Organization of Mouse *Iroquois* Homeobox Genes in Two Clusters Suggests a Conserved Regulation and Function in Vertebrate Development

Thomas Peters,[1,2] Renate Dildrop,[1,2] Katrin Ausmeier,[1,3] and Ulrich Rüther[1,4]

[1]*Entwicklungs- und Molekularbiologie der Tiere, Heinrich-Heine-Universität, 40225 Düsseldorf, Germany*

Iroquois proteins comprise a conserved family of homeodomain-containing transcription factors involved in patterning and regionalization of embryonic tissues in both vertebrates and invertebrates. Earlier studies identified four murine *Iroquois* (*Irx*) genes. Here we report the isolation of two additional members of the murine gene family, *Irx*5 and *Irx*6. Phylogenetic analysis of the *Irx* gene family revealed distinct clades for fly and vertebrate genes, and vertebrate members themselves were classified into three pairs of cognate genes. Mapping of the murine *Irx* genes identified two gene clusters located on mouse chromosomes 8 and 13, respectively. Each gene cluster is represented by three *Irx* genes whose relative positions within both clusters are strictly conserved. Combined results from phylogenetic, linkage, and physical mapping studies provide evidence for the evolution of two *Irx* gene clusters by duplication of a larger chromosomal region and dispersion to two chromosomal locations. The maintenance of two cognate *Irx* gene clusters during vertebrate evolution suggests that their genomic organization is important for the regulation, expression, and function of *Irx* genes during embryonic development.

[The sequence data in this paper have been submitted to the EMBL Nucleotide Sequence Database under accession nos. AJ271053, AJ271054, AJ271055.]

A major force in evolution is clearly the duplication of genes and their subsequent diversification. The duplicated copies constitute well-defined gene families whose members may be clustered together or dispersed on different chromosomes (or a combination of both). Almost all genes (including those that encode transcription factors) are members of families, such as the zinc finger, bHLH, or homeobox genes. With respect to how these families are organized in the genome, dispersion of related members throughout different locations appears to be the more common situation, rather than clustering. For example, zinc finger genes, although existing as multiple discrete subfamilies, have, as yet, never been found in a clustered organization. Similarly, only a few of the genes that encode bHLH transcription factors [e.g., *E(spl)* and *achaete-scute* gene complexes in *Drosophila*] are clustered (Alonso and Cabrera 1988; Knust et al. 1992).

Our knowledge about gene families that are organized in clusters and that have been conserved in evolution is limited. The best-characterized example of a conserved gene cluster is presented by the *Hox* genes, a discrete group of homeobox genes. Here, ~10 genes are duplicated in *cis*, and four such *Hox* clusters have been characterized in mammals. The single ancestral cluster that was common to flies and mammals became duplicated along the evolution of vertebrates (reviewed in Finnerty and Martindale 1998).

The organization of genes in clusters raises the question of the extent to which such genes are under common control. In particular, how does the expression of a gene depend on its context (in *cis*), and what are the consequences of cluster duplication when regulatory elements might be duplicated as well? In principle, the expression profiles of duplicated genes (or genes within duplicated clusters) may be identical or similar to their ancestors unless evolution has selected new expression identities. Indeed, a correlation between the clustered organization of transcription factors with their expression pattern has been shown, in particular for those genes encoded in the four *Hox* clusters mentioned above (reviewed in Krumlauf 1994).

Analysis of the *Drosophila Iroquois* (*Irx*) mutation led to the identification of three tightly linked, highly related homeodomain proteins that are encoded by the *Iroquois*-complex (*Iro-C*) genes *araucan* (*ara*), *caupolican* (*caup*), and *mirror* (*mirr*) (Gómez-Skarmeta et al. 1996; McNeill et al. 1997). During *Drosophila* development these genes are involved in sensory organ formation in the lateral domain of the notum (Leyns et al. 1996; Grillenzoni et al. 1998; Kehl et al. 1998) as well as in the specification of body-wall and wing identity

[2]These authors contributed equally to this paper.
[3]Present address: Institut für Molekularbiologie, MHH, 30625 Hannover, Germany.
[4]Corresponding author.
E-MAIL ruether@uni-duesseldorf.de; FAX 211-811-5113.
Article and publication are at www.genome.org/cgi/doi/10.1101/gr.144100.

(Gómez-Skarmeta et al. 1996; Diez del Corral et al. 1999). In addition, *Iro-C* genes were shown to act as dorsal selector genes in the *Drosophila* eye (Cavodeassi et al. 1999). Recently, multiple homologs of *Iro-C* genes have been isolated from distantly related vertebrate groups such as fish, amphibia, chicken, mouse, and man (Bosse et al. 1997; Bellefroid et al. 1998; Gómez-Skarmeta et al. 1998; Bao et al. 1999; Funayama et al. 1999; Goriely et al. 1999; Lewis et al. 1999; Tan et al. 1999; Bruneau et al. 2000). Functional analyses of vertebrate *Irx* genes are only beginning to appear. As in *Drosophila*, *Xenopus Irx* genes control the expression of proneural genes, whereas the chicken *Irx4* gene regulates the chamber-specific expression of myosin isoforms in the developing heart (Bellefroid et al. 1998; Gómez-Skarmeta et al. 1998; Bao et al. 1999). Four *Irx*-related genes, *Irx1, Irx2, Irx3,* and *Irx4,* have so far been described in the mouse (Bosse et al. 1997; Bruneau et al. 2000). Because of their distinct spatiotemporal expression during embryogenesis, these genes are believed to be involved in processes such as patterning of the central nervous system and regionalization of the otic vesicle, branchial epithelium, heart, and limbs (Bosse et al. 1997; Bruneau et al. 2000).

Here we report the identification of two additional members of the murine *Irx* gene family. Phylogenetic analyses of these in the context of their vertebrate homologs and physical mapping of murine *Irx* genes provide a picture of evolution of the vertebrate gene family in which an ancient three-gene cluster was duplicated and dispersed to different chromosomal locations.

RESULTS

Identification of Two Novel Members of the Murine *Irx* Gene Family

The *Irx* gene family of homeobox-transcription factors was originally described in *Drosophila*, consisting of three clustered genes, *ara*, *caup*, and *mirr* (Gómez-Skarmeta et al. 1996; McNeill et al. 1997). Vertebrate homologs of these have been isolated from several species (Bosse et al. 1997; Bellefroid et al. 1998; Gómez-Skarmeta et al. 1998; Bao et al. 1999; Funayama et al. 1999; Goriely et al. 1999; Lewis et al. 1999; Tan et al. 1999; Bruneau et al. 2000). Until now, five cDNAs have been identified in humans (*IRX1, IRX2, IRX3, IRX4, IRX5*) and four cDNAs have been identified in mice (*Irx1, Irx2, Irx3, Irx4*). This prompted us to perform a database search to identify additional murine *Irx* homologs. We initially identified a set of four overlapping mouse expressed sequence tags (ESTs) (representative EST entryW54596) whose cDNA sequences were similar to the human *IRX2* sequence, and a single mouse EST (AA709522) that is similar to the human *IRX3* sequence.

EST W54596

The conceptual translation of mouse EST W54596 yielded two short open reading frames of 63 and 55 amino acids in length that were similar (94% identity) to the C-terminal part of the human *IRX2* coding sequence translation. The cDNA clone corresponding to EST W54596 was available through the I.M.A.G.E. Consortium (Lennon et al. 1996) and the DNA sequence of the entire insert was determined, yielding a single open reading frame of 410 codons (data not shown). The analyzed clone did not represent a full-length cDNA copy and because no cDNA clones extending into the 5′ direction were available, we isolated the 5′ end of the transcribed sequence from genomic DNA. The correct linkage of sequences from genomic and from cDNA origin in the composite full-length sequence thereby generated was verified by reverse transcriptase-polymerase chain reaction (RT-PCR) and subsequent sequencing of the amplified product (data not shown). The composite cDNA sequence (1592 bp) contained an ATG start codon at position 25, a 1449-bp open reading frame, and a stop codon at position 1474. The derived amino acid sequence yielded a putative protein of 483 amino acids that was most similar (90% identity among 401 overlapping amino acids) to the protein translation of the human *IRX2* database entry (U90304). We have named the mouse cDNA *Irx*5 (GenBank accession: AJ271053, AJ271054).

The N-terminal 82 amino acids of the murine Irx5 protein did not match to the N-terminal 17 amino acids of the human *IRX2* translation, but did well match to other published *Irx* protein sequences. We noticed that the position at which the two sequences start to diverge coincides with an exon–intron boundary in the genomic sequence of the mouse *Irx5* gene (data not shown), which may indicate that the N-terminal 17 amino acids encoded by the human *IRX2* cDNA are either intron derived or else that they result from an alternatively spliced exon. The murine *Irx5* gene was independently isolated by two other groups (Bosse et al. 2000; Cohen et al. 2000)

EST AA709522

The conceptual translation of the cDNA represented by mouse EST AA709522 yielded a single open reading frame of 33 amino acids with homology (66% identity) with the translation of the N-terminal region encoded by human cDNA *IRX3*. The mouse cDNA clone was available through the I.M.A.G.E. Consortium (Lennon et al. 1996) and the DNA sequence of the entire insert was determined. The insert represented a 1750-bp cDNA copy encoding a single, large open reading frame that started at position 11 and terminated at position 1324. The translational product of the encoded open reading frame yielded a protein of 438

amino acids that was most similar (83% identity in 192 amino acids overlap) to the partial protein sequence encoded by human cDNA clone *IRX3* (U90305). The mouse gene represented by EST AA709522 has been named *Irx6* (GenBank accession: AJ271055). The sequence of a recently described murine *Irx* gene, also termed *Irx6* (Cohen et al. 2000), corresponds to the previously published sequence of mouse *Irx2* (Bosse et al. 1997).

The *Irx* gene family

Our initial database search and subsequent sequencing resulted in the identification of cDNAs representing two additional murine *Irx*-related genes (*Irx5* and *Irx6*), which, together with the four mouse cDNAs described previously, constitute the murine family of *Irx* genes. We next performed exhaustive database searches to identify orthologs of the six murine "*Irx* gene" family members across all vertebrate species. Representative database entries (given as GenBank accession numbers) that cover the entire set of entries found are summarized in Figure 1. We subdivided these sequence entries into six separate groups according to the amino acid similarities of their respective coding sequences in comparison to the six protein sequences identified in mice. By this we were able to identify one additional human gene (group 2, assembly: R50645-AI613079-AI654035-AI831283), which, until now, has not been described. Moreover, we were also able to generate sequence assemblies extending partial coding sequences for several human and rodent genes, for which only limited sequence information (mainly homeodomain) existed previously. Finally, two sets of database entries were identified from *Fugu*, encoding homologous sequences of the murine genes *Irx5* (group 5, assembly: AL003277-AL006325-AL003293-AL003273) and *Irx3* (group 3, assembly: AL028548, AL028584).

To simplify the comparison of closely related coding sequences from different species we refer to these sequences according to their group designation. By this criterion, sequence assemblies representing human genes *IRX1* (group 3), *IRX2* (group 5), *IRX3* (group 6), and *IRX5* (group 1) are now referred to as *IRX3, IRX5, IRX6,* and *IRX1,* respectively.

The alignment of protein sequences from 23 vertebrate *Irx* genes, the three *Drosophila* genes described previously, and the single homolog identified in *Caenorhabditis elegans* is shown in Figure 2. Figure 2A displays part of the N-terminal region and the homeodomain. The 64 residues containing the homeodomain (positions 189–252) are highly conserved among the various vertebrate proteins (>89% identity), as are the three *Drosophila* proteins when compared to the set of vertebrate proteins (>91% identity). The homeodomain derived from *C. elegans* gene *C36F7.1* is considerably less conserved (73%–77% identity) when compared to fly and vertebrate members. The overall conservation of amino acid residues in the N-terminal domain of the Irx proteins (positions 1–188) was found to be highest for vertebrate sequences assigned to the same group (averaged value for sequences within a given group >82% identity; range, 71%–94%; data not shown). When comparing N-terminal domains of vertebrate and fly amino acid sequences, conservation

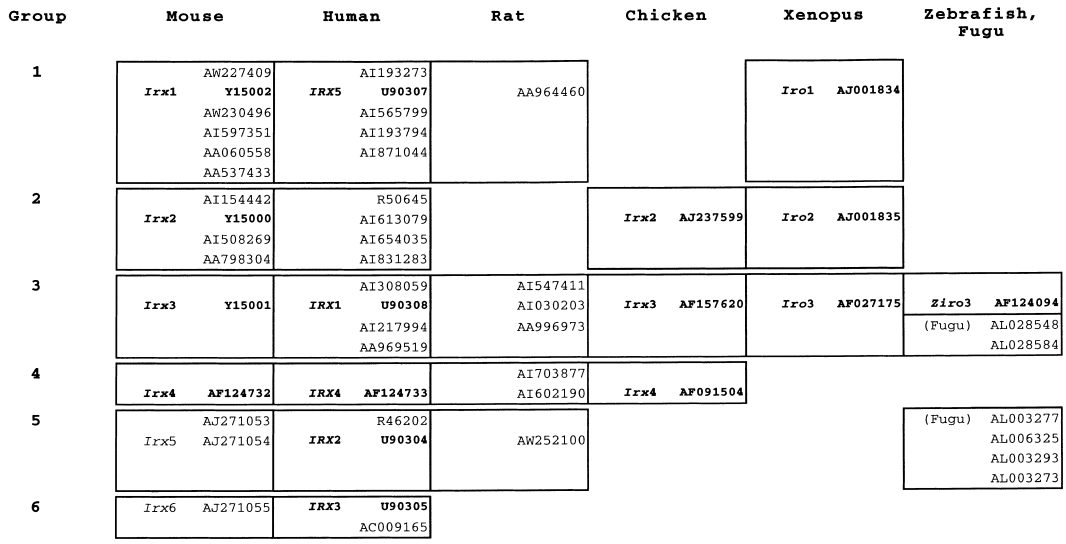

Figure 1 Vertebrate coding sequences (GenBank accession nos) retrieved from the database define six groups of closely related *Iroquois* genes. Sequence assemblies of partial cDNAs (or of exon-containing genomic DNA) representing the same gene are boxed together. Database entries and names of published cDNAs are shown in boldface, all other entries are derived from database searches described in this article. Sequence entries for homologous genes present in *Drosophila* and *Caenorhabditis elegans* are: *caupolican* (X95178), *araucan* (X95179), *mirror* (U95021), and *C36F7.1* (Z81045).

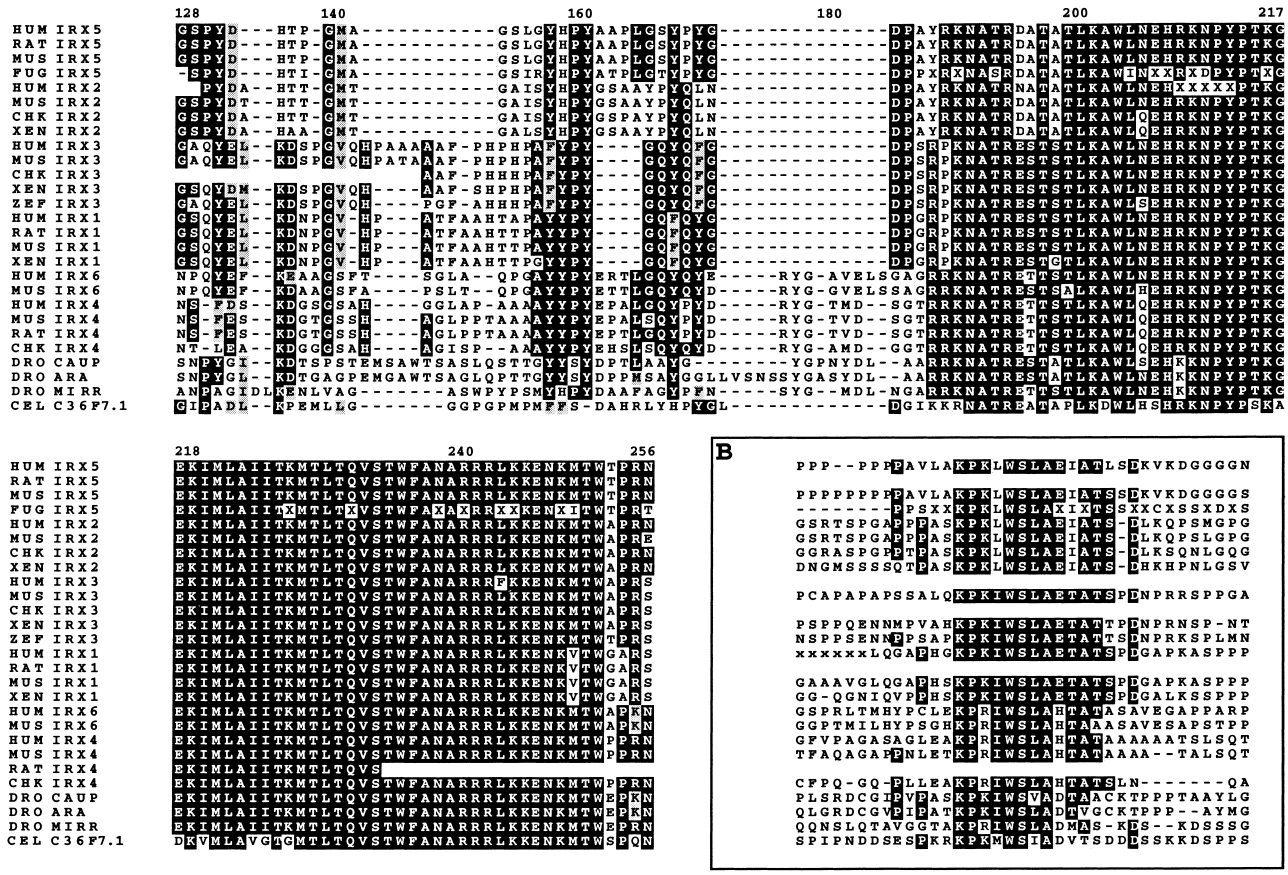

Figure 2 Alignment of vertebrate and invertebrate *Iroquois* protein sequences. Amino acid sequences were deduced from DNA sequences as summarized in Fig. 1, and designation of sequences is according to grouping of sequences therein. (*A*) Alignment of part of the N-terminal domain and the homeodomain (alignment positions 189–252). (*B*) Alignment of part of the C-terminal region encoding the conserved IRO box. Amino acids that are identical in more than half of the sequences are indicated by a solid box, and similar amino acids are indicated by a shaded box. 'X' denotes positions where the corresponding amino acids could not be identified from the DNA sequences; '−' denotes positions where gaps have been introduced into the alignment; 'x' denotes gaps in the actual sequence assemblies. The complete alignment containing the entire N-terminal region and the homeodomain (1–256) is available from the authors on request.

ranges between 20% and; 33% identity (41%–73% when only *Drosophila* sequences are considered). The overall sequence identity of both fly and vertebrate sequences compared to that of the nematode (*C36F7.1*) was <20%.

Within the protein region located toward the C-terminal from the homeodomain (data not shown) the conservation of amino acid sequences is considerably lower than it is in the N-terminal part (>56% identity for vertebrate sequences within a given group, range, 40%–70%). There is, however, a small protein domain of 13 amino acids that was previously identified as IRO box (Bürglin 1997), which has been conserved in all sequences analyzed including sequences of both vertebrate and invertebrate origin (Figure 2B).

Molecular Phylogeny of *Irx* genes

Phylogenetic analyses were performed to assess the evolutionary relationships among the *Irx* gene family by defining orthologous genes (derived from a speciation event) as well as paralogous genes (derived from a duplication event). Gene trees were constructed using the neighbor-joining method (Saitou and Nei 1987) including representative genes from the six vertebrate groups, the three genes from *Drosophila* and the single gene present in *C. elegans*. The nematode gene (*C36F7.1*) was treated as an outgroup to provide a root for the analyses. The necessary assumption that this gene is outside of the insect–vertebrate group was based on homeodomain amino acid identities (see above). The gene tree shown in Figure 3 was constructed from the homeodomain along with part of the N-terminal region (positions 128–256 of the alignment shown in Fig. 2A), covering the region for which the largest set of sequence entries had been retrieved by us from the GenBank database.

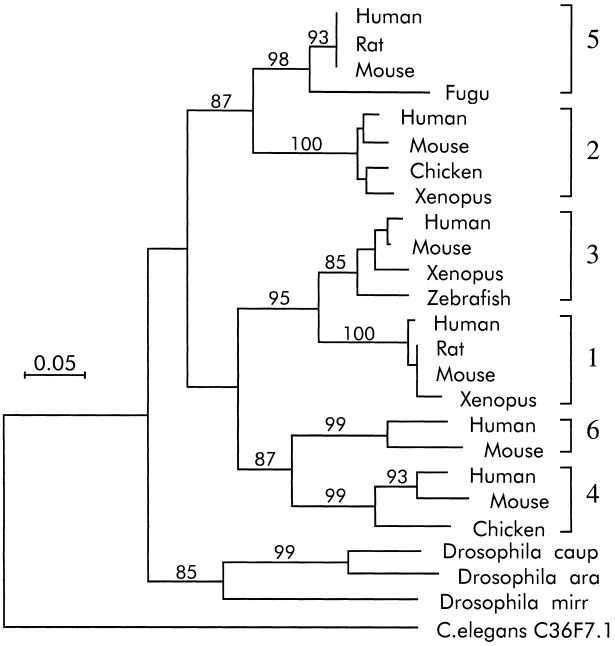

Figure 3 Phylogenetic analysis of *Iroquois* family members. The neighbor-joining tree was calculated from Kimura distances, with scale bars indicating the percentage (x 1/100) of sequence divergence. Numbers above branches indicate the number of times (%) the branch was found in 200 bootstrap replicates. Bootstrap values at short external branches (<2% divergence) are not indicated. The gene tree was inferred from the amino acid alignment in Fig. 2*A* (positions 128–256), analyzing a set of 21 vertebrate genes, 3 genes from *Drosophila* and the *Caenorhabditis elegans* gene *C36F7.1*, which was used as the outgroup. Incomplete sequences (chicken *Irx3*, rat *Irx4*) have not been considered. Brackets at the right highlight the clustering of vertebrate genes into six orthologous groups.

The *Irx* gene phylogeny reveals several notable aspects. First, the six previously defined groups of vertebrate genes form distinct, monophyletic clades in the phylogenetic analysis, indicating that the corresponding genes represent orthologs that evolved from a single ancestral gene. Each of these groups contains a human and a rodent member. In addition to the mammalian homologs, five of these groups contain genes from more distant vertebrate species. Second, the three *Drosophila* genes, *caup*, *ara*, and *mirr* cluster outside the vertebrate *Irx* genes, forming a separate group. The data most probably indicate that the duplications that led to the three genes in *Drosophila* and the duplications that led to the six genes identified in vertebrates were independent events. Third, in addition to the six orthologous groups, the vertebrate genes fall into three larger clades, one of which is characterized by *Irx5* and *Irx2,* another one by *Irx3* and *Irx1,* and the third one by *Irx6* and *Irx4.* The highly related genes within each of these larger clades have evolved apart from a common ancestral gene through a duplication event, thus forming cognate (paralogous) gene pairs.

The same clustering of vertebrate genes (six orthologous groups and three cognate groups), and the clustering of the three *Drosophila* genes into a separate group was obtained when phylogenetic analyses of representative full-length (or nearly full-length) sequences (positions 1–256) were performed (data not shown).

Arrangement of Mouse *Irx* Genes in Two Gene Clusters

To investigate the physical location of *Irx* genes in the mouse genome, we took advantage of the fact that a YAC-based physical map covering ~92% of the mouse genome was recently described by the Whitehead Institute/Massachusetts Institute of Technology Center for Genome Research (WI/MIT CGR; Nusbaum et al. 1999). Thus polymerase chain reaction (PCR) screening of the same YAC library (Haldi et al. 1996) used for physical map construction allowed identification of *Irx*-positive YACs assigned to the integrated genetic and physical map (WI/MIT CGR database).

Our search for genomic clones containing murine *Irx* genes led to the identification of several YAC clones positive for *Irx1, Irx2,* and *Irx4* (Fig. 4A). The finding that *Irx1* and *Irx2* both map to YAC clone 87-C-5 already revealed a tight linkage of these genes. Consultation of the WI/MIT CGR database provided detailed mapping information for YAC clones 87-C-5, 357-B-3 and 396-A-2, which are all located in YAC contig WC-321 at centimorgan position 25 on chromosome 13. In addition, database information for contig WC-321 indicated a physical overlap of YAC clones 87-C-5 and 357-B-3 confirming the tight linkage of *Irx1* and *Irx2.* Most importantly, this database information demonstrated a physical overlap of YAC clones 357-B-3 and 396-A-2 connecting *Irx2* and *Irx4,* respectively. Thus *Irx1, Irx2,* and *Irx4* are tightly linked in this linear order on mouse chromosome 13.

Similarly, YAC clones positive for *Irx3, Irx5,* and *Irx6* were identified (Fig. 4B). Among those, YAC clones 116-F-4, 352-H-11, 388-A-10, and 419-B-6 have been considered by the WI/MIT CGR database and were found to map to YAC contigs WC-1037 and WC-1038, at centimorgan position 43 on chromosome 8. The tight linkage of *Irx5* and *Irx6* was uncovered by analysis of YAC clones 270-C-10 and 388-A-10 containing both *Irx* genes. Likewise, the tight linkage of *Irx3* and *Irx5* was shown by the mapping of both genes to YAC clone 352-H-11. In addition, we assayed for the presence of genomic markers known to be located in this region (Fig. 4B), independently confirming the clustered organization of these genes. Thus *Irx3, Irx5,* and *Irx6* are tightly linked in this linear order on mouse chromosome 8.

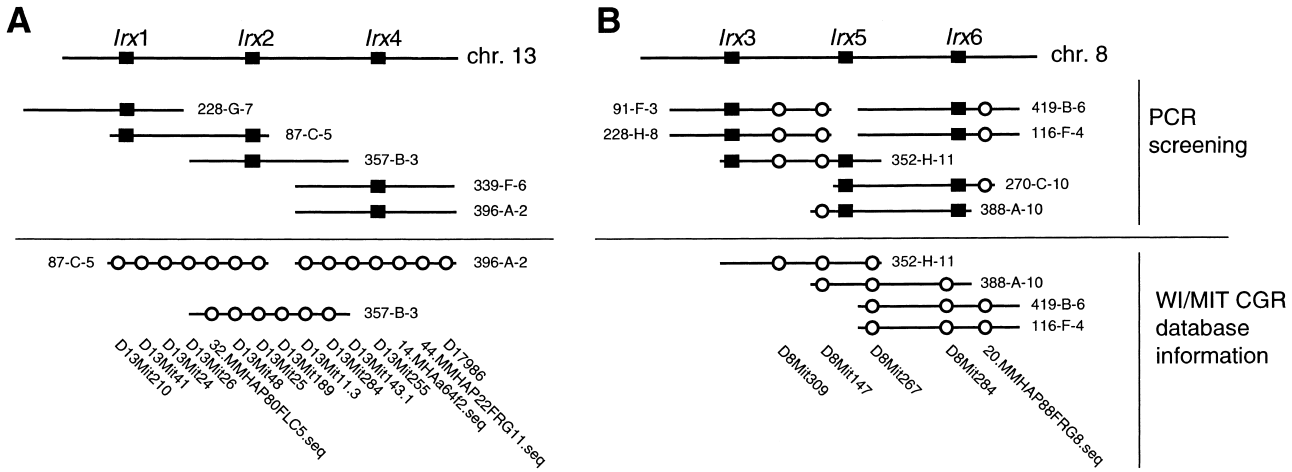

Figure 4 Schematic representation of the physical linkage of murine *Irx* genes in two clusters on mouse chromosomes 13 (*A*) and 8 (*B*). The top half of the figure shows the generation of YAC contigs based on polymerase-chain-reaction (PCR)-screening of the large MIT mouse YAC library (WI/MIT 820). Squares indicate the relative positions of murine *Irx* genes. Single YAC clones are represented as horizontal lines. The name of each clone is indicated. The bottom half represents database information for individual YAC clones provided by the Whitehead Institute/MIT Center for Genome Research (WI/MIT CGR). Underneath individual genomic marker (sequence-tagged sites; open circles) assigned to YAC clones considered by WI/MIT CGR database are shown. Genomic markers D8Mit147, D8Mit309 and 20.MMHAP88FRG8.seq were also tested by PCR. Details on each marker/YAC clone are available on the World Wide Web site at the WI/MIT CGR (http://carbon.wi.mit.edu:8000/cgi-bin/mouse/index).

Iroquois Gene Cluster Duplication By a Large-Scale Chromosomal Duplication

Our linkage data revealed a clustered organization of *Irx* genes in the mouse genome. Interestingly, each of the three genes linked in one of the two identified *Irx* clusters could be assigned to one of the three cognate gene pairs defined by our phylogenetic analysis. The relative position of cognate genes within each of the two clusters is strikingly conserved. This arrangement is intriguing in that it suggests that the two clusters are the result of a single duplication event, revealed by our phylogenetic analysis to have occurred before the divergence of the major vertebrate lineages.

To test whether the chromosomal locations of *Irx* genes were conserved between vertebrate species, we screened the National Center for Biotechnology Information (NCBI) GeneMap '99 for mapping information on human *IRX* genes. These database analyses revealed that human *IRX3* (stSG10226) is positioned on chromosome 16q12, whereas (cognate) *IRX1* (D5S678, stSG51099) was found to map to chromosome 5p15.3. Both chromosomal locations correspond to the location of the two mouse gene clusters on chromosomes 8 (43cM) and 13 (25cM). In addition, BAC clones containing the *IRX6* gene were identified by screening the NCBI database (HTGS division). These BAC clones also contain genomic marker D16S3032, which maps to chromosomal region 16q12 on the sequence-tagged site (STS)-based map of the human genome (WI/MIT CGR; Dib et al. 1996). Human *IRX4* (cognate with *IRX6*) has recently been mapped to chromosome 5p15.3 (Bruneau et al. 2000). Although genomic locations of human *IRX2* and *IRX5* have not been determined so far, it is likely that the clustered organization of *IRX* genes is maintained in humans.

If duplication and dispersion of an ancestral *Irx* cluster occurred en bloc, other genes closely linked to the original *Irx* cluster should have been duplicated as well. To investigate this, we screened the NCBI database (Online Mendelian Inheritance in Man (OMIM), GeneMap 1999) for additional pairs of paralogous genes located in vicinity of human *IRX* clusters. Analysis of the *IRX1*/*IRX4* region on chromosome 5p15.3 identified several genes (e.g., *ADCY2, SLC6A3*) whose paralogs (e.g., *ADCY7, SLC6A2*) were found to map nearby *IRX3*/*IRX6* on chromosome 16q12–q13 (data not shown). The genomic distribution pattern of these genes is consistent with the view that the ancient duplication entailed a large chromosomal region containing additional genes.

DISCUSSION

Irx Genes Are Members of an Ancient Gene Family

Our database search for *Irx*-related genes led to the identification of two additional members of this family of transcriptional regulators, mouse *Irx5* and mouse *Irx6,* which, along with the four previously identified mouse genes, constitute the murine *Irx* gene family. In addition, we have identified one new member (now termed *IRX2*) from the human family, which, together with the five previously identified human genes, also consists of at least six individual genes. At present, homologs from five of the six mammalian family mem-

bers have been identified in other vertebrate groups, including genes from amphibia, birds, and fish. Together with the three *IRO*-C genes previously identified in *Drosophila* and the single homolog identified in *C. elegans*, *Irx* genes form a well-conserved, ancient gene family that can be traced back to a metazoan ancestor that is common to vertebrates, flies, and nematodes. This is revealed by homeodomain amino acid identities (>73%) within the *Irx* gene family, as compared to identities of <56% to members of other homeodomain gene families present in the database (data not shown). In addition, all members of the *Irx* family are characterized by the highly conserved IRO box, which was found to be unique for this family of homeodomain proteins (Bürglin 1997).

Evolution of Vertebrate *Irx* Genes by Cluster Duplication

Our phylogenetic analyses demonstrated that vertebrate *Irx* genes cluster into six orthologous groups. In addition, each two of these orthologous groups form a set of cognate gene pairs. Thus we conclude that *Irx1* and *Irx3* evolved apart from a common ancestral gene through a duplication event, as well as *Irx2*/*Irx5* and *Irx4*/*Irx6,* which represent a second and a third pair of cognate genes. Because our physical mapping analyses revealed the linkage of *Irx3, Irx5,* and *Irx6,* and the linkage of *Irx1, Irx2,* and *Irx4* (discussed below), the data suggest that the three duplications (generating pairs of cognate genes) are the result of a single event. The timing of that duplication event can be estimated to have occurred in a common ancestor of the vertebrate lineage, before the divergence of fish, amphibia, birds, and mammals, most probably in the course of a genome-wide duplication that is believed to have occurred early in the evolution of vertebrates (reviewed in Skrabanek and Wolfe 1998; Aparicio 2000). Exemplified with the case of the clustered organization of the *Hox* genes (mammals possess four *Hox* complexes), at least two such rounds of genome duplication probably coincided with the appearance of vertebrates. This raises the possibility that previous genome duplications resulted in additional *Irx* three-gene clusters that were subsequently lost (Ruddle et al. 1994; Aparicio 2000). Possibly, future investigations of *Irx* gene organization in agnathans or elasmobranchs will shed light on this topic.

Considering the relationships of vertebrate and invertebrate *Irx* genes, the phylogenetic analyses showed that the three *Drosophila* genes cluster outside the major vertebrate clade. The *Irx* genes in the fly have been shown to be physically linked (Gómez-Skarmeta et al. 1996; McNeill et al. 1997; Netter et al. 1998), thus the three genes most likely have evolved by two local duplication events. The very same scenario can be envisioned for the evolution of the single, primordial three-gene cluster that is ancestral to the two three-gene clusters that are present in vertebrates. Our phylogenetic analyses examining *Irx* genes from present-day organisms suggest that the different three-gene clusters (present in flies and vertebrates) evolved independent from one another. Yet we cannot exclude the possibility that mechanisms implied for the concerted evolution of multigene gene families (such as unequal crossing-over and gene conversion) operated on either one of these three-gene clusters, causing the duplicated genes in each linkage group to become progressively more homogenized.

Physical mapping studies uncovered the clustered organization of murine *Irx* genes on mouse chromosomes 8 and 13. Moreover, our mapping studies also revealed that the relative position of cognate genes within both clusters is strictly conserved. The presence of such paralogous gene clusters can be best explained by an en masse duplication, either of the entire genome, or of a regionally restricted subset of the genome (Holland and Garcia-Fernández 1996; Skrabanek and Wolfe 1998). The evolution of *Hox* gene clusters is probably one of the best-studied examples representing paralogous gene clusters. Several studies revealed that duplication of *Hox* gene clusters included nearby developmentally important genes such as members of the *Distal-less, Integrin, Paired box,* and *Wingless* gene families (Ruddle et al. 1994; Holland and Garcia-Fernández 1996; Stock et al. 1996). Likewise, an en masse duplication of several genes was responsible for the creation of two T-box gene clusters on mouse chromosomes 5 and 11 (Agulnik et al. 1996; Ruvinsky and Silver 1997). Analysis of gene loci located near human *IRX* gene clusters on human chromosomes 5 and 16 resulted in the identification of additional pairs of paralogous genes. Thus the existence of cognate *Irx* gene clusters can be most likely attributed to a single duplication event encompassing a larger genomic region. Whether the linkage of genes to the *Irx* clusters is a structural remnant of the ancestral linkage pattern or serves an important biological function remains elusive.

Consequences of Clustered Organization

The linkage of *Irx* genes was demonstrated by their location on overlapping YAC clones identified by screening a large insert mouse YAC library (with an average insert size of 820 kb; Haldi et al. 1996). The fact that no more than two *Irx* genes were found to be located on a single YAC clone suggests that complete (three-gene) clusters cover several hundred kilobases of genomic sequences. However, transcriptional units of *Irx* genes appear to be encoded by <20 kb of genomic sequences (data not shown). Yet the clustered organization and the linear order of *Irx* genes has been conserved in the course of evolution, suggesting a selective advantage to this genomic arrangement.

Indeed, a correlation between the clustered organization of transcription factors and their coordinated expression has been shown for the members of the *Hox* gene clusters. Each of the four mammalian *Hox* gene clusters consists of ~10 transcriptional units that are tightly packed within 150 kb of genomic DNA (Krumlauf 1994). During mammalian embryonic development, *Hox* genes are activated in a spatiotemporal sequence that reflects their clustered organization. In addition, *trans*-paralogous *Hox* genes from different clusters show similar expression patterns consistent with their evolutionary relationships. Recent studies demonstrated that various *cis*-regulatory mechanisms are involved in the coordination of *Hox* gene expression. These include promoter competition or enhancer sharing occurring between nearby located *Hox* genes (for review, see Mann 1997; Duboule 1998) as well as the control of colinear activation of *Hox* genes by a global enhancer element located upstream of the *HoxD* complex (Kondo and Duboule 1999). Whether intergenic regions of *Irx* clusters harbor *cis*-regulatory elements that are essential for *Irx* gene expression remains to be investigated.

The vertebrate *Irx* family currently comprises six individual members. Expression analyses of four of the murine members demonstrated discrete spatially and temporally restricted patterns during embryogenesis (Bosse et al. 1997, 2000; Bruneau et al. 2000; Cohen et al. 2000). Thus, the presence of distinct *cis*-regulatory elements may account for specificity differences between members of different *Irx* gene clusters. However, based on the extreme conservation of their (DNA-binding) homeodomains, individual *Irx* proteins might recognize identical or similar DNA motifs and thus induce the very same or similar genetic programs. Regulation of gene expression is generally mediated by protein complexes, thereby modulating the specificity of DNA recognition. *Irx* proteins harbor a conserved motif downstream from the homeodomain, the IRO box. It will be interesting to determine if the IRO box is involved in the recruitment of cofactors that contribute to the specificity of *Irx* function.

Irx Nomenclature

The relationships among vertebrate *Irx* genes suggested here would be best represented by changes to the nomenclature. Simplification could be attained by adjusting the nomenclature of vertebrate *Irx* genes to the name of their mouse orthologs (see Fig. 1). Considering the genomic organization of vertebrate *Irx* genes, a new system analogous to that in use for the *Hox* gene clusters might be useful. Thus the two complexes might be named *IrxA* and *IrxB*. An identical alphabetic designation might be assigned to *cis*-paralogous *Irx* genes followed by a numeric identifier for each gene to indicate *trans*-paralogous (cognate) *Irx* genes. Reflecting the linear order and the evolutionary relationships of vertebrate *Irx* genes, nomenclature of *Irx1, Irx2,* and *Irx4* on mouse chromosome 13 might be changed to *Irxa1, Irxa2,* and *Irxa3,* and *Irx3, Irx5,* and *Irx6* on mouse chromosome 8 might be changed to *Irxb1, Irxb2,* and *Irxb3*. Consequently, human orthologs on chromosome 5p15.3 (*IRX5, IRX* (EST assembly R50645-AI831283), and *IRX4*) might be changed to *IRXA1, IRXA2,* and *IRXA3,* and human orthologs on chromosome 16q12 (*IRX1, IRX2,* and *IRX3*) might be changed to *IRXB1, IRXB2,* and *IRXB3*.

METHODS

Cloning of Murine *Irx5* 5′ End

The 5′end of murine *Irx5* cDNA was obtained by *Nhe*I/*Not*I restriction fragmentation of a PAC clone positive for *Irx*5. Cloned fragments were analyzed by sequencing, performing virtual alignments of deduced amino acid sequences to members of the *Irx2*/*Irx5* subgroup. One genomic clone was identified containing putative N-terminal sequences of *Irx5*. To evaluate whether the identified sequence was a genuine part of the *Irx5* cDNA we performed RT-PCR analysis using primer Irx5-F30 (5′-ATGTCCTACCCGCAGCTACTTG-3′) and P05951-B1 (5′-CATGATCTTCTCTCCCTTGGTGG-3′) originating from both the genomic clone and the EST clone W54596, respectively. RT-PCR analysis of RNA isolated from whole mouse embryos at E11.5 resulted in the amplification of a 440-bp PCR fragment. Sequencing confirmed the accuracy of the PCR product. PCR amplification using primers Irx5-F30 and P05951-B1 was carried out at 94°C for 30 sec, 57°C for 45 sec, 72°C for 30 sec, for 30 cycles, followed by a 7-min extension at 72°C.

Database Searches and Phylogenetic Analyses

The identification of additional *Irx* family members was performed by database searches using protein sequences derived from published *Irx* genes. Open reading frames were retrieved by TBLASTN (Altschul et al. 1990) searches against the six-frame translations of the GenBank database (NR, dbest, GSS, and HTGS divisions). Sequence assemblies generated from overlapping sequences and gapped assemblies were both guided along the coding frames of full-length sequences reported previously (including manual corrections for frameshifts in several of the primary sequence entries).

Alignment of protein sequences was performed using the CLUSTALW server (http://www.ebi.ac.uk) (and the JALVIEW sequence editor) provided by the EMBL-European Bioinformatics Institute (EBI). Phylogenetic analyses were performed using the PHYLIP server (http://bioweb.pasteur.fr) at the Pasteur Institute. Gene trees were calculated from Kimura (1980) distances employing the neighbor-joining algorithm (Saitou and Nei 1987).

Physical mapping information for identified YAC clones containing *Irx* genes was obtained from the database of the Whitehead Institute for Biomedical Research/MIT Center for Genome Research (WI/MIT CGR) (Nusbaum et al. 1999). Mapping information for human *IRX*, *ADCY2/7,* and *SLC6A2/3* genes was obtained from the NCBI database (GeneMap ’99, OMIM GeneMap).

Isolation of *Irx* Positive YACs

Combinatorial DNA pools of the large insert MIT mouse YAC library (WI/MIT 820) used for the construction of the WI/MIT CGR YAC-based physical map of the mouse genome (Copeland et al. 1993; Dietrich et al. 1994, 1996; Nusbaum et al. 1999) were screened by PCR to identify YAC clones positive for individual *Irx* genes. PCR primers specific for individual murine *Irx* sequences were as follows: *Irx1*: Irx1-F750 (5′-CATAGGCAAGTTTTCCAACTGGAC-3′), Irx1-B1000 (5′-GCGACTTTAACTGTTGTGGGGG-3′); *Irx2*: Irx2-F2050 (5′-GTTTGGGGCTCATTTTTACTGGAG-3′), Irx2-B2150 (5′-GGACGTTTATTTCACTGGCTCTTC-3′); *Irx3*: Irx3-F1 (5′-CCCTATCCAATGTGCTTTCATCAG-3′), Irx3-B1 (5′-GGCTGTCCTTCAGCTCATACTGAG-3′); *Irx4*: Irx4-F1 (5′-GCACGCCACGAACTCAATTC-3′), Irx4-B1 (5′-AGATGCCTCAGAACCATAGGTCAC-3′); *Irx5*: P05951-F1 (5′-TCACCCTTATGCAGCACCTCTG-3′), P05951-B1 (5′-CATGATCTTCTCTCCCTTGGTGG-3′); *Irx6*: Irx6-F1530 (5′-TCAGACTTGGAGGAGAGAAGGTGG-3′), Irx6-B1700 (5′-GACAGTCCTTGGAGACCATTTCAG-3′). Amplification conditions were 94°C for 30 sec, 55° –58°C for 45 sec, 72°C for 30 sec, for 35 cycles, followed by a 7-min extension at 72°C. The three-dimensional pooling scheme of the YAC library allowed the unambiguous identification of individual YAC clones positive for a given gene by only two rounds of PCR screening. For final marker content mapping, YAC DNAs were prepared from 10-mL overnight cultures using standard protocols. The YAC library was initially described by Haldi et al. (1996). YAC clones were reported to have an average insert size of 820 kb. DNA pools of the YAC library and identified YAC clones were provided by the Resource Center/Primary Database of the German Human Genome Project (Library No. 917). Identified YAC clones were as follows: WIBRy917G07228 (228-G-7), WIBRy917C0587 (87-C-5), WIBRy917B03357 (357-B-3), WIBRy917A02396 (396-A-2), WIBRy917F0391D2 (91-F-3), WIBRy917F06339 (339-F-6), WIBRy917H08228D1 (228-H-8), WIBRy917H11352D2 (352-H-11), WIBRy917B06419D2 (419-B-6), WIBRy917-A10388D2 (388-A-10), WIBRy917F04116D2 (116-F-4), WIBRy917C10270D2 (270-C-10).

ACKNOWLEDGMENTS

We thank Bill Martin for discussions. We thank Peter Sikorski for technical assistance and the members of the Rüther lab for critical reading of the manuscript. We thank the Resource Center of the German Human Genome Project for providing the mouse YAC library and clones. This work was supported by the Deutsche Forschungsgemeinschaft (SFB271 and Ru 376/8).

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Agulnik, S.I., Garvey, N., Hancock, S., Ruvinsky, I., Chapman, D. L., Agulnik, I., Bollag, R., Papaioannou, V., and Silver, L.M. 1996. Evolution of mouse T-box genes by tandem duplication and cluster dispersion. *Genetics* **144:** 249–254.

Alonso, M.C. and Cabrera, C.V. 1988. The *achaete-scute* gene complex of *Drosophila melanogaster* comprises four homologous genes. *EMBO J.* **7:** 2585–2591.

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.

Aparicio, S. 2000. Vertebrate evolution: recent perspectives from fish. *Trends Genet.* **16:** 54–56.

Bao, Z.Z., Bruneau, B.G., Seidman, J.G., Seidman, C.E., and Cepko, C.L. 1999. Regulation of chamber-specific gene expression in the developing heart by *Irx4*. *Science* **283:** 1161–1164.

Bellefroid, E.J., Kobbe, A., Gruss, P., Pieler, T., Gurdon, J.B., and Papalopulu, N. 1998. *Xiro3* encodes a *Xenopus* homolog of the *Drosophila Iroquois* genes and functions in neural specification. *EMBO J.* **17:** 191–203.

Bosse, A., Stoykova, A., Nieselt-Struwe, K., Chowdhury, K., Copeland, N.G., Jenkins, N.A., and Gruss, P. 2000. Identification of a novel mouse *Iroquois* homeobox gene, *Irx5*, and chromosomal localization of all members of the mouse *Iroquois* gene family. *Dev. Dyn.* **218:** 160–174.

Bosse, A., Zülch, A., Becker, M.B., Torres, M., Gómez-Skarmeta, J.L., Modolell, J., and Gruss, P. 1997. Identification of the vertebrate *Iroquois* homeobox gene family with overlapping expression during early development of the nervous system. *Mech. Dev.* **69:** 169–181.

Bruneau, B.G., Bao, Z.Z., Tanaka, M., Schott, J.J., Izumo, S., Cepko, C. L., Seidman, J. G., and Seidman, C.E. 2000. Cardiac expression of the ventricle-specific homeobox gene *Irx4* is modulated by *Nkx2–5* and *dHand*. *Dev. Biol.* **217:** 266–277.

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

Cavodeassi, F., Diez del Corral, R., Campuzano, S., and Dominguez, M. 1999. Compartments and organising boundaries in the *Drosophila* eye: The role of the homeodomain Iroquois proteins. *Development* **126:** 4933–4942.

Cohen, D.R., Cheng, C.W., Cheng, S.H., and Hui, C. 2000. Expression of two novel mouse *Iroquois* homeobox genes during neurogenesis. *Mech. Dev.* **91:** 317–321.

Copeland, N.G., Gilbert, D.J., Jenkins, N.A., Nadeau, J.H., Eppig, J.T., Maltais, L.J., Miller, J.C., Dietrich, W.F., Steen, R.G., Lincoln, S.E., et al. 1993. Genome maps IV 1993. *Science* **262:** 67–82.

Dib, C., Faure, S., Fizames, C., Samson, D., Drouot, N., Vignal, A., Millasseau, P., Marc, S., Hazan, J., Seboun, E., et al. 1996. A comprehensive genetic map of the human genome based on 5,264 microsatellites. *Nature* **380:** 152–154.

Dietrich, W.F., Miller, J.C., Steen, R.G., Merchant, M., Damron, D., Nahf, R., Gross, A., Joyce, D.C., Wessel, M., Dredge, R.D., et al. 1994. A genetic map of the mouse with 4,006 simple sequence length polymorphisms. *Nat. Genet.* **7:** 220–245.

Dietrich, W.F., Miller, J., Steen, R., Merchant, M.A., Damron-Boles, D., Husain, Z., Dredge, R., Daly, M.J., Ingalls, K.A., O'Conner, T.J., et al. 1996. A comprehensive genetic map of the mouse genome. *Nature* **380:** 149–152.

Diez del Corral, R., Aroca, P., Gómez-Skarmeta, J.L., Cavodeassi, F., and Modolell, J. 1999. The *Iroquois* homeodomain proteins are required to specify body wall identity in *Drosophila*. *Genes* & *Dev.* **13:** 1754–1761.

Duboule, D. 1998. Vertebrate *hox* gene regulation: Clustering and/or colinearity? *Curr. Opin. Genet. Dev.* **8:** 514–518.

Finnerty, J.R. and Martindale, M.Q. 1998. The evolution of the *Hox* cluster: Insights from outgroups. *Curr. Opin. Genet. Dev.* **8:** 681–687.

Funayama, N., Sato, Y., Matsumoto, K., Ogura, T., and Takahashi, Y. 1999. Coelom formation: Binary decision of the lateral plate mesoderm is controlled by the ectoderm. *Development* **126:** 4129–4138.

Gómez-Skarmeta, J.L., del Corral, R.D., de la Calle-Mustienes, E., Ferre-Marco, D., and 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.

Gómez-Skarmeta, J.L., Glavic, A., de la Calle-Mustienes, E., Modolell, J., and Mayor, R. 1998. *Xiro*, a *Xenopus* homolog of the

Drosophila Iroquois complex genes, controls development at the neural plate. *EMBO J.* **17:** 181–190.
Goriely, A., Diez del Corral, R., and Storey, K.G. 1999. *c-irx2* expression reveals an early subdivision of the neural plate in the chick embryo. *Mech. Dev.* **87:** 203–206.
Grillenzoni, N., van Helden, J., Dambly-Chaudiere, C., and Ghysen, A. 1998. The *iroquois* complex controls the somatotopy of *Drosophila* notum mechanosensory projections. *Development* **125:** 3563–3569.
Haldi, M.L., Strickland, C., Lim, P., VanBerkel, V., Chen, X., Noya, D., Korenberg, J.R., Husain, Z., Miller, J., and Lander, E.S. 1996. A comprehensive large-insert yeast artificial chromosome library for physical mapping of the mouse genome. *Mamm. Genome* **7:** 767–769.
Holland, P.W., and Garcia-Fernández, J. 1996. *Hox* genes and chordate evolution. *Dev. Biol.* **173:** 382–395.
Kehl, B.T., Cho, K.O., and Choi, K.W. 1998. *mirror*, a *Drosophila* homeobox gene in the *Iroquois* complex, is required for sensory organ and alula formation. *Development* **125:** 1217–1227.
Kimura, M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* **16:** 111–120.
Knust, E., Schrons, H., Grawe, F., and Campos-Ortega, J.A. 1992. Seven genes of the Enhancer of split complex of *Drosophila melanogaster* encode helix-loop-helix proteins. *Genetics* **132:** 505–518.
Kondo, T. and Duboule, D. 1999. Breaking colinearity in the mouse *HoxD* complex. *Cell* **97:** 407–417.
Krumlauf, R. 1994. *Hox* genes in vertebrate development. *Cell* **78:** 191–201.
Lennon, G., Auffray, C., Polymeropoulos, M., and Soares, M.B. 1996. The I.M.A.G.E. Consortium: an integrated molecular analysis of genomes and their expression. *Genomics* **33:** 151–152.
Lewis, M.T., Ross, S., Strickland, P.A., Snyder, C.J., and Daniel, C.W. 1999. Regulated expression patterns of *IRX-2*, an *Iroquois*-class homeobox gene, in the human breast. *Cell Tissue Res.* **296:** 549–554.
Leyns, L., Gómez-Skarmeta, J.L., and Dambly-Chaudiere, C. 1996. *iroquois*: a prepattern gene that controls the formation of bristles on the thorax of *Drosophila*. *Mech. Dev.* **59:** 63–72.
Mann, R.S. 1997. Why are *Hox* genes clustered? *Bioessays* **19:** 661–664.
McNeill, H., Yang, C.H., Brodsky, M., Ungos, J., and Simon, M.A. 1997. mirror encodes a novel PBX-class homeoprotein that functions in the definition of the dorsal-ventral border in the *Drosophila* eye. *Genes & Dev.* **11:** 1073–1082.
Netter, S., Fauvarque, M.O., Diez del Corral, R., Dura, J.M., and Coen, D. 1998. *white+* transgene insertions presenting a dorsal/ventral pattern define a single cluster of homeobox genes that is silenced by the polycomb-group proteins in *Drosophila melanogaster*. *Genetics* **149:** 257–275.
Nusbaum, C., Slonim, D.K., Harris, K.L., Birren, B.W., Steen, R.G., Stein, L.D., Miller, J., Dietrich, W.F., Nahf, R., Wang, V., et al. 1999. A YAC-based physical map of the mouse genome. *Nat. Genet.* **22:** 388–393.
Ruddle, F.H., Bentley, K.L., Murtha, M.T., and Risch, N. 1994. Gene loss and gain in the evolution of the vertebrates. *Dev. Suppl.* 155–161.
Ruvinsky, I. and Silver, L.M. 1997. Newly identified paralogous groups on mouse chromosomes 5 and 11 reveal the age of a T-box cluster duplication. *Genomics* **40:** 262–266.
Saitou, N. and Nei, M. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4:** 406–425.
Skrabanek, L. and Wolfe, K.H. 1998. Eukaryote genome duplication—where's the evidence? *Curr. Opin. Genet. Dev.* **8:** 694–700.
Stock, D.W., Ellies, D.L., Zhao, Z., Ekker, M., Ruddle, F.H., and Weiss, K.M. 1996. The evolution of the vertebrate *Dlx* gene family. *Proc. Natl. Acad. Sci.* **93:** 10858–10863.
Tan, J.T., Korzh, V., and Gong, Z. 1999. Expression of a zebrafish *iroquois* homeobox gene, *ziro3*, in the midline axial structures and central nervous system. *Mech. Dev.* **87:** 165–168.

Received April 13, 2000; accepted in revised form August 11, 2000.