLATQLN----------LKTSTVINWFHNYRSRIRREL
RnCDP2          LKKPRVVLA-PEEKEALKRAYQQK-PYPSPKTIEELATQLN----------LKTSTVINWFHNYRSRIRREL
MmCUX           LKKPRVVLA-PEEKEALKRAYQQK-PYPSPKTIEELATQLN----------LKTSTVINWFHNYRSRIRREL
MmCUX-2         AKKPRVVLA-PAEKEALRKAYQLE-PYPSQQTIELLSFQLN----------LKTNTVINWFHNYRSRMRREM
HsCUX-2         IKKPRVVLA-PEEKEALRKAYQLE-PYPSQQTIELLSFQLN----------LKTNTVINWFHNYRSRMRREM
CeCEH-44        KPIQRTVIT-DYQKDTLRFVFVNE-QHPSNELCEQISLKLD----------MSLRTVQNWFHNHRTRSKARE
HsSATB1         KTRPRTKIS-VEALGILQSFIQDVGLYPDEEAIQTLSAQLD----------LPKYTIIKFFQNQRYYLKHHG
MmSATB1         KTRPRTKIS-VEALGILQSFIQDVGLYPDEEAIQTLSAQLD----------LPKYTIIKFFQNQRYYLKHHG
HsKIAA1034      KPRSRTKIS-LEALGILQSFIHDVGLYPDQEAIHTLSAQLD----------LPKHTIIKFFQNQRYHVKHHG
dDVE-HD1        KTRMRTSFDPEMELPKLQKWFADN-PHPSRQQIQTYVVQLNALESRRGRKPLDVNNVVYWFKNARAAQKRAE
CeZK1193.5-HD1  SPTQVQRFNTLYEIPRLDKWFKSD-ANPSKQKMNNYLSQLNQSPFRKNNSKISYQQICNWFTQKRSSSRTLA
dDVE-HD2        RKRNRTFIDPVTEVPKLEQWFAMN-THPSHNLILKYTEDLNTMPYRQKFPRLESKNVQFWFKNRRAKCKRLK
CeZK1193.5-HD2  AARSRLMFDPLTELPVLEKWFEEN-PHPTWMQIDQYTQCLNNCAYRENYPHISQHNVKIWFKNRRAKCKRLL
dAntp           RKRGRQTYT-RYQTLELEKEFHFN-RYLTRRRRIEIAHALC----------LTERQIKIWFQNRRMKWKKEN
                1.......10........20........30........40........50........60........70..
α helixes
```

Since several complete genomes have now become available, we searched the sequence databases for new members to establish comprehensive overviews for these genes and to determine how they evolved. We find several conserved motifs, alternative splicing resulting from a merger with another gene, and another homeobox gene class that we term COMPASS, which may be orthologous to the vertebrate SATB genes.

Results and Discussion

Survey of cut *Superclass Homeobox Genes*

To generate an updated tally of cut superclass homeobox genes we searched the sequence databases with CUX, ONECUT and SATB class members using BLAST and PSI-BLAST. The searches revealed a member of the CUX class in the genome of *C. elegans*, Y54F10AM.4, which has been designated *ceh-44* according to *C. elegans* nomenclature. It had previously not been described, probably because it was in one of the small regions not fully sequenced yet at the time when the complete *C. elegans* genome sequence was announced (Ruvkun and Hobert, 1998; The *C. elegans* Sequencing Consortium, 1998). *ceh-44* is the ortholog of *Drosophila cut*, because of the sequence similarity and the gene structure (see below) and because there are no other similar genes that could be close paralogs in these two species. In vertebrates, two closely related CUX class genes are found (Figs. 1,2). In *Drosophila*, the genome project uncovered a member of the ONECUT class of homeobox genes. In vertebrates, there are two closely related ONECUT families, one represented by the transcription factor HNF-6, and one by ONECUT-2 (OC-2, Jacquemin *et al*., 1999). There are also two closely related SATB genes, but no obvious homologues in flies or *C. elegans* (Figs. 1,2).

The alignment of all the cut domains shows that it is difficult to determine the precise extent of the cut domain (Fig. 1A). Within classes, as well as within particular repeats of the CUX class cut domains, sequence similarity extends beyond a core that is conserved throughout the whole superclass (Fig. 1A). Since the cut domain structure is not known at present, we used the aligned cut domain sequences to perform secondary structure prediction (See experimental protocols). Five alpha helixes are predicted with good scores, but no beta strands. Also, the domain is predicted to be globular in nature. Thus the cut domain is most likely a compact DNA-binding domain composed of alpha helixes, like, for example, the homeodomain or the Paired domain (Qian *et al*., 1989; Xu *et al*., 1995). Since the fourth helix has highly conserved basic residues, we speculate that this helix might lie in the major groove of the DNA and might make sequence-specific contacts.

The alignment of the homeodomain sequences shows that the previous alignment (Lannoy *et al*., 1998) has to be revised, since a better fit can be obtained for the SATB homeodomains if only one amino acid is looped out between helix 1 and helix 2 of the homeodomain (Fig. 1B). In this alignment we also include the COMPASS genes, which encode two highly divergent homeodomains (Fig. 1B).

To determine how the cut superclass homeobox genes may have evolved from each other, we performed phylogenetic analyses using neighbour-joining with the cut and homeodomain sequences (Fig. 2). The analysis of the cut domains shows that the two repeats of the SATB genes are more similar to each other than to other cut domains (Fig. 2A). The same holds true for the three domains of the CUX class genes, which are more similar to each other. In addition, each repeat has been conserved in evolution although fly and worm repeats 1 and 3 have diverged. In particular repeat three of CEH-44 has diverged substantially (see also Fig. 1A). Overall though, it seems clear that the three cut repeats have arisen from a single ancestor and that already at the time of the emergence of the triploblastic animal phyla a CUX gene with three cut domains existed. The second cut domain of the CUX genes is the most highly conserved repeat suggesting that it has been under selective pressure, maybe because this particular repeat contributes most to sequence specific DNA-binding.

The fly and vertebrate ONECUT genes and the *C. elegans* gene C17H12.9 (R07D10.x in Lannoy *et al*., 1998) form a well supported clade (Fig. 2). The other *C. elegans* ONECUT class cut domains are more divergent, in particular F17A9.1. The cut domain of the ascidian gene HNF-6 is an outgroup to vertebrate HNF-6 and OC-2, consistent with the notion that gene duplication occurred in early vertebrate evolution. This also means that ascidian HNF-6 is an orthologue both to vertebrate HNF-6 and OC-2 and its name "HNF-6" is misleading.

Phylogenetic analysis of the homeodomain sequences gives a picture similar to that of the cut domains supporting the view that the homeodomains co-evolved with the cut domains (Fig. 2B). The ONECUT homeodomain sequences form a good clade with exception of the divergent gene F17A9.1, which is the only gene to encode a tyrosine residue at position 48 of the homeodomain instead of a phenylalanine or tryptophan residue (Fig. 1B). Nevertheless, some of the *C. elegans* ONECUT homeodomains are more divergent compared to the fly and vertebrate ONECUT genes (e.g., *ceh-21*, *ceh-39*, *ceh-41*), based on the sequence similarity (Fig. 1B), and the branch lengths. The CUX homeodomain sequences also cluster nicely, except for *C. elegans* CEH-44, which is highly divergent.

Fig.1. Sequence alignment of cut domains and cut class homeodomains. (A) *Aligned cut domain sequences. CUX family genes have three cut domains, repeat 2 and 3 are indicated by the appendix R2 and R3; SATB family genes have two cut domains, the second is indicated by R2. The major core of the cut domain is indicated above the sequences as yellow bar. It is difficult to delineate the boundaries of the domain precisely, as families and individual repeats within families have extended conservation, as indicated by the black bars. Underneath the alignment the secondary structure prediction is shown consisting only of alpha helixes and turns, no beta strands. The first cut repeat of CEH-44 (Y54F10AM.4) was edited by removing three extra residues (at the arrowhead), which resulted from an inappropriately selected splice site in the intron of the first cut repeat (Fig. 4). Species codes for this and subsequent figures: Mm: mouse; Rn: rat; Hs: human; Ce:* C. elegans*; d:* Drosophila melanogaster*; Hr:* Halocynthia roretzi *(sea squirt); Gg: chicken; At:* Arabidopsis thaliana*; Os:* Oryza sativa *(rice); Sc:* Saccharomyces cerevisiae*; Nc:* Neurospora crassa*; Sp:* Schizosaccharomyces pombe. **(B)** *Aligned homeodomain sequences of cut class genes. The alignment of the SATB differs from previous alignments (Lannoy* et al*., 1998), since ClustalX analysis introduced only a gap of one residue in the loop between helix 1 and helix 2 in the typical 60 amino acid homeodomain to accommodate the SATB homeodomains. COMPASS (CMP) genes have two homeodomains (indicated by HD1 and HD2), and extra residues in the loop between helix 2 and helix 3. Helixes are indicated underneath the alignment. The Antennapedia homeodomain is shown for comparison.* C. elegans *genes: F22D3.1 is* ceh-38, ceh-21 *is T26C11.6,* ceh-39 *is T26C11.7,* ceh-41 *is T26C11.5, C17H12.9 was R07D10.x.*

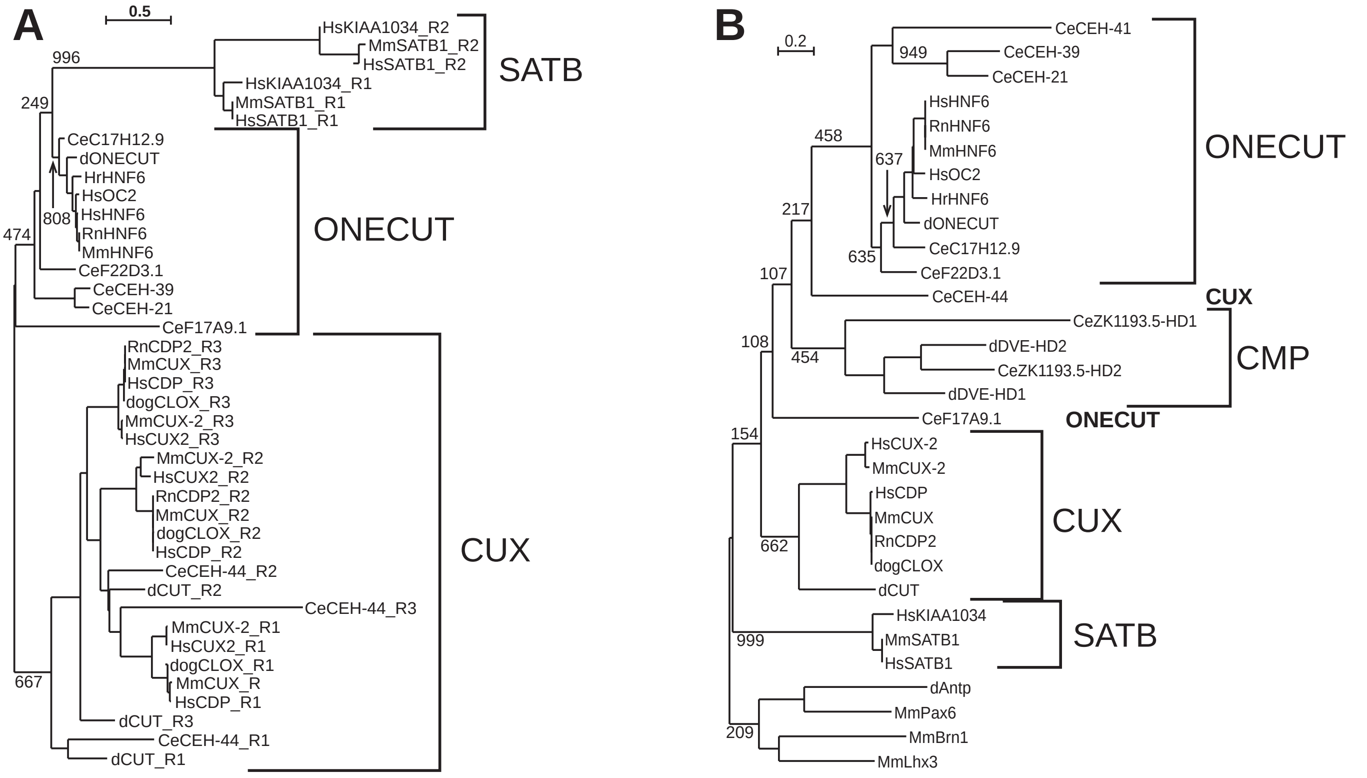


Fig. 2. Phylogenetic tree analysis of cut domains and cut class homeodomains. (A) *Neighbour-joining tree of cut domains from Fig. 1, corrected for multiple substitutions. At selected nodes bootstrap values for 1000 trials are given.* **(B)** *Neighbour-joining tree of cut superclass homeodomains and COMPASS homeodomains corrected for multiple substitutions.* Drosophila *Antennapedia, and mouse Pax-6, Brn-1 and Lhx-3 were used as an outgroup.*

Expansion of C. elegans ONECUT *Genes*

The presence of just one ONECUT gene in flies and two paralogous ones in vertebrates suggests that there was only a single ONECUT gene present in the ancestor of deuterostomes and protostomes. *C. elegans* has multiple ONECUT genes, but it now seems likely that they have arisen by duplication and diversification in the nematode lineage and that no homologues for all of them will be found in other phyla, even though the phylogenetic analysis in Fig. 2 did not result in monophyletic trees. Support for the notion of expansion and diversification of ONECUT genes in nematodes comes from the following observations:

Three of the ONECUT genes occur in a gene cluster (*ceh-39*, *ceh-21*, and *ceh-41*, Fig. 3A), supporting the notion of nematode-specific gene diversification, and the phylogenetic analysis shows that they are more related to each other than to other ONECUT genes (Fig. 2). To determine if these genes are indeed separate genes or if alternative splicing gives rise to genes having several cut domains - at the time no CUX class member with three cut domains was known in *C. elegans* – we isolated several independent cDNAs for *ceh-21*. They resulted in the gene structure shown in Fig. 3A and no evidence, also from all known EST clones, shows that alternative splicing occurs that could link the three genes. Computer ORF prediction of *ceh-21* predicted two extra 5' exons, which apparently are not real (not shown). Inspection of the three genes in this ONECUT cluster (Fig. 3A) also reveals loss of conserved elements. *ceh-41* has a homeodomain related to that of the other two genes (Fig. 2A), but lacks the cut domain. Conversely, there is a small conserved sequence element that we call OCAM (ONECUT-associated motif), which is found only in *ceh-21* and *ceh-41*, but not *ceh-39* (Fig. 3B). Database searches with the OCAM motif revealed one additional gene with this motif in *C. elegans*, T02B5.2 (Fig. 3B). This gene is not related to any other genes in the database, thus most likely it is derived from a ONECUT gene and lost both the cut and homeodomain. It is closely flanked by carboxylesterase genes, which makes it unlikely that computer ORF prediction missed exons.

SATB *Genes and* COMPASS *Domain Genes*

Database searches with SATB genes revealed that in addition to the cut domain and the homeodomain there is significant sequence similarity in the amino-terminus of these genes to *Drosophila defective proventriculus* (*dve*) (Fuss and Hoch, 1998; Nakagoshi *et al*., 1998) and *C. elegans* ZK1193.5 (Fig. 3C). The major sequence similarity extends over a length of about 100 amino acids, and this domain has been termed the COMPASS domain (Fuss and Hoch, 1998). Both *dve* and ZK1193.5 encode two homeodomains (Fig. 1B), which are highly divergent but derived from each other (Fig. 2B). We call this gene family COMPASS (CMP). In comparisons with a large number of homeobox genes, ZK1193.5 does not obviously belong to a particular class of homeobox genes (Ruvkun and Hobert, 1998). Likewise the phylogenetic analysis with cut superclass homeodomains does not reveal close affinity with any class (Fig. 2B), but it shows that these homeodomain sequences have

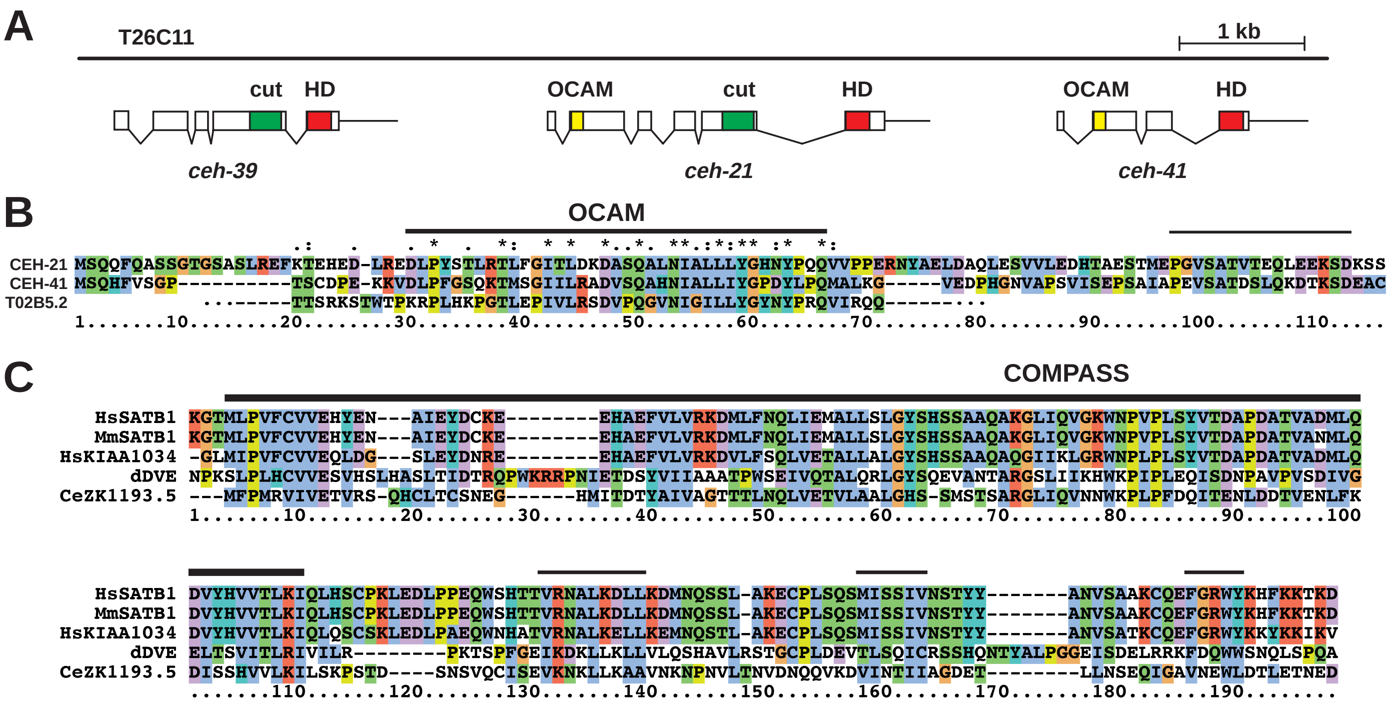

A
T26C11
1 kb
cut
HD
OCAM
cut
HD
OCAM
HD
ceh-39
ceh-21
ceh-41
B
OCAM
CEH-21 MSQQFQASSGTGSASLREFKTEHED-LREDLPYSTLRTLFGITLDKDASQALNIALLLYGHNYPQQVVPPERNYAELDAQLESVVLEDHTAESTMEPGVSATVTEQLEEKSDKSS
CEH-41 MSQHFVSGP----------TSCDPE-KKVDLPFGSQKTMSGIILRADVSQAHNIALLIYGPDYLPQMALKG-----VEDPHGNVAPSVISEPSAIAPEVSATDSLQKDTKSDEAC
T02B5.2 ..------TTSRKSTWTPKRPLHKPGTLEPIVLRSDVPQGVNIGILLYGYNYPRQVIRQQ-------...
1.......10........20........30........40........50........60........70........80........90.......100.......110.....
C
COMPASS
HsSATB1 KGTMLPVFCVVEHYEN---AIEYDCKE--------EHAEFVLVRKDMLFNQLIEMALLSLGYSHSSAAQAKGLIQVGKWNPVPLSYVTDAPDATVADMLQ
MmSATB1 KGTMLPVFCVVEHYEN---AIEYDCKE--------EHAEFVLVRKDMLFNQLIEMALLSLGYSHSSAAQAKGLIQVGKWNPVPLSYVTDAPDATVANMLQ
HsKIAA1034 -GLMIPVFCVVEQLDG---SLEYDNRE--------EHAEFVLVRKDVLFSQLVETALLALGYSHSSAAQAQGIIKLGRWNPLPLSYVTDAPDATVADMLQ
dDVE NPKSLPLHCVVESVHSLHASLTIDTRQPWKRRPNIETDSYVIIAAATPWSEIVQTALQRLGYSQEVANTARGSLIIKHWKPIPLEQISDNPAVPVSDIVG
CeZK1193.5 ---MFPMRVIVETVRS-QHCLTCSNEG------HMITDTYAIVAGTTTLNQLVETVLAALGHS-SMSTSARGLIQVNNWKPLPFDQITENLDDTVENLFK
1.......10........20........30........40........50........60........70........80........90.......100
HsSATB1 DVYHVVTLKIQLHSCPKLEDLPPEQWSHTTVRNALKDLLKDMNQSSL-AKECPLSQSMISSIVNSTYY-------ANVSAAKCQEFGRWYKHFKKTKD
MmSATB1 DVYHVVTLKIQLHSCPKLEDLPPEQWSHTTVRNALKDLLKDMNQSSL-AKECPLSQSMISSIVNSTYY-------ANVSAAKCQEFGRWYKHFKKTKD
HsKIAA1034 DVYHVVTLKIQLQSCSKLEDLPAEQWNHATVRNALKELLKEMNQSTL-AKECPLSQSMISSIVNSTYY-------ANVSATKCQEFGRWYKKYKKIKV
dDVE ELTSVITLRIVILR--------PKTSPFGEIKDKLLKLLVLQSHAVLRSTGCPLDEVTLSQICRSSHQNTYALPGGEISDELRRKFDQWWSNQLSPQA
CeZK1193.5 DISSHVVLKILSKPSTD----SNSVQCISEVKNKLLKAAVNKNPNVLTNVDNQQVKDVINTIIAGDET--------LLNSEQIGAVNEWLDTLETNED
.......110.......120.......130.......140.......150.......160.......170.......180.......190........

Fig. 3. Genomic structure of the *C. elegans* ONECUT gene cluster and sequence alignment of the OCAM motif and the COMPASS domain. (A) *Genomic structure of the homeobox gene cluster on* C. elegans *cosmid T26C11 consisting of the three ONECUT genes* ceh-39, ceh-21 *and* ceh-41. *The location of the cut domains, the homeodomains and a new motif, the OCAM motif is indicated. Based on our cDNA analysis, the 5' end of* ceh-21 *as shown here differs from the computer prediction that has another two 5' predicted exons.* **(B)** *OCAM motif: sequence alignment of the amino-terminal regions of* ceh-21, ceh-41 *and T02B5.2, a* C. elegans *gene that lacks both the cut- and homeodomain. The bold line indicates the OCAM motif.* **(C)** *Sequence alignment of the amino-terminal regions of the vertebrate SATB proteins and COMPASS family homeodomain proteins. The COMPASS domain is indicated with a bold line, smaller conserved regions are marked with a line.*

apparently been subject to rather fast evolution, since the branch lengths are long and the sequence similarity between the fly and *C. elegans* gene is lower than for other homeodomains (Fig. 1B). Searches with COMPASS genes in the databases did not reveal any other genes apart from the SATB genes that have a COMPASS domain. Given the conservation of the COMPASS genes with the SATB class genes through the COMPASS domain, it is possible that the invertebrate COMPASS genes are homologous to the vertebrate SATB genes. However, the COMPASS class homeodomains are rather different from SATB homeodomains in several aspects (Fig. 1B): 1) Extra residues between helix 1 and helix 2 (SATB) versus extra residues between helix 2 and helix 3 (COMPASS). 2) Phenylalanine (in SATB and ONECUT) instead of the standard tryptophan (COMPASS) at the highly conserved position 48 of the homeodomain (Bürglin, 1994). Given the unlikely event that the COMPASS domain associated two times

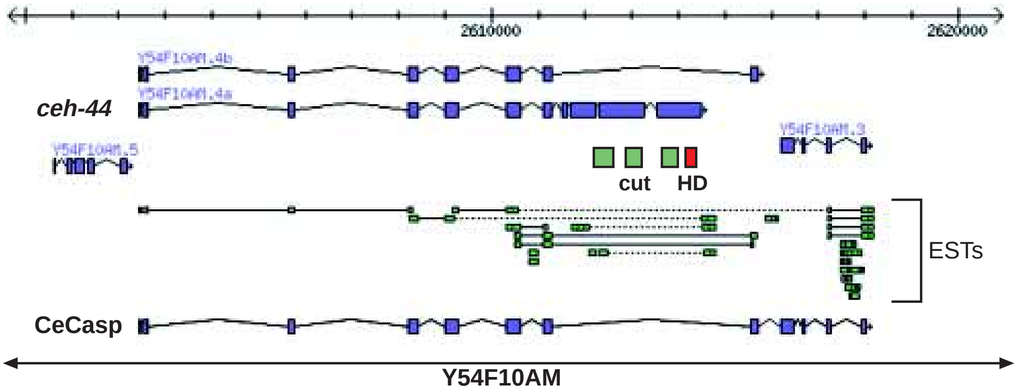


Fig. 4. Genomic structure of the CUX family gene *ceh-44*. *Genomic structure of ORFs on YAC clone Y54F10AM as taken from Wormbase (www.wormbase.org). Several ORFs (purple) are indicated under the sequence ruler; the two splice variants of Y54F10AM.4, a and b are shown. Y54F10AM.4a is the CUX family gene* ceh-44. *The cut domains and the homeodomain are indicated. The green lines under the ORFs indicate ESTs; ESTs linked by a dashed line are from the same cDNA. The CASP genes from other organisms align with Y54F10AM.4b in their amino-termini and with Y54F10AM.3 in their carboxyl-termini and ESTs link Y54F10AM.4b with Y54F10AM.3. The proper prediction for the* C. elegans *CASP gene is shown at the bottom; appropriate splice sites at the end of Y54F10AM.4b and the start of Y54F10AM.3 exist.*

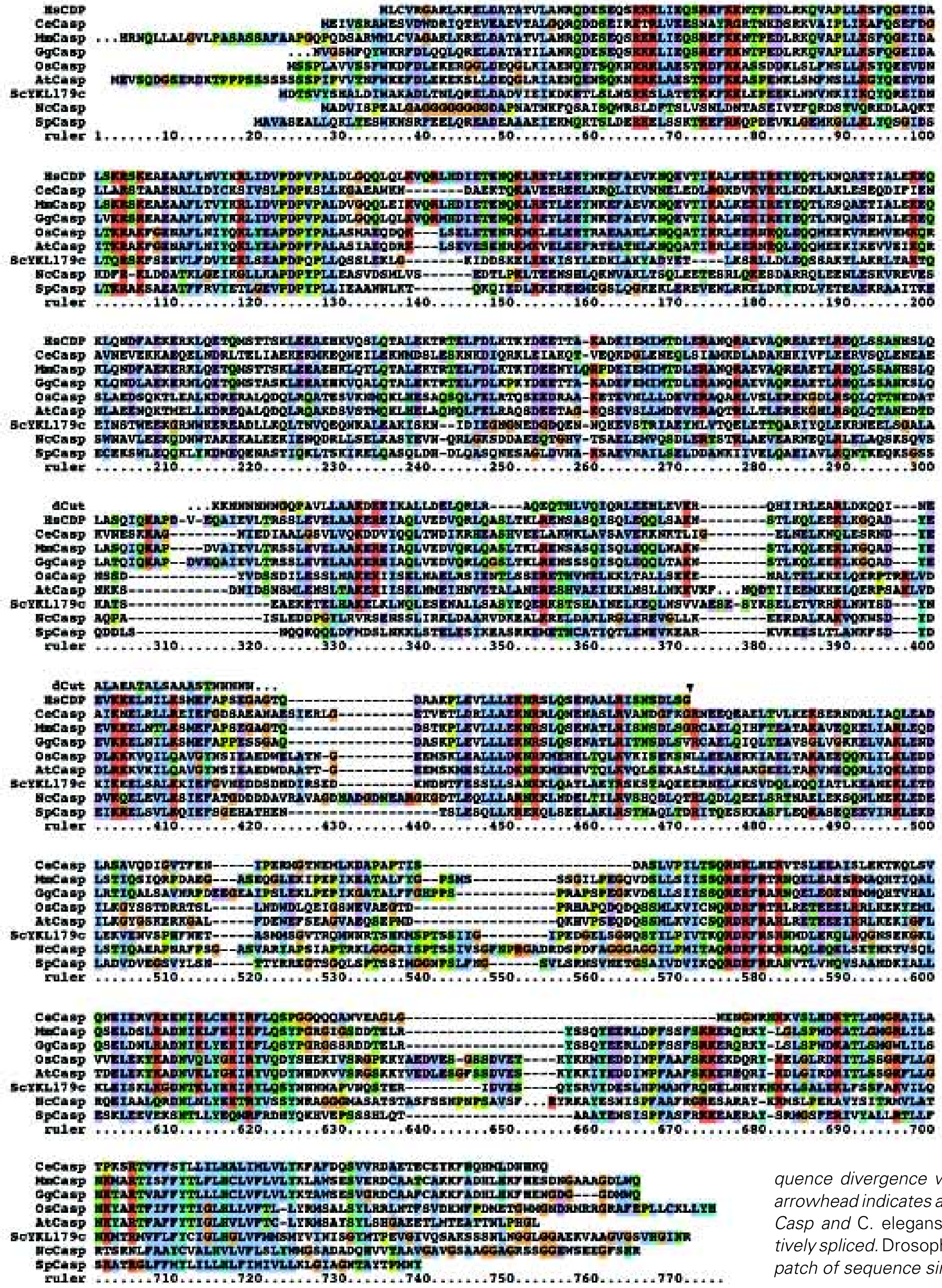


Fig. 5. Sequence alignment of CASP proteins. *The* C. elegans *CASP gene – as predicated from Fig. 4 - shows sequence similarity throughout its whole length to the alternative splice products of the CUX genes – called CASP - and to CASP proteins from plants and fungi. One vertebrate CUX protein, HsCDP, is shown with the amino-terminus lined up to the point of sequence divergence with the CASP proteins. The arrowhead indicates at which point the mouse Cux/ Casp and* C. elegans *ceh-44/CeCasp are alternatively spliced.* Drosophila *Cut has only a very limited patch of sequence similarity left.*

separately in evolution with a homeodomain, we favor the hypothesis that COMPASS genes are derived cut superclass genes and we propose to include the COMPASS family of genes in the cut superclass.

Alternative Splicing of CUX *Genes with* CASP *Genes*

Matrix dotplot comparison of *C. elegans* CEH-44 with vertebrate CUX genes revealed sequence conservation in the amino terminus. Examination of the genomic structure of *ceh-44* (Y54F10AM.4)

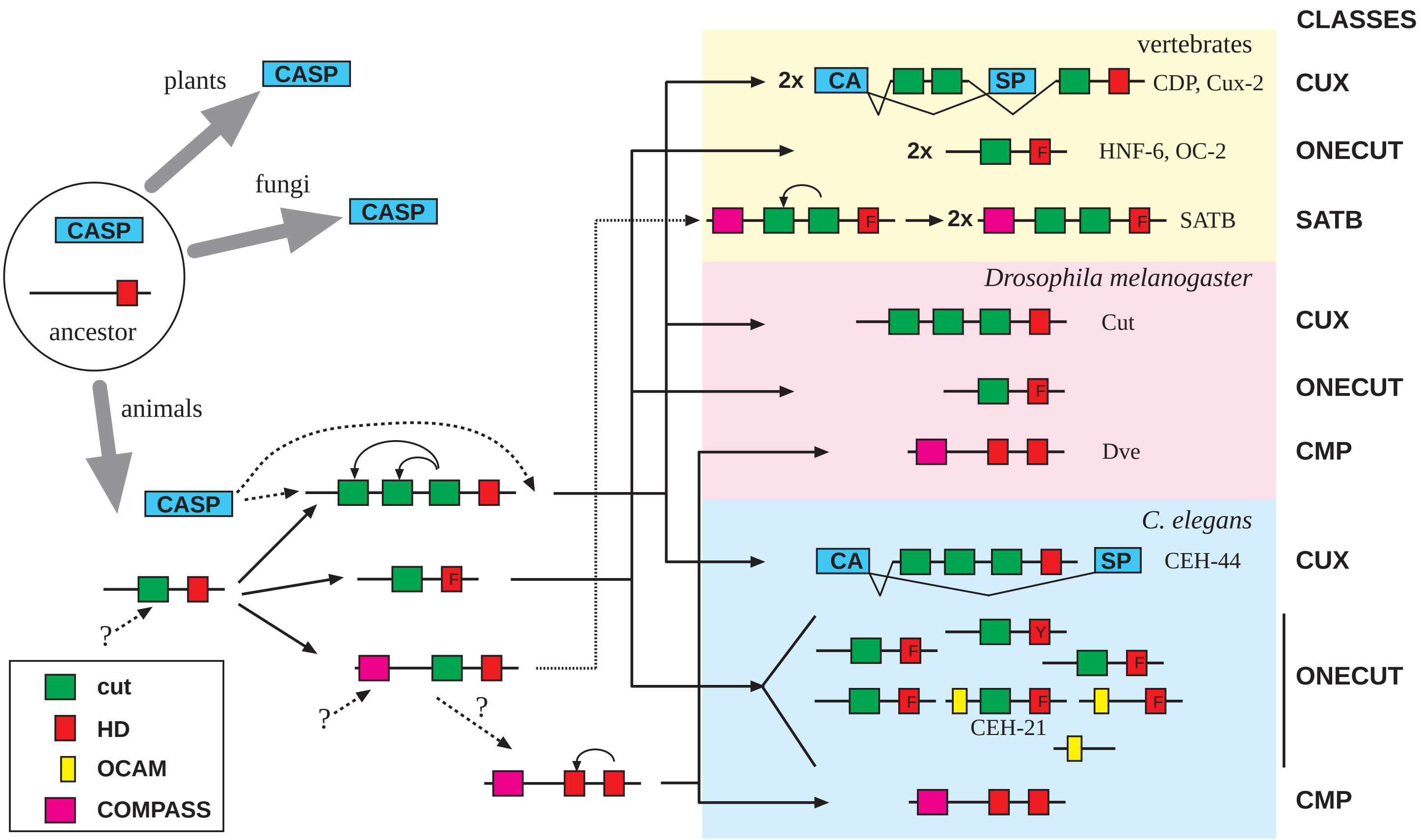


Fig. 6. Evolution of cut class homeobox genes. *Hypothetical model of the evolution of cut superclass homeobox genes from a common ancestor of plants, fungi and animals. The present day classes are shown on the right. The different domains are colour-coded, more uncertain events are indicated by dashed arrows, and question marks. The F and Y in the homeodomain indicates the residue at position 48 of the homeodomain.*

showed alternative splicing supported by EST data, which would give rise to two products, one lacking the cut domains and the homeodomain (Fig. 4). Blast searches with the amino-terminal region of CEH-44 uncovered not only the CUX genes in vertebrates, but also alternative splice forms which lack the cut domains and the homeodomain. These alternative splice forms of human and mouse CDP/Cux-1 have been named CASP (Lievens *et al*., 1997). Blast searches with a full-length CASP protein revealed that *C. elegans* ORF Y54F10AM.3 shows significant sequence similarity with the carboxy-terminus of vertebrate CASP genes (Fig. 5). We conclude that Y54F10AM.3 is the 3' end of the alternative splicing variant of Y54F10AM.4b and this proposed ORF is the *C. elegans* CASP gene. This is further supported by ESTs whose 5' ends lie at the beginning of Y54F10AM.4b and whose 3' ends terminate in Y54F10AM.3 (Fig. 4). A yeast gene with full-length similarity to mammalian CASP genes has previously been discovered (Lievens *et al*., 1997). Further Blast searches with CASP genes also revealed additional single copy genes from plants and fungi with significant sequence similarity throughout their whole length (Fig. 5). Thus the sequence similarity in the amino-terminal region of CUX genes, and between CASP genes is not simply due to conserved domains, but a gene, which was already present before the separation of plants, fungi and animals, is conserved in its entirety. We infer, as already proposed by Lievens *et al.* (1997) that the ancestor of the CUX class inserted into an intron of the CASP gene early in animal evolution and ever since, these genes are linked and produce both CASP and CUX products through alternative splicing. In *Drosophila* the CASP gene seems to have been lost, since the only remaining sequence similarity is confined to a small patch in the amino-terminal region (Fig. 5). Probably CASP gene function has been lost in flies, and the small region of sequence similarity may be the most crucial part of the CASP gene that is necessary for CUX function. The amino-terminal region also contains coil-coil regions with similarity to myosin and laminin, which may be of functional importance, since human CASP interacts with human CDP (Lievens *et al*., 1997; Tufarelli *et al*., 1998). Presently, nothing is known about CASP function in plants or fungi.

Evolution of Cut *Homeobox Genes*

Based on the above analysis, we propose the following hypothesis for the evolution of cut homeobox genes (Fig. 6). In the common ancestor of plants, fungi and animals there were at least two homeobox genes (see Bürglin, 1998) and a CASP gene. At some point in early animal evolution homeobox genes duplicated and one of these homeobox genes became the founding member of the cut superclass by obtaining a cut domain. Several possible scenarios are possible: 1) A domain was acquired by gene fusion from an unknown other gene (similar to the process of CASP/CUX merging). 2) The cut ancestor was derived from another homeobox gene that has associated domains (e.g., paired or POU) and the associated domain was subject to substantial modification so that it is not recognised as a e.g., paired or POU domain anymore. 3) The homeodomain duplicated (like in COMPASS genes) and one diverged substantially to give rise to the cut domain. 4) *De novo* generation of a functional, conserved sequence domain. Presently there is no evidence for any of these scenarios, however the structural prediction for the cut domain shows several alpha helixes which suggests that the cut domain may have been derived from an existing domain.

During further evolution of the founding cut gene, gene duplication must have occurred, since prior to the divergence of triploblastic animals at least two, probably three distinct cut genes existed. One was the ancestral CUX gene, which duplicated its cut domain twice and either before or after that duplication event inserted into an intron within the CASP gene. The second one was the ancestral ONECUT gene, which retained the structural characteristics of the founding cut homeobox gene, except for the change of a tryptophan residue to a phenylalanine at position 48 of the homeodomain. The third gene acquired a COMPASS domain (alternatively, the COMPASS domain could have been present in the primordial cut superclass gene, with subsequent loss in the CUX and ONECUT lineages). This COMPASS/CUT gene gave rise to the SATB genes in the vertebrate lineage, which involved a change of tryptophan to phenylalanine at position 48. It is not clear if the SATB genes are restricted to vertebrates, or if SATB genes were lost in other phyla. In a second branching, the COMPASS/CUT gene lost the cut domain and duplicated its homeodomain. This event could have happened early in the evolution of protostomes so that the COMPASS gene family, as found in flies and *C. elegans*, is directly orthologous to the SATB class and one would not have to propose the loss of SATB genes in invertebrates. Alternatively, COMPASS genes were lost in vertebrates. Presently, it is not possible to determine precisely how the COMPASS encoding genes evolved, though the more parsimonious hypothesis is that the COMPASS genes correspond to the SATB genes.

In conclusion, it seems that the cut genes have undergone more rearrangements and diversification than is usually observed in other homeobox gene classes. In particular the merging of two genes – CUX and CASP – through alternative splicing is a rather unusual event not seen in any other homeobox gene.

Experimental Protocols

Cloning and Sequencing

A *C. elegans* embryonic library (generous gift of Peter Okkema) was screened with the partial cDNA clone cm18e7 (Lannoy*et al*., 1998) to obtain full-length clones. Several clones were obtained, isolated, subcloned and sequenced according to standard methods (Sambrook *et al*., 1989; Cassata *et al*., 2000). The longest cDNA has been submitted to Genbank (Accession Nr AJ427855).

Sequence Analysis

Sequence database searches were carried out using the BLAST and PSI-BLAST Web servers (Altschul *et al*., 1990; Altschul *et al*., 1997; Schaffer *et al*., 2001) at NCBI (http://www.ncbi.nlm.nih.gov/BLAST/). Additional searches were performed at the Blast server of the Sanger Centre (http://www.sanger.ac.uk/Projects/C_elegans/). Sequences were downloaded, converted and compared with PPCMatrix, a dotmatrix program for the MacOS (Bürglin, 1998). Sequences were aligned using ClustalX 1.8 (Thompson *et al*., 1997) as well as by hand. Phylogenetic analyses were carried out using neighbour-joining as implemented in ClustalX 1.8, results were displayed using NJPlot by M. Gouy (http://biom3.univ-lyon1.fr/software/njplot.html). *C. elegans* genomic analysis was done at Wormbase (http://www.wormbase.org). For secondary structure prediction aligned protein sequences were submitted to the PredictProtein Server (http://dodo.cpmc.columbia.edu/predictprotein/) (Rost, 1996).

Acknowledgements

We thank Peter Okkema for the embryonic cDNA library and Patrick Dessi for critical reading of the manuscript. This work was supported by grants from the Swiss National Science Foundation. T.R.B. was recipient of a START Fellowship (NF. 3130-038786.93) and is now supported by the Swedish Foundation for Strategic Research (SSF).

References

ALTSCHUL, S.F., GISH, W., MILLER, W., MYERS, E.W. and LIPMAN, D.J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215: 403-410.

ALTSCHUL, S.F., MADDEN, T.L., SCHÄFFER, A.A., ZHANG, J., ZHANG, Z., MILLER, W. and LIPMAN, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* 25: 3389-3402.

ANDRES, V., NADAL-GINARD, B. and MAHDAVI, V. (1992). *Clox*, a mammalian homeobox gene related to *Drosophila cut*, encodes DNA-binding regulatory proteins differentially expresssed during development. *Development* 116: 321-332.

BLOCHLINGER, K., BODMER, R., JAN, L.Y. and JAN, N.J. (1990). Patterns of expression of Cut, a protein required for external sensory organ development in wild-type and *cut* mutant *Drosophila* embyros. *Genes Dev.* 4: 1322-1331.

BÜRGLIN, T.R. (1994). A comprehensive classification of homeobox genes. In *Guidebook to the Homeobox Genes* (Ed. Duboule, D.) Oxford University Press, Oxford, pp. 25-71.

BÜRGLIN, T.R. (1995). The evolution of homeobox genes. In *Biodiversity and Evolution* (Eds. Arai, R., Kato, M. and Doi, Y.) The National Science Museum Foundation, Tokyo, pp. 291-336.

BÜRGLIN, T.R. (1998). The PBC domain contains a MEINOX domain: Coevolution of Hox and TALE homeobox genes? *Dev. Genes. Evol.* 208: 113-116.

BÜRGLIN, T.R. (1998). PPCMatrix: a PowerPC dotmatrix program to compare large genomic sequences against protein sequences. *Bioinformatics* 14: 751-752.

CASSATA, G., KAGOSHIMA, H., ANDACHI, Y., KOHARA, Y., DÜRRENBERGER, M.B., HALL, D.H. and BÜRGLIN, T.R. (2000). The LIM homeobox gene *ceh-14* confers thermosensory function to the AFD neurons in *Caenorhabditis elegans. Neuron* 25: 587–597.

DUBOULE, D. (Ed.) (1994). *Guidebook to the Homeobox Genes*. Oxford University Press, Oxford.

FUSS, B. and HOCH, M. (1998). *Drosophila* endoderm development requires a novel homeobox gene which is a target of Wingless and Dpp signalling. *Mech. Dev.* 79: 83-97.

HARADA, R., BÉRUBÉ, G., TAMPLIN, O.J., DENIS-LAROSE, C. and NEPVEU, A. (1995). DNA-binding specificity of the Cut repeats from the human Cut-like protein. *Mol. Cell. Biol.* 15: 129-140.

JACQUEMIN, P., LANNOY, V.J., ROUSSEAU, G.G. and LEMAIGRE, F.P. (1999). OC-2, a novel mammalian member of the ONECUT class of homeodomain transcription factors whose function in liver partially overlaps with that of hepatocyte nuclear factor-6. *J. Biol. Chem.* 274: 2665-2671.

LANNOY, V.J., BÜRGLIN, T.R., ROUSSEAU, G.G. and LEMAIGRE, F.P. (1998). Isoforms of hepatocyte nuclear factor-6 differ in DNA-binding properties, contain a bifunctional homeodomain, and define the new ONECUT class of homeodomain proteins. *J. Biol. Chem.* 273: 13552-13562.

LEMAIGRE, F.P., DURVIAUX, S.M., TRUONG, O., LANNOY, V.J., HSUAN, J.J. and ROUSSEAU, G.G. (1996). Hepatocyte nuclear factor 6, a transcription factor that contains a novel type of homeodomain and a single *cut* domain. *Proc. Natl. Acad. Sci. USA* 93: 9460-9464.

LIEVENS, P.M., TUFARELLI, C., DONADY, J.J., STAGG, A. and NEUFELD, E.J. (1997). CASP, a novel, highly conserved alternative-splicing product of the CDP/*cut*/*cux* gene, lacks cut-repeat and homeo DNA-binding domains, and interacts with full-length CDP *in vitro*. *Gene* 197: 73-81.

NAKAGOMI, K., KOHWI, Y., DICKINSON, L.A. and KOHWI-SHIGEMATSU, T. (1994). A Novel DNA-binding motif in the Nuclear Matrix Attachment DNA-Binding protein SATB1. *Mol. Cell. Biol.* 14: 1852-1860.

NAKAGOSHI, H., HOSHI, M., NABESHIMA, Y. and MATSUZAKI, F. (1998). A novel homeobox gene mediates the Dpp signal to establish functional specificity within target cells. *Genes Dev.* 12: 2724-2734.

NEUFELD, E.J., SKALNIK, D.G., LIEVENS, P.M.-J. and ORKIN, S.H. (1992). Human CCAAT displacement protein is homologous to the *Drosophila* homeoprotein *cut*. *Nature Genet.* 1: 50-55.

QIAN, Y.Q., BILLETER, M., OTTING, G., MÜLLER, M., GEHRING, W.J. and WÜTHRICH, K. (1989). The structure of the Antennapedia homeodomain determined by NMR spectroscopy in solution: comparison with prokaryotic repressors. *Cell* 59: 573-580.

ROST, B. (1996). PHD: predicting one-dimensional protein structure by profile based neural networks. *Meth. in Enzym.* 266: 525-539.

RUDDLE, F.H., BARTELS, J.L., BENTLEY, K.L., KAPPEN, C., MURTHA, M.T. and PENDLETON, J.W. (1994). Evolution of *Hox* genes. *Annu. Rev. Genet.* 28: 423-442.

RUVKUN, G. and HOBERT, O. (1998). The taxonomy of developmental control in *Caenorhabditis elegans*. *Science* 282: 2033-2041.

SAMBROOK, J., FRITSCH, E.F. and MANIATIS, T. (1989). *Molecular Cloning. A laboratory manual.* Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.

SCHAFFER, A.A., ARAVIND, L., MADDEN, T.L., SHAVIRIN, S., SPOUGE, J.L., WOLF, Y.I., KOONIN, E.V. and ALTSCHUL, S.F. (2001). Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucl. Acids Res.* 29: 2994-3005.

THE *C. ELEGANS* SEQUENCING CONSORTIUM (1998). Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* 282: 2012-2018.

THOMPSON, J.D., GIBSON, T.J., PLEWNIAK, F., JEANMOUGIN, F. and HIGGINS, D.G. (1997). The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucl. Acids Res.* 25: 4876-4882.

TUFARELLI, C., FUJIWARA, Y., ZAPPULLA, D.C. and NEUFELD, E.J. (1998). Hair defects and pup loss in mice with targeted deletion of the first cut repeat domain of the Cux/CDP homeoprotein gene. *Dev. Biol.* 200: 69-81.

XU, W., ROULD, M.A., JUN, S., DESPLAN, C. and PABO, C.O. (1995). Crystal structure of a paired domain-DNA complex at 2.5Å resolution reveals structural basis for Pax developmental mutations. *Cell* 80: 639-650.