Mammalian Genome 11, 151–158 (2000).

Mammalian Genome
Incorporating Mouse Genome

Evolution of N-terminal sequences of the vertebrate HOXA13 protein

Douglas P. Mortlock,[1,*] Praveen Sateesh,[1] Jeffrey W. Innis[1,2]

[1]Department of Human Genetics, University of Michigan Medical School, Ann Arbor, Michigan 48104-0618, USA
[2]Department of Pediatrics, University of Michigan Medical School, Ann Arbor, Michigan 48109-0318, USA

Received: 8 June 1999 / Accepted: 23 September 1999

Abstract. While the the role of the homeodomain in HOX function has been evaluated extensively, little attention has been given to the non-homeodomain portions of the HOX proteins. To investigate the evolution of the HOXA13 protein and to identify conserved residues in the N-terminal region of the protein with potential functional significance, N-terminal *Hoxa13* coding sequences were PCR-amplified from fish, amphibian, reptile, chicken, and marsupial and eutherian mammal genomic DNA. Compared with fish HOXA13, the mammalian protein has increased in size by 35% primarily owing to the accumulation of alanine repeats and flanking segments rich in proline, glycine, or serine within the first 215 amino acids. Certain residues and amino acid motifs were strongly conserved, and several HOXA13 N-terminal domains were also shared in the paralogous HOXB13 and HOXD13 genes; however, other conserved regions appear to be unique to HOXA13. Two domains highly conserved in HOXA13 orthologs are shared with *Drosophila* AbdB and other vertebrate AbdB-like proteins. Marsupial and eutherian mammalian HOXA13 proteins have three large homopolymeric alanine repeats of 14, 12, and 17–18 residues that are absent in reptiles, birds, and fish. Thus, the repeats arose after the divergence of reptiles from the lineage that would give rise to the mammals. In contrast, other short homopolymeric alanine repeats in mammalian HOXA13 have remained virtually the same length, suggesting that forces driving or limiting repeat expansion are context dependent. Consecutive stretches of identical third-base usage in alanine codons within the large repeats were found, supporting replication slippage as a mechanism for their generation. However, numerous species-specific base substitutions affecting third-base alanine repeat codon positions were observed, particularly in the largest repeat. Therefore, if the large alanine repeats were present prior to eutherian mammal development as is suggested by the opossum data, then a dynamic process of recurring replication slippage and point mutation within alanine repeat codons must be considered to reconcile these observations. This model might also explain why the alanine repeats are flanked by proline, serine, and glycine-rich sequences, and it reveals a biological mechanism that promotes increases in protein size and, potentially, acquisition of new functions.

Introduction

HOX proteins are vertebrate transcription factors that are homologous to the *Drosophila* homeotic genes (Krumlauf 1994). The vertebrate HOX genes are important for assigning identity to cells along the major anterior-posterior axis of the embryo and promoting growth along the proximal-distal axes of the limbs and genital buds. In addition, adult expression in the hematopoietic system, roles in mammalian breast ductal epithelium differentiation during pregnancy and in hair follicle development, and deregulation in certain forms of cancer illustrate the critical role of HOX proteins in physiology (Chen and Capecchi 1999; Godwin and Capecchi 1998; Lawrence et al. 1996).

In mammals, there are four clusters of HOX genes, each thought to have arisen through duplication events from a single ancestral HOX cluster composed of 13 genes (Ruddle et al. 1994). The initial duplication of the HOX clusters, followed by selective gene loss within individual duplicate clusters, has led to the current arrangement of HOX genes in modern mammals. Extant HOX genes can be classified by homology as belonging to one of 13 ancestral paralogous groups (Scott 1992). Of these, HOX paralog groups 9 through 13 are most similar to the *Drosophila* Abdominal-B (*AbdB*) gene.

HOX proteins are composed of a 61-amino acid homeodomain (positioned near the C-terminus) and substantial N-terminal amino acid sequences. Binding and functional specificity is regulated by the homeodomain and by cofactor associations with the homeodomain and its N-terminal arm or other parts of the protein (Gehring et al. 1994; Hayashi and Scott 1990; Shen et al. 1997). The amino acid sequences of the homeodomains are highly conserved. Residues that are characteristic of paralog groups within and around the homeodomains have been identified (Sharkey et al. 1997). Interestingly, the non-homeodomain regions of HOX protein paralogs diverge greatly and may comprise as much as 80% of the protein.

Several studies have shown repressive effects of HOX proteins on transcription (Schnabel and Abate-Shen 1996). On the other hand, when non-homeodomain regions of HOX proteins are attached to heterologous DNA binding domains, transcriptional activation functions can be observed (Schnabel and Abate-Shen 1996; Vigano et al. 1998). Recently, Li and coworkers have shown that cofactor binding may release the transcriptional activation function(s) of the amino-terminal portion of *Drosophila* Dfd from the suppressive effects of the homeodomain (Li et al. 1999). Thus, even if DNA binding occurs, selective functional effects of HOX proteins may not be observed without cofactor-specific interactions that allow for N-terminal domain activities to be released. Our knowledge of the interacting proteins and specific biochemical reactions of the N-terminal domains of HOX proteins is minimal.

The discovery of spontaneous mutations in human and mouse HOX genes has drawn more attention to the role of the N-terminal domains in HOX functions. Alanine expansion mutations in HOXD13 have been described in mouse and human synpolydactyly (Akarsu et al. 1996; Goodman et al. 1997; Muragaki et al. 1996; Johnson et al. 1998), and one family with an alanine expansion in the HOXA13 protein causing hand-foot-genital syndrome has been identified (Goodman et al., personal communication).

Correspondence to: J.W. Innis

[*]*Present address:* Department of Developmental Biology, Beckman Center B300, 279 Campus Drive, Stanford, CA 94305-5329, USA.

The nucleotide sequence data reported in this paper have been submitted to GenBank and have been assigned the accession numbers AF083095–AF083104, and AF107227–AF107229.

Moreover, point mutations or deletions in the N-terminal nonhomeodomain region of HOXD13 give rise to unique dominant phenotypes (Goodman et al. 1999). Finally, the spontaneous mouse mutant *Hypodactyly* is due to a 50-bp deletion in the 5′ end of the first exon of *Hoxa13* (Mortlock et al. 1996). The *Hypodactyly* limb phenotype is more severe than that of the *Hoxa13* “knockout” mouse and results from increased cell death peculiar to *Hypodactyly* embryos (Post and Innis 1999). This mutation simultaneously results in elimination of wild-type HOXA13 protein and production of a novel protein in limb buds consisting of the first 25 amino acids of wild-type HOXA13 followed by 275 amino acids of arginine- and lysine-rich new sequence that lacks a homeodomain (Post, Margulies, Kuo, and Innis, in press). The *Hypodactyly* limb phenotype is similar to human hand-foot-genital syndrome resulting from a protein-truncating point mutation in the HOXA13 homeodomain (Mortlock and Innis 1997). Thus, Hox gene mutations affecting the N-terminal region may produce proteins with altered properties with or without a homeodomain. Indeed, homeodomain-independent functions of this group of transcription factors are well recognized, yet poorly understood (Copeland et al. 1996).

Hoxa13 is the cognate group 13 member of the Hoxa gene cluster (Scott 1992; Mortlock et al. 1996). *Hoxa13* is critical to the development of the autopod, reproductive and extraembryonic structures in mice and humans (Mortlock et al. 1996; Mortlock and Innis 1997; Fromental-Ramain et al. 1996; Warot et al. 1997). In mice, a 386-amino acid protein is produced by conceptual translation of the mRNA, and the amino-terminal nonhomeodomain portions of the mouse and human proteins have long, homopolymeric alanine repeats amidst conserved motifs (Mortlock et al. 1996). To learn more about the evolution of the HOXA13 amino-terminal non-homeodomain protein domains, we used the PCR to amplify the *Hoxa13* N-terminal coding regions from several mammalian and non-mammalian vertebrates and then compared the deduced amino acid sequences.

Materials and methods

Genomic DNAs. Rosy boa (*Charina trivirgata*) genomic DNA was prepared from a small amount of tail tissue from a captive-bred juvenile rosy boa, according to a standard method for preparing mouse genomic DNA (Mortlock et al. 1996). Chick genomic DNA was extracted from day 5 embryos by the same method. Genomic DNA from the lizards *Varanus dumerilli, Agama agama, Sceloporus undulatus*, and *Eumeces inexpectatus* and the snake *Typhlops richardi* were provided by Blair Hedges. Genomic DNA from a South American opossum (*Monodelphis domesticus*) was provided by Paul Samollow. Domestic cat and dog genomic DNAs were provided by Dave Kohrman. *Xenopus laevis* (amphibian) genomic DNA was prepared from a specimen provided by Rich Hume. Salmon (*Oncorhynchus* sp.) DNA was obtained from a commercial source. Stickleback (fish) (*Gasterosteus aculeatus*) genomic DNA was provided by Dae-gwon Ahn and Greg Gibson. Echidna (a monotreme, *Tachyglossus aculeatus*) genomic DNA was kindly provided by Stewart Nicol.

PCR primers. Various primer combinations designed from the murine *Hoxa13* sequence were tested empirically in attempts to amplify the N-terminal coding regions from genomic DNA of different vertebrates. The primer pairs included one of two “forward” primers, F1 (#8394D = 5′-CTATGACAGCCTCCGTGCTC-3′) or F2 (#9764D = 5′- ATCGAGCCCACCGTCATGTTTCTCTACGAC-3′), together with one of two “reverse” primers, R1 (#9765D = 5′-CGAGCTCTGTGCCGTCGCCGAGTAGGGACT-3′) or R2 (#9133D = 5′-TGGTAGAAAGCAAACTCCTTG-3′). These correspond to the following bases of the mouse *Hoxa13* Genbank file #U59322: F1, bases 353–372; F2, bases 388–417; R1, bases 814–843, reverse complement; R2, bases 1023–1043, reverse complement. After completing this study, we discovered that primer R1 has two thymines (underlined in the R1 primer sequence above) that correspond to cytosines in the mouse *Hoxa13* Genbank sequence. The PCR products described in this paper were amplified with the following primer pairs: cat and dog, primers F2 and R1; *Varanus dumerilli*, primers F2 and R2; all other species except fish and *Xenopus*, primers F1 and R2. For fish and *Xenopus* DNA amplification, reverse PCR primers #1131I (5′-AGGGGTAATAGCCGCTGCCGAAGT-3′) or #1132I (5′-GCCCGAGGTGTCCATGTACTTGTC-3′) were designed with *Xenopus Hoxa13* sequence information provided by T. Endo. These primers were used with forward primer F2.

PCR amplification of HOXA13 N-terminal sequences. 100–500 ng of genomic DNA was used for each PCR in 20- or 30-μl reactions. A mixture of 0.5 units *Taq* DNA polymerase and 0.06 units Vent DNA polymerase (New England Biolabs, Beverly, MA) per reaction was used, as this allowed more efficient amplification of specific products. The PCR buffer was as previously described (Mortlock et al. 1996), but with 100 μM each dATP, TTP, and dCTP; 62.5 μM dGTP; 37.5 μM 7-deaza-dGTP; 50 mM KCl; 5 mM NH_4Cl; 10% DMSO; and 3.3% glycerol. The cycling protocol was as follows: after incubation of the reaction at 100°C for 3 min, the reaction was cooled to 95°C and the Taq/Vent enzyme blend was added, followed by 35 cycles of 98°C 30 s, 15 s annealing, 75°C 3 min. In some cases, amplification was successful with Vent polymerase alone. Annealing temperatures were 52°C for primer pairs F1/R2 and F2/R2, and 56°C for F2/R1, except for the chick PCR with pair F1/R2, which used a 53°C annealing temperature. Most PCR products were sequenced after agarose gel purification. Some PCR products were re-amplified, cloned (T-vector, Promega, Madison, Wis.), and sequenced.

DNA sequencing and analysis. Sequencing was performed by the University of Michigan DNA Sequencing Core, or was performed manually with a Sequenase kit (USB). PCR products were sequenced in both directions. Sequence alignments were constructed with the MSA (Multiple Sequence Alignment) program (Lipman et al. 1989) accessed through the Baylor College of Medicine website (http://dot.imgen.bcm.tmc.edu:9331/multi-align/multi-align.html). Alignments were manually modified with the AssemblyLIGN and SeqPup programs. All computer sequence analysis was done with a Macintosh IIsi computer.

Alanine repeats BLASTP survey. The Swissprot database update posted April 25, 1998 was searched for identical matches to a query sequence of a perfect repeat of ten alanines, with the BLASTP program accessed through the NCBI website. The parameters were set as follows: V = 500, B = 50, H = 0, E = 1000, filter = none. This reported perfect matches as having a score of 40. Genpept reports of perfect matches were manually inspected to record alanine repeat numbers and lengths.

Results

PCR amplification of N-terminal Hoxa13 DNA sequences. Our previous amplification of human HOXA13 sequences with primers derived from the mouse genomic sequence suggested that this

Fig. 1. Alignment of HOXA13 ortholog amino-terminal sequences. Alignment of deduced amino-terminal amino acid sequences of the HOXA13 protein. Invariant amino acids are shown in red and summarized under A13c. Amino acids conserved among ≥75% of the available sequences are shown in blue. The alanine repeats are shown in green. Purple residues at amino acids 29, 117, and 216 of the mouse protein represent positions not meeting the 75% threshold but having two very similar residues of like charge or nature. For mouse and human, the sequence from the initiatior methionine to amino acid 297 is shown; for the remainder, the initiator methionine is not shown since the primers for PCR overlapped this genomic coding region. Only N-terminal region sequences are shown. Mus = mouse; hum = human; cat = cat; dog = dog; opm = opossum; aga = agamma; eum = eumeces; sce = sceloporus; var = varanus; typ = typhlops; lic = rosy boa; chk = chicken; xen = xenopus; stk = stickleback; sal = salmon; zeb = zebrafish. The zebrafish hoxa13b sequence from amino acid 1–201 was obtained from GenBank accession number AF071242. Sequence similarity becomes much greater after alanine repeat VII. Regions discussed in the text are labeled above the sequence. I–VII refer to variable length homopolymeric alanine stretches. (−) indicates gaps introduced for optimal alignment.

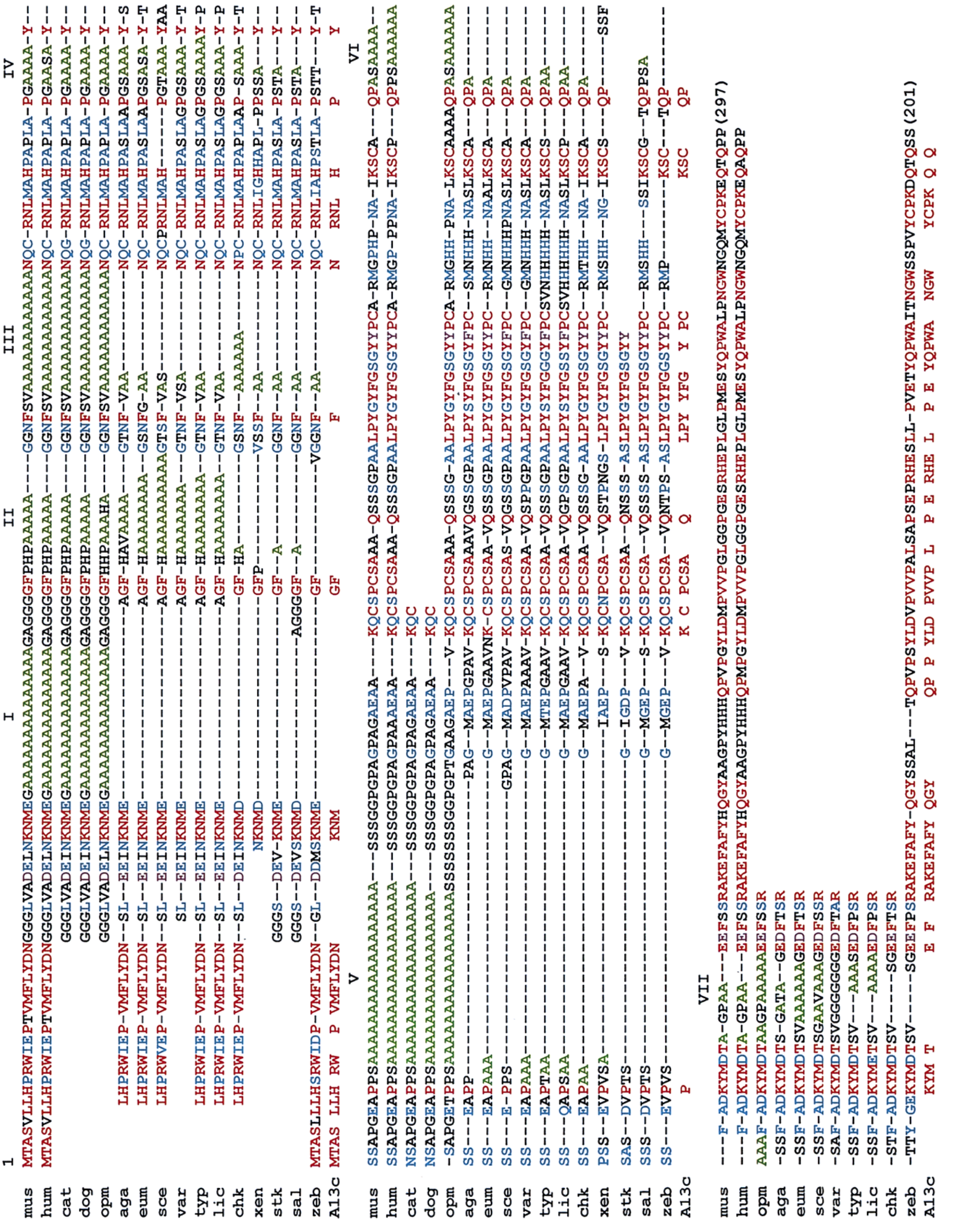

1
I
II
III
IV
V
VI
VII
mus
hum
cat
dog
opm
aga
eum
sce
var
typ
lic
chk
xen
stk
sal
zeb
A13c
(297)
(201)

Fig. 1.

approach could be used successfully on other species (Mortlock and Innis 1997). We obtained productive and specific amplifications from dog, cat, opossum, chick, five species of lizard, two species of snake, *Xenopus,* and two teleost fish. Agarose gel analysis revealed that the PCR products from teleost fish, *Xenopus,* reptiles, and chick were approximately 100–200 bp shorter than those of mouse or human when similar primer pairs were used (data not shown). The start site for translation of the mouse HOXA13 protein has recently been defined in vitro and in vivo and corresponds to the first methionine in the sequence shown for mouse HOXA13 in Fig. 1 (Post, Margulies, Kuo, and Innis, in press). Amplification of DNA other than mouse or human incorporated a forward primer that overlapped or was positioned slightly downstream of the initiator methionine as defined in the mouse protein. The conceptual translations of the newly amplified sequences revealed that they were true orthologs of HOXA13 as opposed to other HOX paralogs, suggesting that the bona fide *Hoxa13* gene was amplified from these organisms. The presence of multiple unique, reproducible nucleotide substitutions also identified the cloned products as originating from distinct species.

Alignment of N-terminal sequences from HOXA13 orthologs. Figure 1 shows the alignment of the HOXA13 N-terminal conceptual translations, along with the previously characterized mouse and human protein sequences. The shortest protein was from stickleback or zebrafish (teleost fish), and the longest was from opossum. Zebrafish (*Danio rerio*) HOXA13b was reported by Amores and associates (1998) and is presented at the bottom of the figure. Zebrafish have at least seven Hox clusters (Amores et al. 1998), including two that are believed to be orthologous to the mammalian *Hoxa* cluster; the zebrafish HOXA13b protein is clearly orthologous to mammalian HOXA13, although another ortholog may also be present in fish. The mammalian N-terminal regions are larger than the fish orthologs by approximately 70% if comparison is confined to the homologous sequences N-terminal to mouse HOXA13 amino acid 215. Overall, protein size has increased by 35% from fish to mammals. Several amino acids and short peptides were strongly conserved between all species examined (Red letters = 100% identity among available sequences; blue letters, ≥75% identity). These conserved residues are generally interrupted by more variable amino acid sequences consisting of either alanine repeats (green) or regions rich in alanine, glycine, proline, or serine, which are responsible for the increased size of the mammalian proteins. The extreme N-terminal region of HOXA13 from amino acids 1–36 (aa1–36) relative to the mouse, human, and zebrafish proteins was strongly conserved. After a long alanine repeat (I) present only in mammals, there are two phenylalanines separated by another alanine repeat (II), a N–RNL–H peptide motif followed by a proline-rich region and a single tyrosine. Except for two motifs each with an acidic residue (D or E) close to a proline, the next 45 amino acids are not strongly conserved. This region includes another long homopolymeric alanine repeat present only in mammals. This is followed by K-C-PCSA–Q, a highly aromatic peptide motif LPY-YFG–YY/FPC, and then KSC–QP; there is an interesting similarity between these two lysine-containing motifs. Between the aromatic motif and KSC–QP is a histidine-rich region present in fish, amphibians, reptiles, and birds, but lost in the mammals represented here. The KSC–QP conserved region is separated by another alanine repeat, present in mammals, from the F/YD/EKYMD/ET peptide sequence. Amino acids beyond the arginine at position 221 were compared only for the mouse, human, and zebrafish proteins.

The alanine repeats indicated as domains I, III, and V, previously known from the mouse and human sequences, were found to be conserved in the cat, dog, and the South American opossum. Repeats I and III were, comparatively, of the same length in these animals; however, repeat V is 18 alanines in mouse, human, and cat, yet 17 alanines in the dog and opossum. These long repeats were not conserved in the reptilian, chick, or fish sequences. Repeats I and V are absent from the fish, amphibian, reptile, and chick, with at most three consecutive alanines present in the corresponding locations. Repeat III is similarly reduced in length in the reptile sequences, while a repeat of six alanines is in this location in the chick. In reptiles, repeat II has 4–10 alanines, while teleost fish, *Xenopus,* and chick had 0–1 alanine. Interestingly, unlike repeats I, III, and V, repeats II, IV, and VII have not increased in size during evolution of the mammals. The opossum sequence also contains a repeat of nine alanines (VI) that is homologous to a repeat of 4 and 6 alanines in mouse and human, respectively. In addition, the opossum sequence contains a repeat of seven serines immediately C-terminal to alanine repeat V, whereas the eutherian mammals have a homologous repeat of only three serines. The amino acid sequences C-terminal to repeat V are rich in serine, glycine, proline, and alanine residues in the mammals only. Similarly, following alanine repeat I is a glycine-rich segment that, except for salmon, is present only in mammals.

Alignment of HOX group 13 paralogs. To identify paralog group 13-conserved residues, we aligned the mouse HOXA13 protein with the human HOXB13 and HOXD13 proteins (Fig. 2). The highly conserved region of all HOXA13 proteins including the first 36 amino acids is not identifiable in HOXB13 or HOXD13. The two phenylalanines separated by 11 amino acids in HOXA13 also are not present in B13; however, HOXD13 has two phenylalanine residues separated by five amino acids further C-terminal. The R-L-A motif of HOXA13 is conserved in both HOXB13 and HOXD13, with some degeneracy in position of the leucine and alanine residues in HOXD13 owing to the insertion of a phenylalanine. The histidine is similarly located in HOXB13, but is displaced in the C-terminal direction in HOXD13. HOXB13 has no alanine repeats longer than two codons, whereas HOXD13 has repeats of 5 and 15 alanines that are not clearly homologous with those of HOXA13.

Several other regions of conservation among the HOX group 13 proteins exist. This includes the K-C–P and YGY-FG–YY-C-R motifs. The latter is rich in aromatic residues, with five tyrosine or phenylalanines within a 9- to 10-amino acid stretch, suggesting this domain may provide hydrophobic side chains important for proper folding. An examination of the AbdB protein revealed that it also has an aromatic-rich peptide region immediately followed by charged residues, YGSGYYDRK, between the N-terminus and the homeodomain (Fig. 2). After this extended common motif, the paralogs share an identical lysine and a tyrosine residue surrounded by at least one acidic (D or E) amino acid. The frequency of identical residues increases among the paralogs as the comparison shifts toward the homeodomain in the C-terminus of the proteins.

MEIS-1 protein has been shown to interact with an N-terminal domain of HOXA9 (Shen et al. 1997). Since MEIS-1 also binds to other AbdB-like HOX proteins, we compared the invariant HOXA13 amino acids to conserved N-terminal regions of various HOX paralog group 9 proteins. A comparison of mouse and pufferfish HOXA9, mouse HOXB9, mouse and pufferfish HOXC9, and mouse HOXD9 revealed an identical YYVDS peptide motif, nine amino acids from the initiator methionine in each group 9 sequence. For simplicity, only mouse HOXA9 and HOXB9 are shown in Fig. 2. The FLYDN peptide sequence from the HOXA13 N-terminal domain and the YYVDS peptide of group 9 may, therefore, represent a conserved motif of three hydrophobic amino acids followed by aspartic acid and a polar uncharged amino acid. A similar peptide, VMYED, is present in the N-terminal region of *Drosophila* AbdB (Swissprot ID# P09087, amino acid positions 95–99).

In summary, several N-terminal amino acids or motifs are

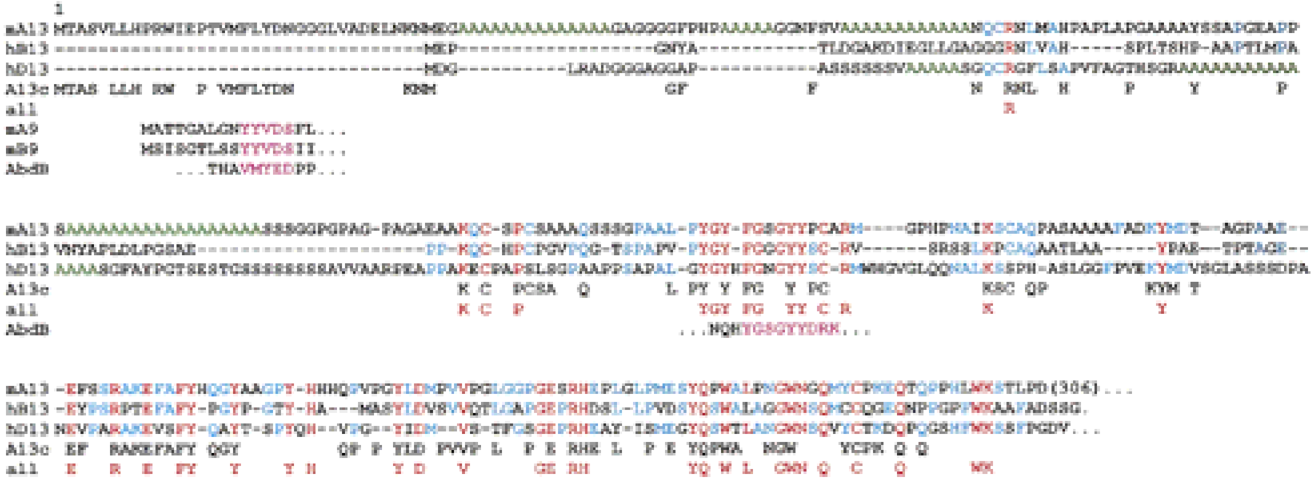


Fig. 2. Paralog group 13 amino-terminal alignment. The mouse HOXA13 sequence (mA13), human HOXB13 sequence (hB13), the human HOXD13 sequence (hD13) are aligned together. The invariant HOXA13 amino acids (A13c) are shown to indicate the distribution of ortholog conservation in contrast to the paralogs. Invariant residues are displayed in red (all). Residues in blue represent those in common with two out of the three paralogs, limited to A13 consensus regions. Alanine repeats are shown in green. The mouse HOXA9 (mA9) and the mouse HOXB9 (mB9) amino-terminal residues in common with the group 13 conserved residues are also shown, as are two amino-terminal amino acid motifs of the AbdominalB protein of *Drosophila* that are similar to those conserved in the mammalian group 13 genes. Significant motif homologies among these are pink.

shared among the known paralog group 13 genes that represent a subset of those highly conserved in the HOXA13 orthologs. HOXA13 shares a highly aromatic N-terminal domain with AbdB-like mouse proteins HOXA9 and HOXB9 and, apparently, AbdB; however, these are not shared with HOXB13 or HOXD13. Unless the single tyrosine at amino acid position 6 of HOXB13 represents this latter homology, it appears that AbdB-like HOX proteins have diverged significantly, sharing some residues and not others since gene duplication. A conserved amino acid array similar to the HOX paralog group 13 YGY-FG–YY-C-R motif also is present in AbdB. Peptide domains conserved among AbdB-like proteins probably represent common regulatory, structural, or protein interaction domains for the HOX group 13 paralogs. Differences in conserved amino acids between group 13 paralogs may represent divergent functional capabilities.

Evolution of HOXA13 alanine repeats. The overall size of the HOXA13 protein has increased substantially (35%) from fish to mammals. Almost all of the increase in length (approximately 100 amino acids) can be attributed to alanine-rich sequences or stretches rich in glycine, proline, or serine confined to the first 215 amino acids of the mammalian proteins. Three large homopolymeric alanine repeats coded by the first exon appeared before the divergence of marsupials from the ancestors of the eutherian mammals (Fig. 3). Since their divergence, the large alanine repeats of mammalian HOXA13 orthologs have remained virtually unchanged in length.

HOXA13 alanine expansions have been confined to certain regions of the protein during evolution. Repeats I, III, and V underwent substantial increases in length, while repeats II, IV, and VII remained the same or decreased in length. In addition, repeat V is variable in length among the mammals.

Close examination of the amino acids flanking the repeats reveals interesting differences. Alanine repeats I, III, and V vary in the number of flanking non-conserved proline-, serine-, or glycine-rich sequences. The number of flanking residues is greatest for repeat V, followed by repeat I and then repeat III. The distribution of flanking residues strongly favors the C-terminal region for repeats I and V, yet it is confined to two residues on the amino-terminal side of repeat III.

It is believed that homopolymeric amino acid repeats generally arise through “replication slippage” mechanisms, such that individual codons are duplicated in tandem, resulting in increased repeat length (Dover 1995; Primmer et al. 1996). Since the alanine codon is GCX, alanine repeats generated by replication slippage should initially have tandemly repeated codons with identical third-base nucleotides. In the third position nucleotides within the alanine repeat codons of the mammalian genes, evidence for replication slippage was seen in each of the HOXA13 alanine repeats. Repeat I is mostly comprised of GCG codons in the mouse, dog, and cat, with dog and cat each possessing a $(GCG)_8$ repeat. Repeat III of mouse has a $(GCC)_7$ repeat, whereas in dog repeat III has $(GCG)_{10}$. Repeat V has tandem GCC repeats of six or more codons in human, mouse, dog, and cat, but in opossum repeat III has a $(GCA)_8$ repeat at its 5′ end.

Relative to position within the large repeats, some codon third nucleotides are identical among the mammals. This similarity appears to be greatest for alanine repeat I (Fig. 4). The variable length of repeat V complicates the comparison. Perhaps more important, there is also significant alanine codon third nucleotide divergence, indicating that third position base substitutions followed by replication slippage events may have occurred independently within some mammalian groups after alanine repeats became established in an ancestral HOXA13 gene. Additional third-base substitutions appear to have occurred as well.

Despite the variability in alanine codon third-position nucleotides, the homopolymeric nature and length of alanine repeats I and III in the mammals tested are identical, whereas the length of repeat V is different. The occurrence of third-base substitutions without similar first- and second-base mutations may indicate that there is selection pressure for alanine repeats. However, such conservation may be, in part, a consequence of replication slippage.

Discussion

Conserved motifs within HOXA13 orthologs and between group 13 paralogs.

HOXA13 orthologs: Amino acid sequence alignment of the amino-terminal sequences of the HOXA13 orthologs revealed strong conservation of several motifs between teleost fish and

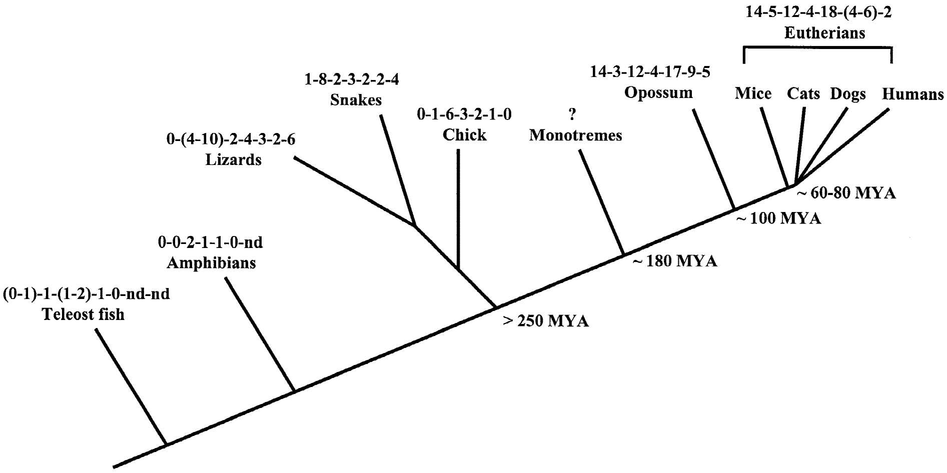
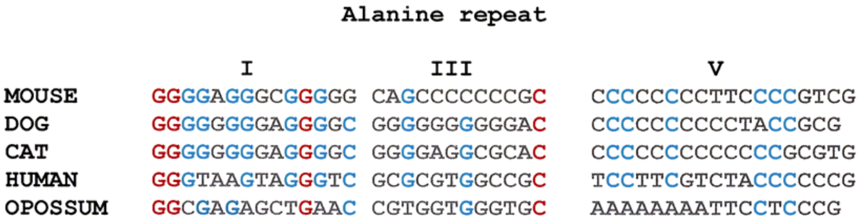


Fig. 3. Evolution of HOXA13 alanine repeats. The number of alanine codons in each repeat I–VII is shown relative to the evolutionary tree. Time since divergence is presented below the tree. nd = no data.



Alanine repeat

	I	III	V
MOUSE	GGGGAGGGCGGGGG	CAGCCCCCCCCGC	CCCCCCCCTTCCCCGTCG
DOG	GGGGGGGGAGGGGC	GGGGGGGGGGAC	CCCCCCCCCCTACCGCG
CAT	GGGGGGGGAGGGGC	GGGGAGGCGCAC	CCCCCCCCCCCCCGCGTG
HUMAN	GGGTAAGTAGGGTC	GCGCGTGGCCGC	TCCTTCGTCTACCCCCCG
OPOSSUM	GGCGAGAGCTGAAC	CGTGGTGGGTGC	AAAAAAAATTCCTCCCG

Fig. 4. Alanine repeat codon third position nucleotides. Shown are the third-base position nucleotides only, for the codons of the large HOXA13 alanine repeats in mammals. Repeats I, III, and V are 14, 12, and 18 alanines in length, respectively, in the mouse protein. Identical third-base nucleotides are in red; those positions with four identical bases out of five are shown in blue; gray nucleotides indicate fewer than four in common.

eutherian mammals, the most distant vertebrate comparison. These motifs are probably essential to ancient, conserved functions of this protein. The alanine repeat motifs of the marsupial and eutherian mammals are not shared with more distant vertebrates, suggesting either that they are nonessential or that their appearance was accompanied by newly acquired functions in mammals.

Group 13 paralogs: The most closely related genes to HOXA13 in vertebrates are the three other members of paralog group 13. Despite their related phylogeny, there is substantial variation in the amino-terminal sequences. Comparison with the HOXB13 and HOXD13 proteins revealed shared motifs that constitute a subset of those described for HOXA13 orthologs and probably were present before the duplication of the Hox clusters. The mouse HOXC13 protein shares similar motifs, including the extreme amino-terminal residues of HOXA13, but lacks the alanine repeats characteristic of HOXA13 (A. Godwin, personal communication). The mouse and human HOXB13 genes (Zeltser et al. 1996) lack alanine repeats altogether. The large alanine repeat of HOXD13 does not appear to be homologous to any of the repeats in HOXA13, considering the position of conserved residues between the two proteins. Therefore, the repeats in HOXA13 and HOXD13 probably arose independently after the duplication of the progenitor of these genes. The dramatic differences in the development of alanine repeats in these two transcription factors are remarkable given the involvement of both genes in limb and urogenital development and their ancestral relationship.

AbdB-like conservation: We assume that a very high degree of sequence conservation is present in the first 25 amino acids of all vertebrate HOXA13 proteins because of the success of upstream primers used in the PCR amplification. This conclusion is supported by the high degree of homology in this region between mammalian and zebrafish HOXA13 proteins. This N-terminal segment of HOXA13 protein shares aromatic residues with other AbdB-like proteins such as mouse HOXA9 and HOXB9, and AbdB itself. Interestingly, these residues are not present in the HOXB13 and HOXD13 proteins, indicating potentially divergent biological activities of these transcription factors. The YGYFGYY motif is an additional region of amino-terminal conservation between vertebrate HOXA13 proteins and *Drosophila* AbdB that is also shared with HOXD13 and HOXB13 and probably represents a very important functional domain.

Alanine repeats

HOXA13 alanine repeats: In marsupials and eutherian mammals, three homopolymeric alanine repeats of 12, 14, and 18 alanines exist in HOXA13. These large repeats are absent in the reptiles, birds, and teleost fish, indicating that the repeats appeared after the divergence of reptiles from the lineage that would give rise to the monotremes, marsupials, and eutherian mammals. The echidna, a monotreme or egg-laying mammal, diverged from the ancestors of the marsupial/eutherian lineage after the appearance of the earliest known mammals about 180 MYA (Fig. 3; Radinsky 1987). Repeated attempts to obtain *Hoxa13* PCR products from echidna genomic DNA were unsuccessful; however, this sequence information would be potentially very helpful in trying to determine the rate of appearance of the alanine repeats.

A similar homopolymeric amino acid expansion, with conservation of the repeats in mammals, has been documented for class III POU transcription factors (Nakachi et al. 1997). In addition, an alanine repeat in HOXD13 apparently also has undergone expansion through evolution (Muragaki et al. 1996). The human and mouse HOXD13 protein each have a homopolymeric repeat of 15 alanines in their N-terminal regions. This repeat is absent from the zebrafish *Hoxd13* gene (GenBank #X87752), whereas the chick HOXD13 protein has a 9-alanine repeat. It would be interesting, in relation to the work presented here, to determine when the HOXD13 alanine expansion occurred. It has been suggested that the HOXD13 alanine repeat might have arisen during the evolution of tetrapod limbs for a novel functional purpose (Muragaki et al. 1996). If this were true, the alanine repeats in HOXA13 might share similar functional characteristics with those of HOXD13, since these two factors are expressed in similar regions of the distal limb and genital bud during development (Zakany et al. 1997).

What determines where a homopolymeric alanine repeat will form? It is curious that alanine repeat expansions have developed in some regions and not others of the HOXA13 protein. Examination of the protein sequence of HOXA13 shows that repeats II, IV, and VII did not change appreciably in size, while the other three repeats lengthened considerably. This could mean that there is a functional purpose for increases in the size of certain repeats. Previous studies have shown that homopolymeric amino acid repeats are enriched in transcription factors of both *Drosophila* and mammals (Karlin and Burge 1996), and that alanine repeats or alanine-rich regions can function as repressor domains (Hanna-Rose and Hansen 1996). Therefore, some alanine repeat expansions may be favored because of a selective functional advantage.

On the other hand, alanine repeat length may be constrained because of deleterious effects on protein function above a certain limit. A BLASTP search of the Swissprot database to identify proteins containing at least one repeat of 10 or more alanines produced 72 proteins meeting those criteria. 55 (76%) were previously characterized transcription factors, of which 25 were non-orthologous vertebrate proteins. The longest perfect alanine repeat (21 amino acids) was found in the *Drosophila* Ovo protein, while large single repeats of 19 (human FKHL15), 18 (HOXA13), and 17 (mouse GSH-2) were also identified, each within transcription factors. From this search, it was revealed that alanine repeats are enriched in transcription factors, and there appears to be a natural limit to the length of uninterrupted alanine repeats, at approximately 20. In addition, expansions of alanine repeats are associated with human diseases including human and mouse synpolydactyly (HOXD13; Akarsu et al. 1996; Goodman et al. 1997; Muragaki et al. 1996; Johnson et al. 1998), cleidocranial dysplasia (*CBFA1;* Mundlos et al. 1997), oculopharyngeal muscular dystrophy (*PABP2;* Brais et al. 1998), and holoprosencephaly (*ZIC2;* Brown et al. 1998). These mutations are dominant and, at least for synpolydactyly, progressive increases in the alanine repeat length correlate with greater phenotypic severity (Goodman et al. 1997). Therefore, there are constraints on the length of alanine repeats reflected in the natural distribution of homopolymeric repeat length in proteins and by the occurrence of disease or malformation when selected repeats become too long.

The variability in third, but not first or second, base usage in codons within these large repeats supports the hypothesis that there is a selective advantage for the repeats to be composed of alanine residues. In vitro and in vivo systems capable of reporting HOXA13 transcription factor function could be used to explore the consequences of deletion, expansion, or substitution within the alanine repeats as well as the conserved residues of the protein. An alternative view based on the notion that the repeats do not provide an advantageous function is that they resulted from a propensity of the replication machinery to make slippage errors in GC-rich triplet codons. This view is supported by the existence of identical third-base nucleotides in tandemly repeated codons. Consequently, in this model repeats capable of expansion without interference in HOXA13 function have accumulated. Further expansion may result in genetic disease as revealed by the occurrence of hand-foot-genital syndrome as a result of an expansion of repeat V (F. Goodman, personal communication). It would be interesting to determine whether expansion of the other repeats would also lead to developmental abnormalities.

The flanking amino acid contexts of alanine repeats I and V are rich in serine, glycine, proline, and additional alanine residues and are very well conserved only in the mammals. This suggests that they arose in mammalian HOXA13 along with the alanine repeats. A dynamic mechanism incorporating both slippage and base substitution might reconcile this observation. If first and second nucleotide substitutions occurred within a progenitor repeat that changed alanine to proline, serine, or glycine and then replication slippage restored or expanded the length of the remaining alanine repeat, then the nonalanine codons would tend to be shifted gradually to either side of the repeat and the protein would increase in size. Under this model, the rates of expansion of these repeats may be variable secondary to underlying sequence characteristics or to effects on the function of HOXA13. In addition, alanine repeats with larger numbers of flanking serine, glycine, or proline may be sites that have a higher frequency of replication slippage or perhaps were the first to emerge in a HOXA13 progenitor.

Acknowledgments. The authors wish to thank Allan Force and John Postlethwait for generously sharing data regarding the zebrafish HOXA13 gene, Alan Godwin for the mouse *Hoxc13* sequence, and T. Endo for *Xenopus* genomic sequences. We also thank the following people for kindly providing genomic DNA samples: Blair Hedges (reptile samples), Paul Samollow (opossum), Stewart Nicol, and Dave Kohrman (domestic cat and dog). We thank Jani Lewis and Nick Plummer for providing chick embryos and a rosy boa tissue sample, respectively. D.P. Mortlock was supported by a postdoctoral fellowship from the University of Michigan Center for Organogenesis (Grant # 5-T32-HD0750502). This work was also partially funded by the University of Michigan Multipurpose Arthritis and Musculoskeletal Diseases Center (Center Grant 5 P60 AR20557) and NIH HD34057.

References

Akarsu AN, Stoilov I, Yilmaz E, Sayli BS, Sarfarazi M (1996) Genomic structure of HOXD13 gene: a nine polyalanine duplication causes synpolydactyly in two unrelated families. Hum Mol Genet 5, 945–952

Amores A, Force A, Yan Y-L, Wang Y-L, Fritz A et al. (1998) Zebrafish hox clusters and vertebrate genome evolution. Science 282, 1711–1714

Brais B, Bouchard J-P, Xie Y-G, Rochefort DL, Chretien N et al. (1998) Short GCG expansions in the PABP2 gene cause oculopharyngeal muscular dystrophy. Nat Genet 18, 164–167

Brown SA, Warburton D, Brown LY, Yu C-y, Roeder ER et al. (1998) Holoprosencephaly due to mutations in *ZIC2*, a homologue of *Drosophila odd-paired.* Nat Genet 20, 180–183

Chen F, Capecchi MR (1999) Paralogous mouse *Hox* genes, *Hoxa9, Hoxb9, Hoxd9,* function together to control development of the mammary gland in response to pregnancy. Proc Natl Acad Sci USA 96, 541–546

Copeland JWR, Nasiadka A, Dietrich BH, Krause HM (1996) Patterning of the *Drosophila* embryo by a homeodomain-deleted Ftz polypeptide. Nature 379, 162–165

Dover G (1995) Slippery DNA runs on and on . . . Nat Genet 10, 254–256

Fromental-Ramain C, Warot X, Messadecq N, LeMeur M, Dolle P et al. (1996) *Hoxa-13* and *Hoxd-13* play a crucial role in the patterning of the limb autopod. Development 122, 2997–3011

Gehring WJ, Qian YQ, Billeter M, Furukubo-Tokunaga K, Schier AF et al. (1994) Homeodomain-DNA recognition. Cell 78, 211–223

Godwin AR, Capecchi MR (1998) *Hoxc13* mutant mice lack external hair. Genes Dev 12, 11–20

Goodman FR, Mundlos S, Muragaki Y, Donnai D, Giovannucci-Uzielli ML et al. (1997) Synpolydactyly phenotypes correlate with size of expansions in HOXD13 polyalanine tract. Proc Natl Acad Sci USA 94, 7458–7463

Goodman F, Giovannucci-Uzielli ML, Hall C, Reardon W, Winter R et al. (1999) Deletions in *HOXD13* segregate with an identical, novel foot malformation in two unrelated families. Am J Hum Genet 63, 992–1000

Hanna-Rose W, Hansen W (1996) Active repression mechanisms of eukaryotic transcription repressors. Trends Genet 12, 229–234

Hayashi S, Scott MP (1990) What determines the specificity of action of *Drosophila* homeodomain proteins? Cell 63, 883–894

Johnson KR, Sweet HO, Donahue LR, Ward-Bailey P, Bronson RT et al. (1998) A new spontaneous mouse mutation of *Hoxd13* with a polyalanine expansion and phenotype similar to human synpolydactyly. Hum Mol Genet 7, 1033–1038

Karlin S, Burge C (1996) Trinucleotide repeats and long homopeptides in genes and proteins associated with nervous system disease and development. Proc Natl Acad Sci USA 93, 1560–1565

Krumlauf R (1994) Hox genes in vertebrate development. Cell 78, 191–201

Lawrence HJ, Sauvageau G, Humphries RK, Largman C (1996) The role of HOX genes in normal and leukemic hematopoiesis. Stem Cells 14, 281–290

Li X, Murre C, McGinnis W (1999) Activity regulation of a Hox protein and a role for the homeodomain in inhibiting transcriptional activation. EMBO J 18, 198–211

Lipman DA, Altschul SF, Kececioglu JD (1989) A tool for multiple sequence alignment. Proc Natl Acad Sci USA 86, 4412–4415

Mortlock DP, Innis JW (1997) Mutation of HOXA13 in hand-foot-genital syndrome. Nat Genet 15, 179–180

Mortlock DP, Post LC, Innis JW (1996) The molecular basis of Hypodactyly (*Hd*): a deletion in *Hoxa13* leads to arrest of digital arch formation. Nat Genet 13, 284–289

Mundlos S, Otto F, Mundlos C, Mulliken JB, Aylsworth AS et al. (1997) Mutations involving the transcription factor CBFA1 cause cleidocranial dysplasia. Cell 89, 773–779

Muragaki Y, Mundlos S, Upton J, Olsen BR (1996) Altered growth and branching patterns in synpolydactyly caused by mutations in HOXD13. Science 272, 548–551

Nakachi Y, Hayakawa T, Oota H, Sumiyama K, Wang L, Ueda S (1997) Nucleotide compositional constraints on genomes generate alanine-, glycine-, and proline-rich structures in transcription factors. Mol Biol Evol 14, 1042–1049

Post LC, Margulies EH, Kuo A, Innis JW Severe limb defects in Hypodactyly mice result from the expression of a novel, mutant HOXA13 protein. Dev Biol, in press

Post LC, Innis JW (1999) Altered *Hox* expression and increased cell death distinguish *Hypodactyly* from *Hoxa13* null mice. Int J Dev Biol 43, 287–294

Primmer CR, Ellgren H, Saino N, Moller AP (1996) Directional evolution in germline microsatellite mutations. Nat Genet 13, 391–393

Radinsky LB (1987) *The Evolution of Vertebrate Design.* (Chicago: University of Chicago Press)

Ruddle FH, Bartels JL, Bentley KL, Kappen C, Murtha MT et al. (1994) Evolution of *HOX* genes. Annu Rev Genet 28, 423–442

Schnabel CA, Abate-Shen C (1996) Repression by HoxA7 is mediated by the homeodomain and the modulatory action of its N-terminal arm residues. Mol Cell Biol 16, 2678–2688

Scott MP (1992) Vertebrate homeobox gene nomenclature. Cell 71, 551–553

Sharkey M, Graba Y, Scott MP (1997) *Hox* genes in evolution: protein surfaces and paralog groups. Trends Genet 13, 145–151

Shen W-F, Montgomery JC, Rozenfeld S, Moskow JJ, Lawrence HJ et al. (1997) AbdB-like Hox proteins stabilize DNA binding by the Meisl homeodomain proteins. Mol Cell Biol 17, 6448–6458

Vigano MA, Di Rocco G, Zappavigna V, Mavilio F (1998) Definition of the transcriptional activation domains of three human HOX proteins depends on the DNA-binding context. Mol Cell Biol 18, 6201–6212

Warot X, Fromental-Ramain C, Fraulob V, Chambon P, Dolle P (1997) Gene dosage-dependent effects of the *Hoxa-13* and *Hoxd-13* mutations on morphogenesis of the terminal parts of the digestive and urogenital tracts. Development 124, 4781–4791

Zakany J, Fromental-Ramain C, Warot X, Duboule D (1997) Regulation of number and size of digits by posterior *Hox* genes: a dose-dependent mechanism with potential evolutionary implications. Proc Natl Acad Sci USA 94, 13695–13700

Zeltser L, Desplan C, Heintz N (1996) *Hoxb13:* a new Hox gene in a distant region of the HOXB cluster maintains colinearity. Development 122, 2475–2484