Developmental Biology **227,** 239–255 (2000)
doi:10.1006/dbio.2000.9883, available online at http://www.idealibrary.com on IDEAL®

REVIEW

Phylogeny of the SOX Family of Developmental Transcription Factors Based on Sequence and Structural Indicators

Josephine Bowles,[1] Goslik Schepers,[1] and Peter Koopman[2]
Institute for Molecular Bioscience, University of Queensland, Brisbane 4072, Australia

Members of the SOX family of transcription factors are found throughout the animal kingdom, are characterized by the presence of a DNA-binding HMG domain, and are involved in a diverse range of developmental processes. Previous attempts to group *SOX* genes and deduce their structural, functional, and evolutionary relationships have relied largely on complete or partial HMG box sequence of a limited number of genes. In this study, we have used complete HMG domain sequence, full-length protein structure, and gene organization data to study the pattern of evolution within the family. For the first time, a substantial number of invertebrate SOX sequences have been included in the analysis. We find support for subdivision of the family into groups A–H, as has been suggested in some previous studies, and for the assignment of two new groups, I and J. For vertebrate genes, it appears that relatedness as suggested by HMG domain sequence is congruent with relatedness as indicated by overall structure of the full-length protein and intron–exon structure of the genes. Most of the SOX groups identified in vertebrates were represented by a single *SOX* sequence in each invertebrate species studied. We have named anonymous sequences and, where appropriate, have suggested systematic names for some previously identified sequences. In addition, we identify an HMG domain signature motif which may be considered representative of the SOX family. Based on our data, we propose a robust phylogeny of *SOX* genes that reflects their evolutionary history in metazoans.

***Key Words: SOX; SRY*-related; HMG box; transcription factors; phylogeny; nomenclature.**

INTRODUCTION

The SOX family of transcription factors was first identified in mammals in 1990 based on conservation of the HMG box of the gene for the mammalian testis-determining factor SRY (Gubbay *et al.,* 1990). The HMG domain, a DNA binding motif of approximately 79 amino acids, characterizes the HMG domain superfamily which is composed of two subfamilies—representatives of the TCF/SOX/MATA group typically contain single sequence-specific HMG domains while members of the HMG/UBF group have multiple HMG domains which are less sequence-specific in their binding (Laudet *et al.,* 1993; Grosschedl *et al.,* 1994; Soullier *et al.,* 1999). Members of the TCF/SOX/MATA family tend to bind, in the minor groove of the DNA, to variants of the sequence motif $^{A}/_{T}$ $^{A}/_{T}$ CAAAG (Laudet *et al.,* 1993, and references therein). *SOX* genes appear to be restricted to animals (Soullier *et al.,* 1999; Wegner, 1999) and have now been identified in birds, reptiles, amphibians, fish, insects, and nematodes, with at least 30 members currently recognized in mammals (for recent reviews see Pevny and Lovell-Badge, 1997; Wegner, 1999). SOX transcription factors show both classical and architectural modes of action and have diverse tissue-specific expression patterns during early development. They have been implicated in cell fate decisions in numerous developmental processes (Pevny and Lovell-Badge, 1997; Wegner, 1999). By convention, HMG domains of SOX proteins are at least 50% identical to the HMG domain of SRY. Outside the HMG domain, SOX sequences are quite variable, although common nonbox domains can be identified among a number of SOX proteins, suggesting recent

[1] These authors contributed equally to this work.
[2] To whom correspondence should be addressed. Fax: +61 7 3365 4388. E-mail: p.koopman@imb.uq.edu.au.

0012-1606/00 $35.00

shared ancestry. Among vertebrates, orthologous genes (representatives of the same gene in different species) are generally very similar to each other in terms of both HMG box homology and nonbox structural conservation.

To better understand the role of SOX proteins in development, and in order to facilitate comparative analyses, it is of interest to determine evolutionary relationships among *SOX* family genes. *SOX* genes are dispersed throughout the mouse and human genomes (Wegner, 1999), arguing against a purely tandem duplication model of SOX family expansion. One theory is that the family has arisen from a common ancestor via ancient duplication, dispersal, mutation, and acquisition mechanisms. We presume that at various times throughout metazoan evolution, HMG box-containing sequences duplicated, in each case leaving one redundant copy which was then free to evolve a new function or else be lost from the genome (Force *et al.*, 1999). Rather than relying on slow genetic drift, it is likely that "spare" HMG box-containing fragments recruited preexisting functional domains and hence formed mosaic proteins capable of rapidly taking on novel functions (Ohno, 1970; Patthy, 1991, 1994; Holland *et al.*, 1994). Since the HMG domain is a DNA binding domain with a recognition site that is largely conserved throughout the family, it is likely to have diverged by gradual drift. Hence, the HMG domain may be considered an independent evolutionary unit, and we predict that HMG domain variation will be an accurate marker of the pattern of evolution of the family, provided that sufficient information is present and that back mutations have not obscured its history. Since the remainder of the protein is likely to have evolved in a more stochastic and erratic fashion, we expect that non-HMG domain SOX sequences will be of limited usefulness in molecular phylogenetic studies at the level of the SOX family. We anticipate, however, that the structure of the non-HMG domain regions of SOX proteins will be informative in a more classical phylogenetic sense.

A number of earlier studies have addressed the evolutionary history of the SOX family. Wright *et al.* (1993) conducted a pairwise comparison of partial HMG domain sequences among 15 known mouse *Sox* genes and used identity scores to assign six provisional groups within the SOX family. These were A, *Sry;* B, *Sox1, -2, -3,* and *-14;* C, *Sox4, -11,* and *-12;* D, *Sox5, -6,* and *-13;* E, *Sox8, -9,* and *-10;* and F, *Sox7.* This was later expanded to seven groups (A–G) with the discovery of *Sox15* and *-20* (van de Wetering and Clevers, 1993; Meyer *et al.*, 1996). An eighth group (H) was suggested recently to accommodate *SOX30* (Osaki *et al.*, 1999). Laudet *et al.* (1993) constructed distance-based phylogenies using both full and partial HMG domain sequences as available. In some cases, the various *SOX* genes identified within a particular species or species group were observed to cluster together. In other instances, clustering of putative orthologues was observed, as might be expected. It is unclear whether this inconsistent result represents sampling or PCR artifact or whether it illustrates that very rapid duplication and divergence has occurred within certain lineages, resulting in the evolution of species- or lineage-specific genes. In a more recent phylogenetic study (Soullier *et al.*, 1999), six SOX groups which differed from the seven groupings defined previously (Wright *et al.*, 1993) were identified. These authors recognized groups SOX5/6, SRY, SOX2/3, SOX14, SOX4/22, and SOX9/18, although both SOX9/18 and SRY groups were paraphyletic in their analyses. In contrast to the groupings suggested by Wright *et al.* (1993), this arrangement combines groups E and F and creates a group solely for the *Drosophila* sequence SOX14 (Accession No. X65667; Denny *et al.*, 1992).

In this study, we aimed to reconstruct the evolutionary history of the SOX family using the large number of HMG domain sequences and full-length protein sequences currently available. We examined whether phylogenies based solely on HMG domain sequence data are congruent with expectations of relatedness based on structural and functional conservation of the entire protein. An additional aim was to revise and simplify nomenclature and classification of *SOX* genes. In contrast to earlier studies, only complete HMG domain sequences were used, none of which was generated by PCR amplification. Invertebrate sequences from *Caenorhabditis elegans* (roundworm, phylum Nematoda), *Drosophila melanogaster* (fruit-fly, phylum Arthropoda), and *Strongylocentrotus purpuratus* (sea urchin, phylum Echinodermata) were included in the analysis. These represent two of the three large monophyletic clades within the metazoans. *C. elegans* and *D. melanogaster* are ecdysozoans while *S. purpuratus*, like vertebrates, are deuterostomes (Aquinaldo *et al.*, 1997).

Our data support subdivision of the family into groups A–H, as has been suggested previously, as well as the recognition of two new groups, I and J. We identified invertebrate sequences associated with most of these groups. For vertebrate SOX sequences, we show that relatedness as suggested by HMG domain sequence is congruent with relatedness as suggested by overall gene and protein structure. Our HMG domain sequence data set thus allows us to construct SOX family phylogenies to reflect the evolutionary history of this gene family in metazoans. In addition, this study has allowed us to assign meaningful names to previously anonymous sequences, recommend suitable names for ambiguously named sequences, and define robust criteria for *SOX* gene identification and classification.

MATERIALS AND METHODS

Extraction of SOX Sequences

Most SOX sequences used in this study were obtained from various databanks (GenBank, EMBL, SwissProt) using BLAST 2.0 searches for homology to known SOX HMG domains. *C. elegans* SOX sequences were obtained from a Wu-Blast version 2.0 search of the "Wormpep" peptide database (http://www.sanger.ac.uk/projects/C_elegans/) using a SOX HMG domain consensus motif as the query. Novel *D. melanogaster SOX* genes were obtained from

a tBlastX search of the High Throughput Genomic Sequence database and the *Drosophila* genome database (http://www.ncbi.nlm.gov/BLAST/), using either mo-*Sox1* or mo-*Sox8* HMG box nucleotide sequence as the query. All accession numbers are given in Fig. 1.

SOX HMG Domain Alignment

Multiple sequence alignments were made using CLUSTALW, Genetics Computer Group (GCG) Wisconsin Package, Version 8, 1994 (Thompson *et al.*, 1994). Only full-length HMG box (nucleotide) and domain (amino acid) sequences are included except where the sole example of a particular SOX protein contains only a partial HMG domain (mo-SOX12 and mo-SOX16).

Construction of Phylogenetic Trees

Distance matrices were computed from amino acid sequence data using EProtdist (GCG), which is based on the Dayhoff PAM 250 matrix. Phylogenetic analyses were performed using distance-based GCG programs EFitch, EKitsch, and ENeighbour. Phylogenetic analyses were also performed (on amino acid sequence and DNA sequence data) using the principles of maximum parsimony (EProtPars; GCG) and maximum likelihood with molecular clock (EDNAMLK; GCG). Bootstrapping was carried out on 100 replicates using ESeqboot (GCG). Phylogenetic trees were displayed using the WebANGIS helper application TREEVIEW (http:taxonomy.zoology.gla.ac.uk/rod/treeview.html). Incomplete HMG domain sequences were not used in phylogenetic analyses. In some cases trees were rooted with established outgroups fu-MATA1, mo-TCF1, or mo-LEF1 (Laudet *et al.*, 1993).

RESULTS

Alignment and Grouping of SOX HMG Domain Sequences

All available full-length HMG domain sequences are presented in aligned form in Fig. 1, along with two incomplete sequences. Because the HMG domain is highly conserved among SOX proteins, the alignment was unambiguous with no gaps or insertions. A SOX HMG domain consensus sequence was generated from all the full-length HMG domain sequences available. Dashes in Fig. 1 represent homology to the consensus sequence so that variability among sequences can be easily surveyed.

In this figure, sequences are presented arranged into groups as previously defined (Wright *et al.*, 1993; Hiraoka *et al.*, 1998b; Osaki *et al.*, 1999; Wegner, 1999) and as refined in this study. It is obvious from sequence comparison alone that different SOX groups can be invoked, each associated with characteristic amino acid strings (e.g., group D, $A^{K}/^{D}_{R}/_{E}$ERR; group E, AQAARR; and group F, $AK^{I}/_{D}$ERK; for positions 14–19). It is also evident that vertebrate orthologues are highly conserved in HMG domain sequence, although this conservation falls off considerably outside the HMG domain (Jay *et al.*, 1997; Kamachi *et al.*, 1998; Lefebvre *et al.*, 1998; Uchikawa *et al.*, 1999; Schepers *et al.*, 2000). The sequence motif RPMNAF (positions 5–10) is conserved for all SOX sequences but not for outgroup sequences fu-MATA1, mo-LEF1, or mo-TCF1. An extended version of this, RPMNAFMVW (positions 5–13), is conserved for all SOX sequences other than SRY and may represent a useful signature by which *SOX* genes can be recognized.

We have allocated names to previously anonymous sequences where appropriate and renamed some sequences, to better reflect their phylogenetic status, to simplify nomenclature, or to resolve ambiguity (Table 1). In doing so we recognize that, at least for the three invertebrate representatives included here, a single *SOX* gene usually corresponds to an entire group in vertebrates. Based on this observation, we propose that vertebrate genes be numbered, in approximate order of discovery, regardless of their grouping, but that invertebrate genes be given a letter designation reflecting their grouping. New names (*) and revised names (#) are marked in Fig. 1, and details relating to revised SOX nomenclature are presented in Table 1.

Intron Conservation within and between SOX Groups

Intron gains and losses represent major genetic rearrangements and are therefore relatively rare events when compared with sequence changes. Because it is highly unlikely that an intron would arise at the same sequence position in different lineages, scoring for the presence or absence of introns at particular positions is often used as a phylogenetic tool to assess relatedness.

No HMG box introns have been reported for vertebrate members of SOX groups A, B, C, or G or for sea urchin group B sequences. However, based on comparison of available genomic and cDNA sequences, seven of the eight *Drosophila* and *C. elegans* members of groups B and C do have introns in their HMG boxes. In some cases the positions of these are conserved, suggesting that introns have been lost in the deuterostome lineage of group B (Fig. 1). Where HMG box introns are known for members of groups D, E, and F, their positions are conserved within the group. The position of the HMG box intron for group D and group F sequences is conserved between the two groups, which might suggest recent shared ancestry; however, such a relationship is not supported by further phylogenetic analysis (Figs. 2 and 3). The intron positions in groups H and J (see below) differ from each other and from all other groups, supporting their classification as separate groups.

Phylogenetic Analysis of SOX Protein HMG Domain Sequences

Of the 79 positions in the aligned SOX HMG domain sequences, 59 were phylogenetically informative (at least two different amino acids are found in that position, and each of the two was present in at least two different sequences). We used distance-based methods to analyze phylogenetic relationships among these sequences and re-

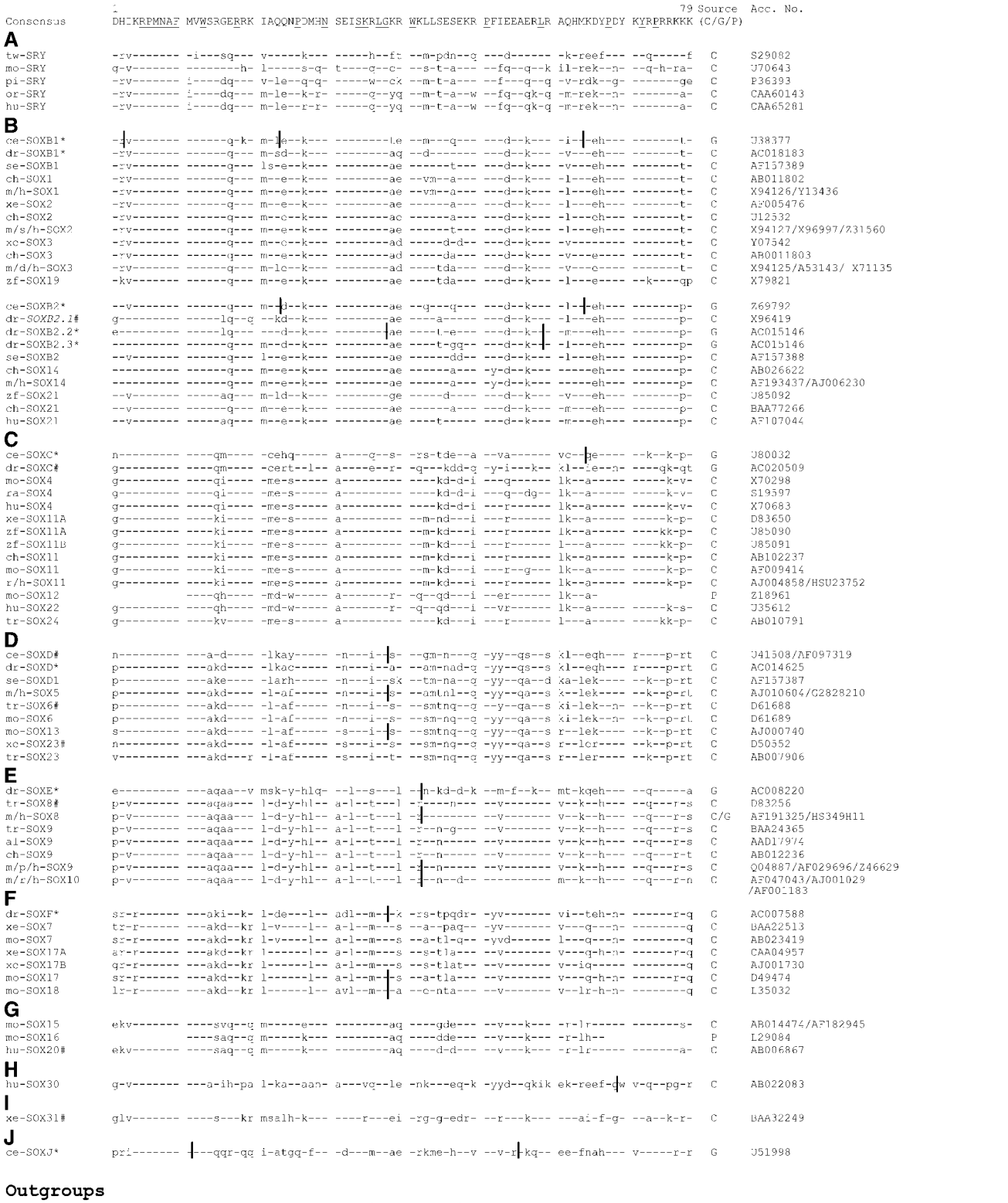


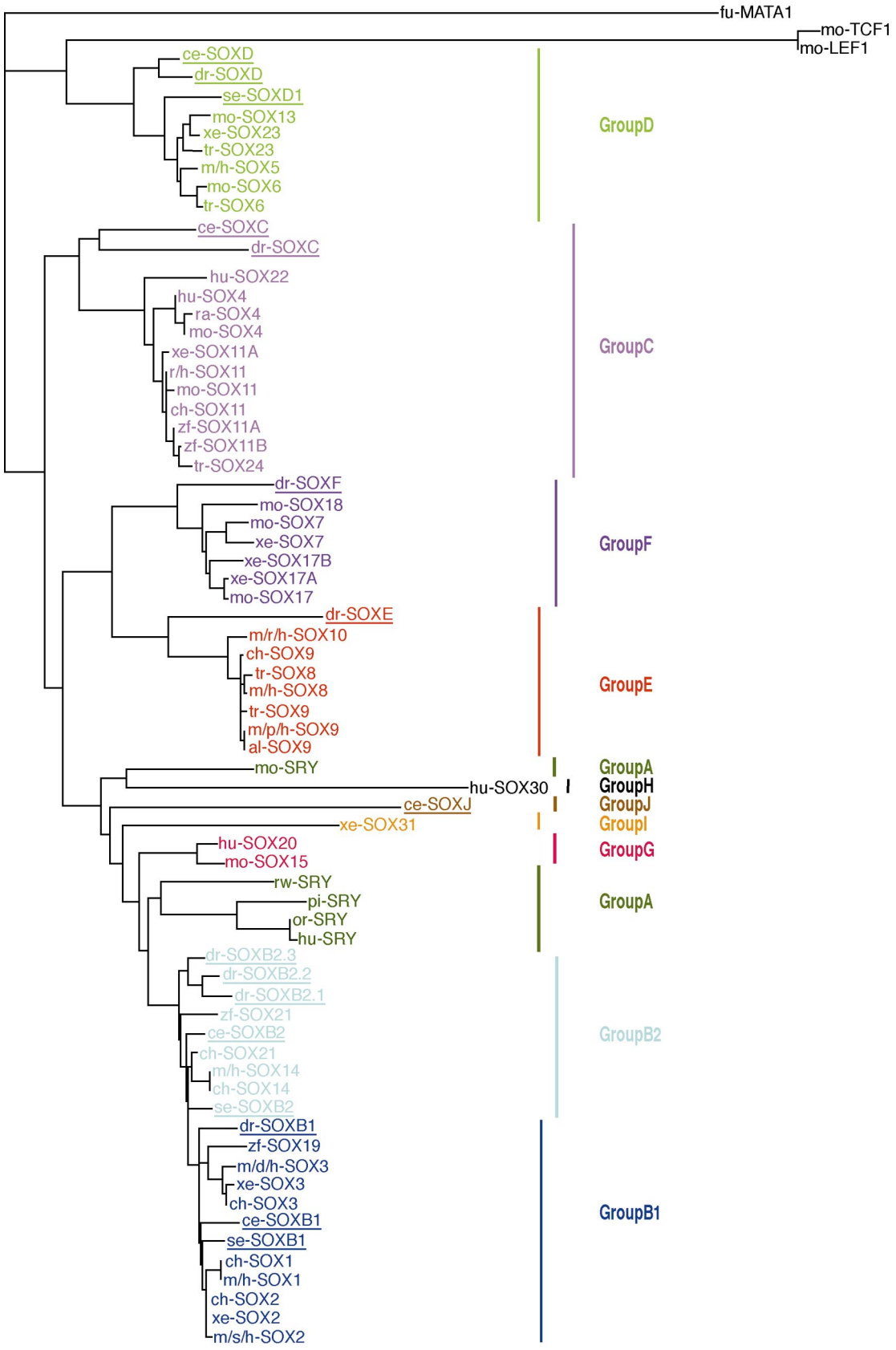
TABLE 1

Preferred Names for Anonymous or Previously Designated Sequences as Recommended by This Study

Preferred name	Species	Other name(s)	Group	Accession No.	Reference
SOXB1	*C. elegans*	Anon.	B1	U38377	Wegner, 1999
SOXB2	*C. elegans*	Anon.	B2	Z69792	Wegner, 1999
SOXC	*C. elegans*	Anon.	C	U80032	Wegner, 1999
SOXD	*C. elegans*	COG-2	D	U41508	Hanna-Rose and Han, 1999
SOXJ	*C. elegans*	Anon.	J	U51998	
LEF/TCF	*C. elegans*	Anon.	LEF/TCF	AF043702	
SOXB1	*D. melanogaster*	SOX29F	B1	AC018183	S. Russell, unpublished
		SOXNeuro		AJ252124	Crémazy *et al.*, 2000
SOXB2.1	*D. melanogaster*	*dichaete*,	B	X96419	Russell *et al.*, 1996
		fish-hook,		U68056	Nambu and Nambu, 1996
SOX70D		SOX70D		AA246637	F. Girard, unpublished
SOXB2.2	*D. melanogaster*	Anon.	B2	AC015146	
SOXB2.3	*D. melanogaster*	Anon.	B2	AC015146	
SOXC	*D. melanogaster*	SOX14	C	AJ252125	Denny *et al.*, 1992
		SOX60B			F. Girard, unpublished
SOXD	*D. melanogaster*	SOX102D	D	AC01916	S. Russell, unpublished
SOXE	*D. melanogaster*	SOX100B	E	AJ251580	Loh and Russell, 2000
				AC008220	
SOXF	*D. melanogaster*	SOX15	F	AJ250955	Denny *et al.*, 1992
		SOX50E		AC007588	S. Russell, unpublished
SOX23	*X. laevis*	SOX12	D	D50552	Komatsu *et al.*, 1996
SOX6	*O. mykiss*	SOXLZ	D	D61688	Takamatsu *et al.*, 1995
SOX8	*O. mykiss*	SOXP1	E	D83256	Ito *et al.*, 1995
SOX20	*H. sapiens*	SOX12	G	AB006867	Gozé *et al.*, 1993
SOX21	*H. sapiens*	SOX30	B2	AB022083	Osaki *et al.*, 1999
SOX31	*X. laevis*	SOXD	I	BAA32249	Mizuseki *et al.*, 1998

Note. Accession numbers and references are included. Previously named SOX sequences (where full-length HMG box sequence is available) have been renamed systematically where deemed appropriate based on evolutionary relationships, for consistency, or to avoid confusion. For invertebrate sequences, we recommend that genes be named to reflect the SOX subgroup they appear to represent. For this reason, SOXD, indicating membership of group D, is preferred to COG-2, SOXB1 is preferred to SOXNeuro, and SOXB2.1 is preferred to *dichaete*, *fish-hook*, or SOX70D. *Drosophila* SOX14 is referred to here as SOXC based on its association with group C and to distinguish it from the mouse, human, and chicken SOX14 sequences of group B2. *Drosophila* SOX100B is referred to as SOXE, and SOX15 as SOXF (mouse SOX15 is in group G). Based on sequence conservation *Xenopus* SOX12 is a likely orthologue of rainbow trout SOX23 (data not shown), rainbow trout SOXLZ is a likely orthologue of mouse SOX6 (Wegner, 1999), and rainbow trout SOXP1 is a likely orthologue of mouse SOX8 (Schepers *et al.*, 2000). Human SOX12 is identical to SOX20 while human SOX30 is identical to SOX21. To avoid confusion with group D, we consider SOX31 a more appropriate name for xe-SOXD.

FIG. 1. Alignment of SOX protein HMG domains. HMG domain sequences inferred for most known *SOX* genes from a range of species are given, arranged in related groups as surmised based on sequence similarity. Accession numbers are shown in each case (Acc. No.). Only two incomplete HMG domain sequences have been included—these are mo-SOX12 (group C) and mo-SOX16 (group G), the only sequences derived purely from PCR amplification data. The source of sequence information is shown (G, genomic; C, cDNA; P, PCR). Outgroup sequences are included. Dashes represent identity to the consensus sequence (top). Where known, the positions of box introns are marked (|). ce-LEF/TCF sequence contains a 2-amino-acid insertion (nn) after position 26. al, alligator, *Alligator mississippiensis;* ce, nematode, *Caenorhabditis elegans;* ch, chicken, *Gallus gallus;* dr, fruit-fly, *Drosophila melanogaster;* du or d, dunnart (marsupial), *Sminthopsis macroura;* fu, fungi, *Saccharomyces cerevisiae;* hu or h, human, *Homo sapiens;* mo or m, mouse, *Mus musculus;* or, orangutan, *Pongo pygmaeus;* pi or p, pig, *Sus scrofa;* ra or r, rat, *Rattus norvegicus;* tw, tammar wallaby (marsupial), *Macropus eugenii;* sh or s, sheep, *Ovis aries;* tr, rainbow trout, *Oncorhynchus mykiss;* se, sea urchin, *Strongylocentrotus purpuratus;* xe, frog, *Xenopus laevis;* zf, zebrafish, *Danio rerio.* *Sequences identified from database searches as *SOX* genes and first named in this report; #sequences previously identified as *SOX* genes with our preferred name.



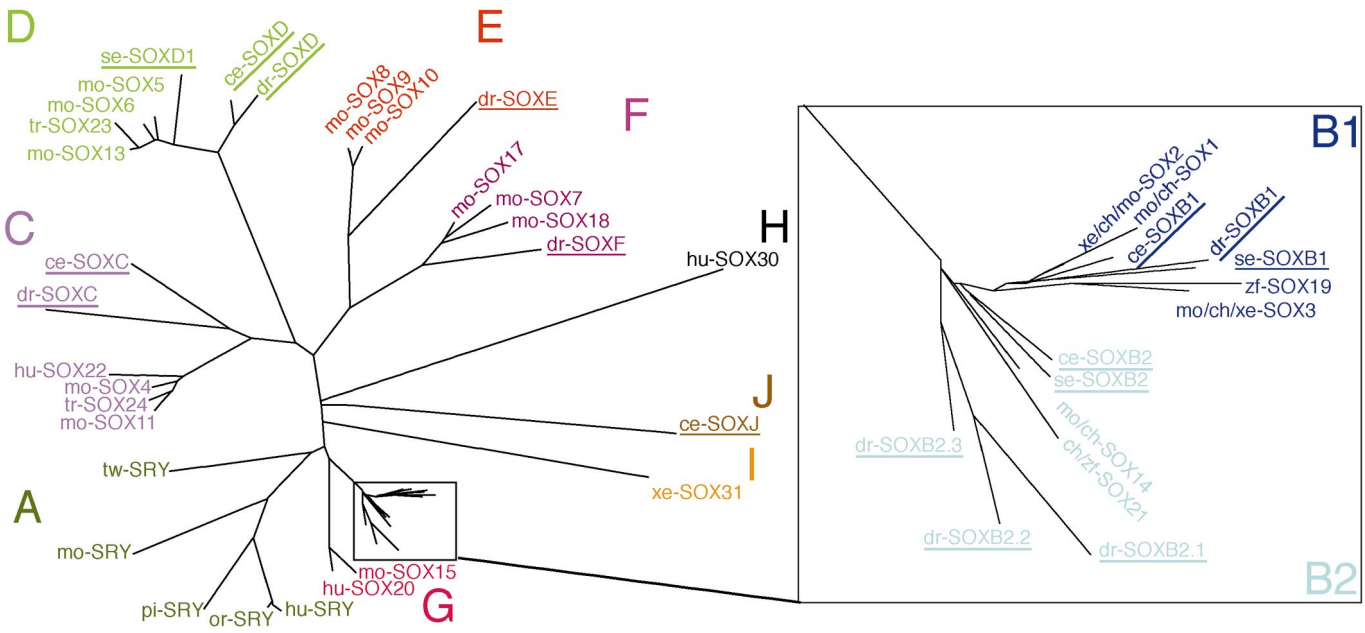


FIG. 3. An unrooted phylogeny for the SOX HMG domain, computed using the distance method FITCH (GCG). Branch lengths are representative of the extent of divergence. For groups of presumed mammalian orthologues (other than group A—Sry), only one representative has been included. The various groups are highlighted by use of color. Inset shows group B in enlarged format. B1 and B2 clades are colored differently. Invertebrate sequences are underlined. Abbreviations as given for Fig. 1.

sults for one such analysis (using eFITCH; Fitch, 1981) are shown in Fig. 2. In this representation, branch lengths are indicative of the extent of divergence, and the order of branching demonstrates that SOX proteins segregate into distinct groups. This is one of a number of possible trees that can be generated from the data. For this reason, the fine branching detail of the tree is not reliable although the clustering of sequences into groups is robust and stable. This type of information is usually presented by way of a bootstrapped or resampled tree—in this case bootstrap analysis is difficult because of the large number of sequences under consideration and the small size of the HMG box. Except for *Drosophila* group B2 genes, we found no evidence of species-specific clustering that has been reported previously (Laudet *et al.*, 1993). In most instances, invertebrate sequences (underlined) appear ancestral to clusters of related vertebrate sequences. For group D the expected Ecdysozoa/Deuterostomia clustering pattern is observed. This is not always the case for groups B1 and B2, probably reflecting an inability of the HMG sequence analysis to adequately resolve these evolving groups. Members of group A (mammalian SRY proteins) do not form a monophyletic group in this particular phylogenetic analysis although this unexpected result may reflect their relatively rapid rate of divergence (see below).

Human SOX30 does not fall within any of the previously defined groups A–G and so is classified, as previously suggested, as the sole member of group H (Osaki *et al.*, 1999). Similarly, *Xenopus* sequence SOX31 (previously xe-SOXD, Table 1; Mizuseki *et al.*, 1998) is not represented within any of the defined groups, although it clusters among the SOX sequences, and it has been assigned here to group I. A fifth *C. elegans* sequence (U51998) clearly represents a SOX protein, yet it is quite distinct in sequence from all known SOX HMG domains. Here we have assigned it to group J and called it ce-SOXJ.

Since vertebrate orthologues are highly conserved in HMG domain sequence (Fig. 1), single representative verte-

FIG. 2. Rooted phylogeny for the SOX HMG domain sequences. The tree was computed using the distance method FITCH (GCG). fu-MATA1 was specified as the outgroup and additional outgroup sequences mo-LEF1 and mo-TCF1 are also included. Previously defined SOX family groups A–G are supported. hu-SOX30, xe-SOX31, and ce-SOXJ are included—they do not fall into any of the previously defined groups and so are each designated the sole member of group H, I, and J, respectively. Branch lengths are proportional to evolutionary distance or extent of divergence. Invertebrate sequences are underlined. Abbreviations and sequence sources as for Fig. 1.

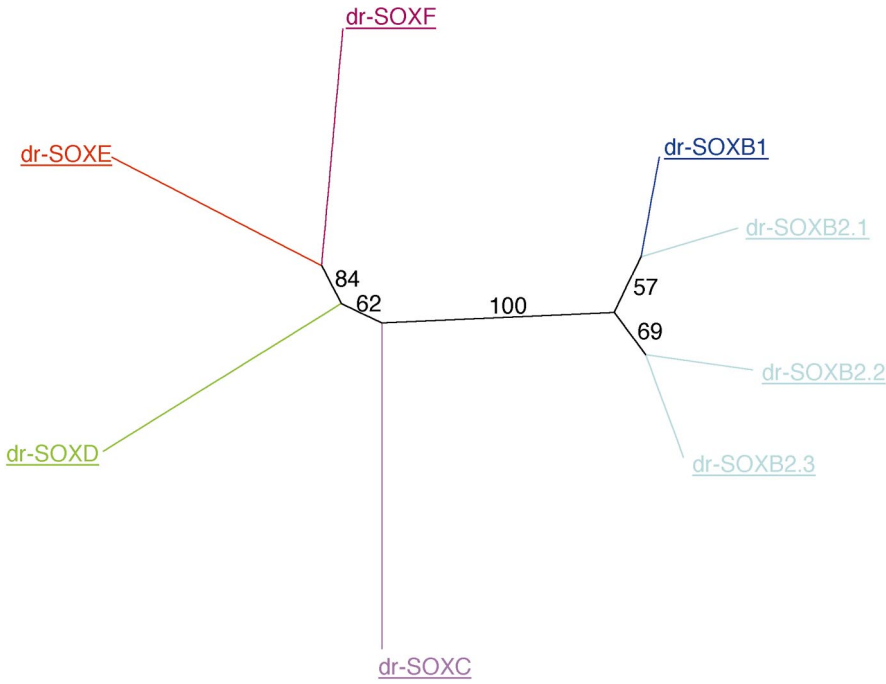


FIG. 4. An unrooted phylogeny for *Drosophila SOX* gene HMG box sequences. DNA sequence data were analyzed using the principle of maximum likelihood with molecular clock (eDNAMLK; GCG). To assess the robustness of branching, 100 bootstrap replicates were carried out (eSeqboot; GCG) and % support values are marked.

brate members of each SOX protein were phylogenetically analyzed along with all available invertebrate sequences (Fig. 3). The results are illustrated in an unrooted tree format with branch lengths representative of the extent of divergence. Because of the large number of group B sequences, this group is shown in enlarged format. The proposed division of group B into B1 and B2 (Uchikawa *et al.*, 1999) is illustrated. Similar results were obtained using other methods of phylogenetic analysis (neighbor-joining, parsimony) and when HMG box nucleotide sequences were used (results not shown). SOX family groups A–J are supported in these analyses although, because of the large number of sequences considered and the limited number of informative amino acid positions in the HMG domain, bootstrap values were low (data not shown).

Our analysis included five *C. elegans*, eight *Drosophila*, and three sea urchin sequences. *C. elegans* proteins ce-SOXB1, ce-SOXB2, ce-SOXC, and ce-SOXD appear, by virtue of their HMG domain sequences, to be true members of the SOX family and are closely associated with subgroups B1 and B2 and groups C and D, respectively. *Drosophila* SOX proteins dr-SOXB1, dr-SOXB2.1, dr-SOXB2.2, dr-SOXB2.3, dr-SOXC, dr-SOXD, dr-SOXE, and dr-SOXF appear to be associated with groups B1, B2, C, D, E, and F, as indicated by their names. Sea urchin sequences se-SOXB1, se-SOXB2, and se-SOXD1 are associated with subgroups B1 and B2 and group D, respectively.

Since representatives of most SOX groups have been identified in *Drosophila*, we reasoned that phylogenetic analysis of these sequences could provide a clearer picture of evolutionary relationships within the SOX family (Fig. 4). Since the number of taxa to be analyzed was relatively small, a maximum likelihood method was used and bootstrap values have been shown for all branch points. The close relationship among group B *Drosophila* sequences was supported in 100% of replicates although the exact branching pattern within the group was not resolved. The overall group topology was consistent with that found in distance-based analyses (Figs. 2 and 3) with bootstrap values supporting the relationships between SOX groups.

Comparison of Full-Length Structures for Vertebrate SOX Proteins

Rather than relying entirely on HMG domain sequences to reconstruct the evolutionary history of the SOX family, we also considered non-HMG domain sequences where available. Full length SOX sequences and structural and/or functional features were compiled from published information. Figure 5 shows representative full-length vertebrate SOX proteins in schematic form, arranged into groups as defined by phylogenetic analysis of HMG domain sequences (Wright *et al.*, 1993, this study, Figs. 1, 2, and 3). Since mammalian SOX orthologues are reasonably well

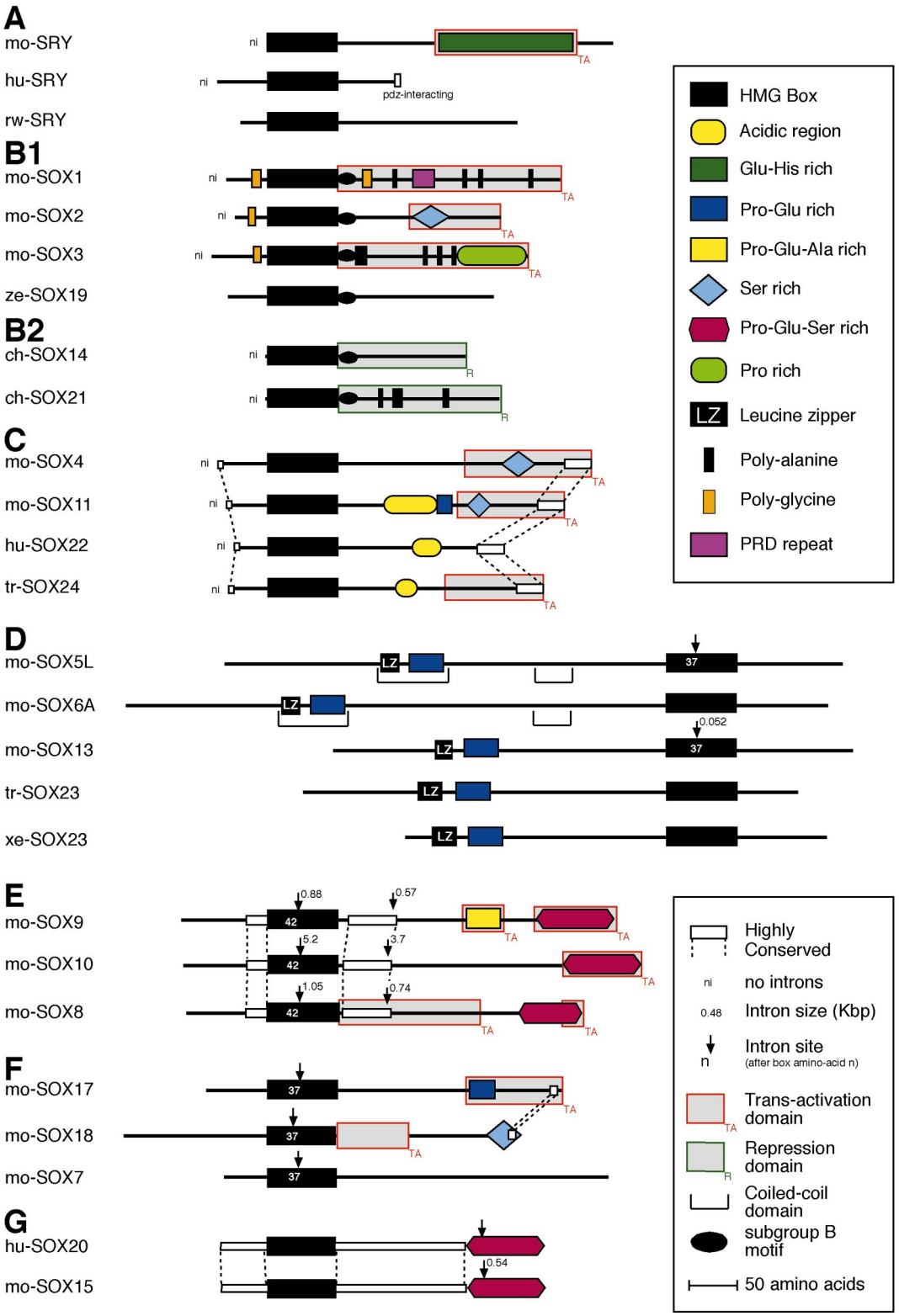


FIG. 5. Schematic representation of SOX proteins highlighting conservation within SOX family groups. Proteins are arranged in groups as defined by HMG domain sequence. Various structural features, motifs, and functional regions (demonstrated or putative) are shown along with intron positions and sizes where known. Genomic structures are known in some cases—"ni" (no intron) indicates that an

conserved, only one mammalian representative (usually mouse) of each orthologous group is included. Where available, we have included structural domains, functional domains, regions of high conservation, intron positions, and the exact positions and sizes of HMG box introns.

It is clear that a characteristic arrangement of structural and functional domains is associated with each group, with the exception of group A (SRY). The most obvious feature of conservation within groups is the position of the HMG domain within the protein and the overall length of the protein. Also notable are regions of group-specific conservation. For example, the leucine zippers and proline/glutamine-rich regions are conserved within group D (Hiraoka *et al.*, 1998a; Roose *et al.*, 1999), a group B motif of approximately 8 amino acids (Uchikawa *et al.*, 1999) is conserved just C-terminal of the HMG domain, and regions of conservation are also apparent in groups C, E, and G. The positions of the HMG box introns are invariably conserved within groups (see also Fig. 1), although their sizes are variable. In cases in which transcriptional activator or repressor domains have been identified, there is usually some conservation both in structure and in position within proteins of each group. This analysis shows that vertebrate members of the same group share significant and characteristic structural and organizational similarity outside of the HMG domain, providing the strongest evidence to date of their common evolutionary origin.

DISCUSSION

A Robust Grouping System for SOX Genes

The SOX family groupings we define in this study are in broad agreement with those proposed previously (Laudet *et al.*, 1993; Wright *et al.*, 1993; Soullier *et al.*, 1999; Wegner, 1999). Other than for *Drosophila* group B2 sequences (see below), we do not find evidence of species- or species group-specific clusters of sequences as has been reported (Laudet *et al.*, 1993). Results of that study may reflect the inclusion of PCR-generated sequences—we used only complete HMG domain sequences generated by means other than PCR. The groupings we recommend are not entirely in agreement with those proposed by Soullier and co-workers (1999). Specifically, we find no evidence to support the joining of groups E and F into a single SOX9/18 group, instead supporting those groups originally proposed by Wright and co-workers (1993). We consider the conventional groupings warranted, largely on the basis of differences in HMG box intron positions and non-HMG domain protein structures, but also due to the evidence of HMG domain sequence phylogeny and the finding of distinct *Drosophila* group E and F representatives. Further, we do not recognize the "SOX14" group (Laudet *et al.*, 1993 which included just one member, *Drosophila* SOX14; the HMG domain sequence of this SOX protein (Denny *et al.*, 1992, X65667; Crémazy *et al.*, 2000, AJ252125) is now complete and has been included here as *Drosophila* SOXC and found to be representative of group C.

Consideration of the full-length sequences and functional roles of some vertebrate members of group B (*SOX1*, *-2*, *-3*, *-14*, and *-21*) led to the suggestion that they would be more correctly partitioned into two distinct subgroups, B1 and B2 (Uchikawa *et al.*, 1999; Fig. 5). The most compelling evidence for this partitioning comes from the demonstration that SOX14 and -21 (subgroup B2) proteins can act as transcriptional repressors, while SOX1, -2, and -3 (subgroup B1) are transcriptional activators. Although molecular phylogenetic analysis of the HMG domain alone is insufficient to conclusively define groups B1 and B2, we do find some clustering within group B (Figs. 2 and 3). There is support for an early subdivision of group B to subgroups B1 and B2 in that both SOXB1- and SOXB2-like sequences have been identified in *Drosophila, C. elegans,* and sea urchin.

We have confirmed SOX30 as the sole member of group H (Osaki *et al.*, 1999) and have defined two additional groups, I and J, to accommodate two SOX sequences which do not fall clearly into any of the previously defined SOX groups. Currently, each of the groups H, I, and J have only one member. It is possible that orthologues of these singletons remain to be found or may have become extinct because they failed to secure an essential and nonredundant role. Alternatively, these genes may be highly divergent orthologues and in fact belong in a single group.

What Defines a SOX Gene?

By convention, SOX proteins are more than 50% identical to SRY in the HMG domain. This definition is now

intronless structure has been reported. Abbreviations as for Fig. 1. Schematics were compiled using the following information: A—mo-SRY (Dubin and Ostrer, 1994); hu-SRY (Poulat *et al.*, 1997); rw-SRY (Foster *et al.*, 1992). B—mo-SOX1, -2, -3 (Collignon *et al.*, 1996); zf-SOX19 (Vriz and Lovell-Badge, 1995); mo-SOX2 (Yuan *et al.*, 1995); ch-SOX14 and -21 (Uchikawa *et al.*, 1999); ch-SOX21 (Rex *et al.*, 1997); mo-SOX1, -2, -3 (Kamachi *et al.*, 1999); mo-SOX2 (Kamachi *et al.*, 1998). C—hu-SOX22 (Jay *et al.*, 1997); tr-SOX24 (Kanda *et al.*, 1998); mo-SOX4 (van de Wetering *et al.*, 1993); mo-SOX11 (Hargrave *et al.*, 1997); ra-SOX11 (Kuhlbrodt *et al.*, 1998). D—mo-SOX5 (Hiraoka *et al.*, 1998a); moSOX13 (Kido *et al.*, 1998); ze-SOX6 (Komatsu *et al.*, 1996); mo-SOX5 and -6 (Lefebvre *et al.*, 1998); tr-SOX23 (Yamashita *et al.*, 1998). E—mo-SOX9 (Wright *et al.*, 1995); mo-SOX10 (Tani *et al.*, 1997); mo-SOX10 (Pusch *et al.*, 1998); mo-SOX9 (McDowall *et al.*, 1999); mo-SOX9 (Ng *et al.*, 1997); mo-SOX8 (Schepers *et al.*, 2000). F—mo-SOX17 (Kanai *et al.*, 1996); mo-SOX18 (Hosking *et al.*, 1995); mo-SOX7 (Taniguchi *et al.*, 1999). G—hu-SOX20 (Meyer *et al.*, 1996; Hiraoka *et al.*, 1998b); mo-SOX15 (Miyashita *et al.*, 1999).

inaccurate, with new *SOX* genes which do not conform to this rule being identified. For instance, the group H sequence hu-SOX30 is 48% and ce-SOXJ is only 46% identical to hu-SRY in the HMG domain. By comparison, mo-LEF1 is 24% identical to hu-SRY. Based on these figures, it seems that classification based on a strict 50% identity to SRY may not be a suitable indicator of SOX family membership. Reference to the SRY sequence was a historical, arbitrary, and, in retrospect, poor choice for such SOX family comparisons, since SRY has arisen only in the mammalian lineage and is clearly very divergent. It may be more appropriate to compare identity to another SOX sequence or to the SOX consensus sequence. However, even if this is done, a 50% cutoff value may be too stringent (e.g., human SOX30, 46% identity to the SOX consensus).

Our results provide an alternative criterion to define *SOX* genes using the conservation of key motifs within the HMG domain. The HMG domain sequence RPMNAF (positions 5–10) appears to be conserved for all SOX sequences, including those of groups H, I, and J, but not for the most closely related outgroups fu-MATA1, mo-LEF1, and mo-TCF1. However, this sequence is also present in a recently defined SOX-like gene in *Drosophila,* capicua (*cic*), which has apparent orthologues in *C. elegans* and humans, suggesting that this 6-amino-acid motif is insufficient to strictly define *SOX* genes (Jiménez *et al.,* 1999). The extended version, common to all non-SRY SOX members (RPMNAFMVW), appears to be the most reliable signature of the SOX family.

Invertebrate SOX Genes Confirm Vertebrate Groupings

We have identified invertebrate representatives of most of the SOX groups thus far identified. *C. elegans* proteins ce-SOXB1, ce-SOXB2, ce-SOXC, and ce-SOXD (known as COG-2; Hanna-Rose and Han, 1999) are associated with subgroups B1 and B2 and groups C and D, respectively. No *C. elegans* genes encoding proteins with homology to groups E, F, G, H, or I have been detected. An additional *C. elegans* SOX protein which cannot be assigned to any of the existing groups has been provisionally allocated it to a new group J. A putative *C. elegans* orthologue of LEF/TCF (ce-LEF/TCF) was also identified. *Drosophila* SOX proteins dr-SOXB1, dr-SOXB2.1, dr-SOXB2.2, dr-SOXB2.3, dr-SOXC, dr-SOXD, dr-SOXE, and dr-SOXF are associated with groups B1, B2, C, D, E, and F as indicated by their names. No *Drosophila* sequences have been found for groups G, H, I, or J. Similarly, sea urchin sequences se-SOXB1, se-SOXB2, and se-SOXD1 are associated with subgroups B1 and B2 and group D. It is entirely likely that additional sea urchin *SOX* genes remain to be identified—this genome is not yet completely sequenced. It should be noted that many of the *C. elegans* and some of the *Drosophila* sequences have been retrieved from genomic databases (see Fig. 1 and Table 1) and thus we have no evidence that they are functional. The exceptions are ce-SOXD (also known as COG-2), dr-SOXB2.1 (also known as *dichaete,* dr-SOX70D, or *fish-hook*), dr-SOXC (also known as SOX14), dr-SOXB1 (published as SOXNeuro; Crémazy *et al.,* 2000), and dr-SOXE (published as SOX100B; Loh and Russell, 2000).

For each invertebrate species examined, only one representative sequence has been identified for each group—the exception to this is group B (see below). This suggests that for each of the currently recognizable SOX groups, a single ancestral form existed before the origin of the vertebrate lineage. In contrast to this general trend, four group B SOX sequences have been identified in *Drosophila.* These include a single group B1 representative (*SoxB1*) along with three group B2 genes (*SoxB2.1, SoxB2.2,* and *SoxB2.3;* see Table 1). The three *SoxB2* genes are physically linked and it is possible that lineage-specific duplication and diversification have occurred in this case. In support of this possibility, the HMG boxes of *SoxB2.2* and *SoxB2.3* are approximately 50 and 70 kb downstream of the HMG box of *SoxB2.1* (*dichaete*) on chromosome 3 (AC015146). This relatively recent divergence is confirmed by maximum likelihood analysis—the four group B sequences clustered in 100% of analyses (Fig. 4).

Based on phylogenetic considerations, it is not possible to define the invertebrate orthologues of specific mammalian genes. It has been suggested that *dichaete* (here called SOXB2.1) is the *Drosophila* equivalent of mammalian SOX2 (Pevny and Lovell-Badge, 1997). Although mouse SOX2 can functionally substitute for *dichaete* (Sanchez Soriano and Russell, 1998) our analysis suggests that the *Drosophila* protein might more reasonably be considered to represent an ancestral form of the entire SOXB or SOXB2 group (Figs. 2 and 3, see below). *dichaete* is similar to vertebrate SOX2 sequences only in the HMG domain and a short C-terminal region which does not appear to be essential for its function (Sanchez Soriano and Russell, 1998; Mukherjee *et al.,* 2000), suggesting that rescue of the *dichaete* mutant might be possible also with other vertebrate group B proteins.

SOX Gene Duplication: A Model of Recent Evolution

Evolution is continuous, occurring both in the past and in the present. It seems likely that the physical linkage of *Drosophila* SOXB2.1, -B2.2, and -B2.3 is due to relatively recent duplication events. SOXB2.2 or -B2.3 were identified purely from genomic sequence and may represent *SOX* genes that have not yet secured a functional role and are redundant. It remains to be seen what, if any, role these genes play in *Drosophila* development, and these genes may prove a useful model for functional studies of SOX gene evolution.

It is interesting to note that there are two separate *Sox11* genes in zebrafish (Rimini *et al.,* 1999). The proteins they encode differ slightly in size (354 and 368 amino acids) and in one position in the HMG domain but are very similar to hu-SOX11 and ch-SOX11 throughout the protein, suggesting that they are both true orthologues (Rimini *et al.,* 1999). Two

distinct versions of *SOX9* have also been identified (B.-C. Chung, personal communication). The existence of two *SOX11* and two *SOX9* orthologues in zebrafish is consistent with evidence that a genome-wide duplication occurred in the teleost fish lineage at some point, leading in some cases to duplication of functional genes (Amores *et al.*, 1998). The duplication–degeneration–complementation model would predict that SOX11 and SOX9 functions will, in each case, be shared by the two orthologues (Force *et al.*, 1999). It will be of interest to determine whether other *SOX* genes are represented in two copies in zebrafish.

Ancient Intron Conservation in SOX Groups

The position and size of introns can be considered phylogenetic characters insomuch as they can lend support to theories of common ancestry. SOX groups D, E, and F have intron positions that are conserved between *Drosophila* and vertebrates, providing evidence that these represent ancient introns that were present before the divergence of the vertebrate lineage (Foster *et al.*, 1994; Wagner *et al.*, 1994; Kanai *et al.*, 1996; Wunderle *et al.*, 1996; Pingault *et al.*, 1998; Hanna-Rose and Han, 1999; Roose *et al.*, 1999; Taniguchi *et al.*, 1999; B. Hosking, personal communication; Fig, 1; Schepers *et al.*, 2000). Further, the position of the HMG box intron is conserved between groups D and F, suggesting that these groups had a common ancestor distinct from the other groups. Intron positions are also conserved between *C. elegans* group B and C *SOX* genes; however, these seem to have been lost during evolution as they are no longer present in *Drosophila*, sea urchin, or vertebrate orthologues.

SRY—Common Ancestor or Species-Specific Evolution?

SRY is unlike other *SOX* genes in that the various mammalian orthologues retain very little homology outside the HMG box region. This may result from its location on the rapidly evolving Y chromosome. There is evidence that non-HMG domain sequences may perform different mechanistic functions in various species and therefore it is possible that species-specific divergence has occurred (Dubin and Ostrer, 1994; Poulat *et al.*, 1997; Desclozeaux *et al.*, 1998; Bowles *et al.*, 1999). The rate of divergence within the HMG box is also high in comparison with other *SOX* genes—i.e., the observed differences in the various SRY proteins do not just relate to non-HMG domain regions, but reflect an unusually rapid divergence of the entire protein (Tucker and Lundrigan, 1993; Whitfield *et al.*, 1993; Miller *et al.*, 1995; Pamilo and O'Neill, 1997). In our analyses, group A is not robustly monophyletic using the HMG domain sequence data set. In some analyses SRY sequences cluster together to the exclusion of all other SOX sequences while in others group A is paraphyletic as has previously been reported (this study, Figs. 2 and 3; Soullier *et al.*, 1999). Monophyletic grouping would support the theory that the Y-chromosomal sex-determining *SOX* gene has not arisen independently in various mammalian lineages but that it arose once and then evolved rapidly. The inconsistency of clustering of SRY sequences most likely reflects the limitations of the use of HMG domain sequences in some situations.

A recent study addressed specifically the question of the evolution of the mammalian Y-chromosomal *SOX* gene, *SRY*, from the X-linked gene *SOX3* (Katoh and Miyata, 1999). Using a sophisticated heuristic maximum likelihood method, those workers were able to construct a tree in which the appearance of SRY was consistent with its mammalian origin and in which the various SRY genes were consistently monophyletic. They suggest, and we agree, that the aberrant behavior of SRY in phylogenetic analyses using less rigorous methods of analysis is likely to relate to its remarkably high evolutionary rate.

Relationship of Phylogeny to SOX Protein Function

Only a relatively small number of *SOX* genes and proteins have been well characterized but some functional data have emerged from expression and molecular studies, targeted gene deletions, existing mouse mutations, and human disease mapping (Table 2). For some *SOX* groups, conservation of sequence and structure does correlate with similarity in function. In other instances such correlation is not apparent, with closely related proteins having adopted diverse functions.

Members of group B1 show functional similarity, all being involved in central nervous system (CNS) development and, in particular, in regulation of the neuronal phenotype (Collignon *et al.*, 1996). There is some evidence for functional redundancy among vertebrate SOX1, SOX2, and SOX3, although it appears that these proteins have also taken on additional unique roles (Ambrosetti *et al.*, 2000; Kamachi *et al.*, 1998; Nishiguchi *et al.*, 1998). This may represent an example of the duplication–degeneration–complementation model discussed above (Force *et al.*, 1999). Group B2 genes (*SOX14* and *-21*) are also expressed in the CNS but, in contrast to those of group B1, these genes appear to act as repressors rather than activators (Uchikawa *et al.*, 1999). A more specific role has been postulated for SOX14 than for SOX21—the specification of a particular subset of neurons rather than neuronal development in general (Hargrave *et al.*, 2000). Group D (*SOX5*, *-6*, *-13*, and *-23*) includes some members which appear quite similar in expression profile and potential function. *SOX5* and *SOX6* both appear to be involved in spermatogenesis and chondrogenesis, with potential for redundancy of function (Connor *et al.*, 1995; Lefebvre *et al.*, 1998). This group is characterized by the presence of a leucine-zipper motif and all vertebrate members have been shown to homo- and/or heterodimerize (Table 2). In contrast, members of group E (*SOX8*, *-9*, and *-10*) are very similar in structure, yet their expression patterns and their involvement in a range of human diseases indicate that they have diverged consider-

TABLE 2
Functions of *Sox* Genes in Development

Group	Gene	Functional data	Targets	References
A	*Sry*	Mammalian testis-determining gene.		Koopman *et al.*, 1991
B1	*SoxB1*	Expressed in early developing CNS in *Drosophila* (also called *SoxNeuro*). Neuroectoderm expression may be controlled by the zygotic dorsoventral patterning genes (*dpp, sog, brk, twi*).		Crémazy *et al.*, 2000
	Sox1	Expressed in mouse embryonic CNS, lens; putative neural differentiation gene. Knockout mice show microphthalmia, cataracts, seizures.	*γ-crystallin*	Nishiguchi *et al.*, 1998; Pevny *et al.*, 1998
	Sox2	Expressed in pluripotent lineages of preimplantation mouse embryo, developing CNS, lens.	*Fgf4*, *β-* and *γ-crystallin*	Fraidenraich *et al.*, 1998; Kamachi *et al.*, 1998
	Sox3	Expressed in mouse embryonic CNS, lens. No direct functional data. May act redundantly with *Sox1* and *-2*.		Collignon *et al.*, 1996; Kamachi *et al.*, 1998
	Sox19	Expressed in early CNS and lens in zebrafish. No direct functional data.		Vriz *et al.*, 1996
B2	*SoxB2.1*	Also called *dichaete*. Involved in early embryo segmentation and brain development and is required for the correct differentiation of the hindgut.	*labial, zfh-2, wingless, engrailed*	Sanchez-Soriano and Russell, 2000
	Sox14	May specify a subset of ventral interneurons in the spinal cord and neuronal subtypes in the brain. Transcriptional repressor.		Hargrave *et al.*, 2000; Uchikawa *et al.*, 1999
	Sox21	Expressed in developing CNS. Transcriptional repressor.		Uchikawa *et al.*, 1999
C	*Sox4*	Roles in cardiac outflow tract development and B-cell development revealed by phenotype of knockout mice.		Schilham *et al.*, 1996
	Sox11	Expressed in maturing neurons in CNS, also in PNS and sites of epithelial–mesenchymal interaction. Involved in oligodendrocyte differentiation.		Hargrave *et al.*, 1997; Kuhlbrodt *et al.*, 1998
	Sox12	Not known.		
	Sox22	Expressed in CNS and many other tissues. No direct functional data.		Jay *et al.*, 1997
	Sox24	Expressed in oocytes. Not studied in detail. No direct functional data.		Kanda *et al.*, 1998
D	*SoxD*	Also called *COG-2* (connection of gonad); regulates late-stage uterine seam cell differentiation and fusion.		Hanna-Rose and Han, 1999
	Sox5	Expressed during spermatogenesis and chondrogenesis. Involvement in chondrogenesis supported by interaction with SOX9. Homo- and heterodimerization with SOX6.		Denny *et al.*, 1992; Lefebvre *et al.*, 1998
	Sox6	Expressed during spermatogenesis and chondrogenesis, and in CNS. Involvement in chondrogenesis supported by interaction with SOX9. May act redundantly with SOX5.		Connor *et al.*, 1995; Lefebvre *et al.*, 1998
	Sox13	Expressed in developing arteries and thymus, and widely in adult human tissues. Homodimerization.		Kasimiotis *et al.*, 2000; Roose *et al.*, 1998
	Sox23	Expressed in embryonic ovary and brain. Homodimerization.		Yamashita *et al.*, 1998
E	*SoxE*	Also called *Sox100B*. Expressed in large intestinal cells, in basophilic cells in the midgut, in the Malpighian tubules, and at the posterior cap of gonadal mesoderm.		Loh and Russell, 2000
	Sox8	Expressed in fetal CNS, brain, branchial arches, limb, heart, dorsal root ganglia, and testes. Deleted in ATR-16 patient. No direct functional data.		Pfeifer *et al.*, 2000; Schepers *et al.*, 2000
	Sox9	Key regulator of chondrogenesis and sex determination, roles also in heart, kidney, and brain development, as revealed by phenotype of campomelic dysplasia patients. Cells lacking SOX9 cannot form chondrocytes.	*Col2a1*, *aggrecan*, other matrix protein genes, *Amh*	Bell *et al.*, 1997; Bi *et al.*, 1999; de Santa Barbara *et al.*, 1998; Sekiya *et al.*, 1997
	Sox10	Regulator of neural crest cell differentiation. Mutation leads to neurocristopathy in humans and mice.	P_0	Peirano *et al.*, 2000; Pingault *et al.*, 1998; Southard-Smith *et al.*, 1998

TABLE 2—*Continued*

Group	Gene	Functional data	Targets	References
F	*Sox7*	Not known.		
	Sox17	Regulator of endoderm development and spermatogenesis. Role in endoderm induction shown by protein function interference in *Xenopus.*		Hudson *et al.*, 1997; Kanai *et al.*, 1996
	Sox18	Blood vessel and hair follicle development, as demonstrated by mutations in *ragged* mice.		Pennisi *et al.*, 2000
G	*Sox20*	Expressed in fetal testes but expression and function not studied in detail.		Hiraoka *et al.*, 1998
	Sox15	Inhibitor of myoblast differentiation, as revealed by experiments involving cultured myoblasts.		Beranger *et al.*, 2000
	Sox16	Not known.		
H	*Sox30*	Expressed in germ cells of embryonic testes.		Osaki *et al.*, 1999
I	*Sox31*	Expressed in late blastula, gastrula, and neural tissues. Dominant negative experiments in *Xenopus* demonstrate role in neural induction.		Mizuseki *et al.*, 1998

ably in function. In all cases, however, it appears that these proteins act to induce cell-type differentiation (for details and references see Table 2). It remains to be seen whether conservation of SOX protein function will be revealed at the molecular level once mechanistic details of their regulatory roles, including involvement of accessory or partner proteins, are more clearly understood.

CONCLUDING REMARKS

We have conducted an exhaustive phylogenetic study of the SOX family in order to examine its evolutionary history and to clarify relationships among its various members. We based our approach on the hypothesis that the HMG domain alone should be a good marker of the history of evolution of the family providing that it has diverged sufficiently (i.e., it contains enough phylogenetic information) but not to the point at which information has been lost because it was "written over." Based on HMG domain sequence alone we find, as have others, that vertebrate SOX family members fall into clear groups. In general, the previously suggested groupings are supported. Two new groups have been assigned to accommodate SOX proteins which do not fall into any of the established groups. The inclusion of invertebrate sequences provides added insight into the evolution of the SOX family and confirms the groups defined for vertebrate *SOX* genes. Based on HMG domain sequence analysis of newly identified and previously defined invertebrate *SOX* genes, it appears that most of the vertebrate SOX groups are represented by a single gene in simpler organisms. This is true for all groups apart from group B in which lineage-specific duplication seems to have occurred. We also examined the non-HMG domain sequences of vertebrate SOX proteins. Despite relatively poor primary sequence conservation, we note strong conservation in terms of structural features, motifs, and functional domains within each group. Based on this analysis we find that for vertebrate SOX proteins, the groupings defined by HMG domain sequences are upheld when full-length sequences and genomic structure are considered.

The SOX family shows evidence of evolution by both slow divergence and the recruitment of preexisting functional elements. The HMG domain, the ancestral motif which forms the core of SOX family proteins, would be expected to have gradually accumulated sequence changes under the selection pressure of retaining sequence-specific DNA-binding function. In contrast, variability of SOX proteins outside of the HMG domain indicates that sudden and stochastic evolutionary changes must also have occurred apparently via co-option of functional domains and motifs resulting in the formation of "evolutive chimeras" (Ohno, 1970; Patthy, 1991, 1994; Holland *et al.*, 1994). Such changes may, at least to some extent, mark the origin of the various SOX groups. Subsequent to these major changes, additional duplication and divergence events must have occurred, resulting in the range of SOX proteins present in vertebrates today.

ACKNOWLEDGMENTS

We thank Robin Lovell-Badge, Jenny Graves, and Claire Larroux for critically reading the manuscript, and Steve Russell, Franck Girard, Bon-Chu Chung, and Brett Hosking for access to unpublished data. P.K. is an Australian Research Council Australian Senior Research Fellow.

REFERENCES

Aguinaldo, A. M., Turbeville, J. M., Linford, L. S., Rivera, M. C., Garey, J. R., Raff, R. A., and Lake, J. A. (1997). Evidence for a clade of nematodes, arthropods and other moulting animals. *Nature* **387,** 489–493.

Amores, A., Force, A., Yan, Y.-L., Joly, L., Amemiya, C., Fritz, A., Ho, R. K., Langeland, J., Prince, V., Wang, Y. L., Westerfield, M., Ekker, M., and Postlethwait, J. H. (1998). Zebrafish *hox* clusters and vertebrate genome evolution. *Science* **282,** 1711–1714.

Bell, D. M., Leung, K. K. H., Wheatley, S. C., Ng, L. J., Zhou, S., Ling, K. W., Sham, M. H., Koopman, P., Tam, P. P. L., and Cheah, K. S. E. (1997). SOX9 directly regulates the type-II collagen gene. *Nat. Genet.* **16,** 174–178.

Beranger, F., Mejean, C., Moniot, B., Berta, P., and Vandromme, M. (2000). Muscle differentiation is antagonized by SOX15, a new member of the SOX protein family. *J. Biol. Chem.* **275,** 16103–16109.

Bi, W., Deng, J. M., Zhang, Z., Behringer, R. R., and de Crombrugghe, B. (1999). Sox9 is required for cartilage formation. *Nat. Genet.* **22,** 85–89.

Bowles, J., Cooper, L., Berkman, J., and Koopman, P. (1999). Sry requires a CAG repeat domain for male sex determination in Mus musculus. *Nat. Genet.* **22,** 405–408.

Collignon, J., Sockanathan, S., Hacker, A., Cohen-Tannoudji, M., Norris, D., Rastan, S., Stevanovic, M., Goodfellow, P. N., and Lovell-Badge, R. (1996). A comparison of the properties of Sox-3 with Sry and two related genes, Sox-1 and Sox-2. *Development* **122,** 509–520.

Connor, F., Wright, E., Denny, P., Koopman, P., and Ashworth, A. (1995). The Sry-related HMG box-containing gene Sox6 is expressed in the adult testis and developing nervous system of the mouse. *Nucleic Acids Res.* **23,** 3365–3372.

Crémazy, F., Berta, P., and Girard, F. (2000). *SoxNeuro,* a new *Drosophila Sox* gene expressed in the developing central nervous system. *Mech. Dev.* **93,** 215–219.

de Santa Barbara, P., Bonneaud, N., Boizet, B., Desclozeaux, M., Moniot, B., Südbeck, P., Scherer, G., Poulat, F., and Berta, P. (1998). Direct interaction of SRY-related protein SOX9 and steroidogenic factor 1 regulates transcription of the human anti-Müllerian hormone gene. *Mol. Cell. Biol.* **18,** 6653–6665.

Denny, P., Swift, S., Brand, N., Dabhade, N., Barton, P., and Ashworth, A. (1992). A conserved family of genes related to the testis determining gene, *SRY*. *Nucleic Acids Res.* **20,** 2887.

Desclozeaux, M., Poulat, F., de Santa Barbara, P., Capony, J.-P., Turowski, P., Jay, P., Mejean, C., Miniot, B., Boizet, B., and Berta, P. (1998). Phosphorylation of an N-terminal motif enhances DNA-binding activity of the human SRY protein. *J. Biol. Chem.* **273,** 7988–7995.

Dubin, R. A., and Ostrer, H. (1994). Sry is a transcriptional activator. *Mol. Endocrinol.* **8,** 1182–1192.

Fitch, W. M. (1981). A non-sequential method for constructing trees and hierarchical classifications. *J. Mol. Evol.* **18,** 30–37.

Force, A., Lynch, M., Pickett, F. B., Amores, A., Yan, Y. L., and Postlethwait, J. (1999). Preservation of duplicate genes by complementary degenerative mutations. *Genetics* **151,** 1531–1545.

Foster, J. W., Brennan, F. E., Hampikian, G. K., Goodfellow, P. N., Sinclair, A. H., Lovell-Badge, R., Selwood, L., Renfree, M. B., Cooper, D. W., and Graves, J. A. M. (1992). Evolution of sex determination and the Y chromosome: *SRY*-related sequences in marsupials. *Nature* **359,** 531–532.

Foster, J. W., Dominguez-Steglich, M. A., Guioli, S., Kwok, C., Weller, P. A., Weissenbach, J., Mansour, S., Young, I. D., Goodfellow, P. N., Brook, J. D., and Schafer, A. J. (1994). Campomelic dysplasia and autosomal sex reversal caused by mutations in an *SRY*-related gene. *Nature* **372,** 525–530.

Fraidenraich, D., Lang, R., and Basilico, C. (1998). Distinct regulatory elements govern *Fgf4* gene expression in the mouse blastocyst, myotomes, and developing limb. *Dev. Biol.* **204,** 197–209.

Grosschedl, R., Giese, J., and Pagel, J. (1994). HMG domain proteins: Architectural elements in the assembly of nucleoprotein structures. *Trends Genet.* **10,** 94–100.

Gubbay, J., Collignon, J., Koopman, P., Capel, B., Economou, A., Münsterberg, A., Vivian, N., Goodfellow, P., and Lovell-Badge, R. (1990). A gene mapping to the sex-determining region of the mouse Y chromosome is a member of a novel family of embryonically expressed genes. *Nature* **346,** 245–250.

Hanna-Rose, W., and Han, M. (1999). COG-2, a sox domain protein necessary for establishing a functional vulval–uterine connection in Caenorhabditis elegans. *Development* **126,** 169–179.

Hargrave, M., Karunaratne, A., Cox, L., Wood, S., Koopman, P., and Yamada, T. (2000). The HMG box transcription factor gene Sox14 marks a novel subset of ventral interneurons and is regulated by sonic hedgehog. *Dev. Biol.* **219,** 142–53.

Hargrave, M., Wright, E., Kun, J., Emery, J., Cooper, L., and Koopman, P. (1997). Expression of the Sox11 gene in mouse embryos suggests roles in neuronal maturation and epithelio-mesenchymal induction. *Dev. Dyn.* **210,** 79–86.

Hiraoka, Y., Ogawa, M., Sakai, Y., Kido, S., and Aiso, S. (1998a). The mouse Sox5 gene encodes a protein containing the leucine zipper and the Q box. *Biochim. Biophys. Acta* **1399,** 40–46.

Hiraoka, Y., Ogawa, M., Sakai, Y., Taniguchi, K., Fujii, T., Umezawa, A., Hata, J., and Aiso, S. (1998b). Isolation and expression of a human SRY-related cDNA hSOX20. *Biochim. Biophys. Acta* **1396,** 132–137.

Holland, P. W. H., Garcia-Fernandez, J., Williams, N. A., and Sidow, A. (1994). Gene duplications and the origins of vertebrate development. *Development Suppl.* **1994,** 125–133.

Hosking, B. M., Muscat, G. E. O., Koopman, P. A., Dowhan, D. H., and Dunn, T. L. (1995). *Trans*-activation and DNA-binding properties of the transcription factor, Sox-18. *Nucleic Acids Res.* **23,** 2626–2628.

Hudson, C., Clements, D., Friday, R. V., Stott, D., and Woodland, H. R. (1997). Xsox17alpha and -beta mediate endoderm formation in Xenopus. *Cell* **91,** 397–405.

Jay, P., Sahly, I., Goze, C., Taviaux, S., Poulat, F., Couly, G., Abitbol, M., and Berta, P. (1997). Sox22 is a new member of the Sox gene family, mainly expressed in human nervous tissue. *Hum. Mol. Genet.* **6,** 1069–1077.

Jiménez, G., Guichet, A., Ephrussi, A., and Casanova, J. (1999). Relief of gene repression by Torso RTK signaling: Role of *capicua* in *Drosophila* terminal and dorsoventral patterning. *Genes Dev.* **14,** 224–231.

Kamachi, Y., Cheah, K. S., and Kondoh, H. (1999). Mechanism of regulatory target selection by the SOX high-mobility-group domain proteins as revealed by comparison of SOX1/2/3 and SOX9. *Mol. Cell. Biol.* **19,** 107–120.

Kamachi, Y., Uchikawa, M., Collignon, J., Lovell-Badge, R., and Kondoh, H. (1998). Involvement of Sox1, 2 and 3 in the early and subsequent molecular events of lens induction. *Development* **125,** 2521–2532.

Kanai, Y., Kanai-Azuma, M., Noce, T., Saido, T. C., Shiroishi, T., Hayashi, Y., and Yazaki, K. (1996). Identification of two Sox17 messenger RNA isoforms, with and without the high mobility group box region, and their differential expression in mouse spermatogenesis. *J. Cell Biol.* **133,** 667–681.

Kanda, H., Kojima, M., Miyamoto, N., Ito, M., Takamatsu, N., Yamashita, S., and Shiba, T. (1998). Rainbow trout Sox24, a novel

member of the Sox family, is a transcriptional regulator during oogenesis. *Gene* **211,** 251–257.

Kasimiotis, H., Myers, M. A., Argentaro, A., Mertin, S., Fida, S., Ferraro, T., Olsson, J., Rowley, M. J., and Harley, V. R. (2000). Sex-determining region Y-related protein SOX13 is a diabetes autoantigen expressed in pancreatic islets. *Diabetes* **49,** 555–561.

Katoh, K., and Miyata, T. (1999). A heuristic approach of maximum likelihood method for inferring phylogenetic tree and an application to the mammalian *SOX-3* origin of the testis-determining gene *SRY*. *FEBS Lett.* **463,** 129–132.

Kido, S., Hiraoka, Y., Ogawa, M., Sakai, Y., Yoshimura, Y., and Aiso, S. (1998). Cloning and characterization of mouse *mSox13* cDNA. *Gene* **208,** 201–206.

Komatsu, N., Hiraoka, Y., Shiozawa, M., Ogawa, M., and Aiso, S. (1996). Cloning and expression of Xenopus laevis xSOX12 cDNA. *Biochim. Biophys. Acta* **1305,** 117–119.

Koopman, P., Gubbay, J., Vivian, N., Goodfellow, P., and Lovell-Badge, R. (1991). Male development of chromosomally female mice transgenic for *Sry*. *Nature* **351,** 117–121.

Kuhlbrodt, K., Herbarth, B., Sock, E., Enderich, J., Hermans-Borgmeyer, I., and Wegner, M. (1998). Cooperative function of POU proteins and SOX proteins in glial cells. *J. Biol. Chem.* **273,** 16050–16057.

Laudet, V., Stehelin, D., and Clevers, H. (1993). Ancestry and diversity of the HMG box superfamily. *Nucleic Acids Res.* **21,** 2493–2501.

Lefebvre, V., Li, P., and de Crombrugghe, B. (1998). A new long form of Sox5 (L-Sox5), Sox6 and Sox9 are coexpressed in chondrogenesis and cooperatively activate the type II collagen gene. *EMBO J.* **17,** 5718–5733.

Loh, S. Y. H., and Russell, S. (2000). A *Drosophila* group E *Sox* gene is dynamically expressed in the embryonic alimentary canal. *Mech. Dev.* **93,** 185–188.

McDowall, S., Argentaro, A., Ranganathan, S. W. P., Mertin, S., Mansour, S., Tolmie, J., and Harley, V. (1999). Functional and structural studies of wild type SOX9 and mutations causing campomelic dysplasia. *J. Biol. Chem.* **274,** 24023–24030.

Meyer, J., Wirth, J., Held, M., Schempp, W., and Scherer, G. (1996). SOX20, a new member of the SOX gene family, is located on chromosome 17p13. *Cytogenet. Cell Genet.* **72,** 246–249.

Miller, K. E., Lundrigan, B. L., and Tucker, P. K. (1995). Length variation of CAG repeats in Sry across populations of *Mus domesticus*. *Mamm. Genome* **6,** 206–208.

Miyashita, A., Shimizu, N., Endo, N., Hanyuu, T., Ishii, N., Ito, K., Itoh, Y., Shirai, M., Nakajima, T., Odani, S., and Kuwano, R. (1999). Five different genes, Eif4a1, Cd68, Supl15h, Sox15 and Fxr2h, are clustered in a 40 kb region of mouse chromosome 11. *Gene* **237,** 53–60.

Mizuseki, K., Kishi, M., Shiota, K., Nakanishi, S., and Sasai, Y. (1998). SoxD: An essential mediator of induction of anterior neural tissues in Xenopus embryos. *Neuron* **21,** 77–85.

Mukherjee, A., Shan, X., Mutsuddi, M., Ma, Y., and Nambu, J. R. (2000). The *Drosophila Sox* gene, *fish-hook,* is required for postembryonic development. *Dev. Biol.* **217,** 91–106.

Ng, L. J., Wheatley, S., Muscat, G. E., Conway-Campbell, J., Bowles, J., Wright, E., Bell, D. M., Tam, P. P., Cheah, K. S., and Koopman, P. (1997). SOX9 binds DNA, activates transcription, and coexpresses with type II collagen during chondrogenesis in the mouse. *Dev. Biol.* **183,** 108–121.

Nishiguchi, S., Wood, H., Kondoh, H., Lovell-Badge, R., and Episkopou, V. (1998). *Sox1* directly regulates the gamma-crystallin genes and is essential for lens development in mice. *Genes Dev.* **12,** 776–781.

Ohno, S. (1970). "Evolution by Gene Duplication." Springer-Verlag, Heidelberg.

Osaki, E., Nishina, Y., Inazawa, J. N. G. C., Gilbert, D. J., Jenkins, N. A., Ohsugi, M., Tezuka, T., Yoshida, M., and Semba, K. (1999). Identification of a novel Sry-related gene and its germ cell-specific expression. *Nucleic Acids Res.* **27,** 2503–2510.

Pamilo, P., and O'Neill, R. J. (1997). Evolution of the Sry genes. *Mol. Biol. Evol.* **14,** 49–55.

Patthy, L. (1991). Modular exchange principles in proteins. *Curr. Opin. Struct. Biol.* **1,** 351–361.

Patthy, L. (1994). Introns and exons. *Curr. Opin. Struct. Biol.* **4,** 383–392.

Peirano, R. I., Goerich, D. E., Riethmacher, D., and Wegner, M. (2000). Protein zero gene expression is regulated by the glial transcription factor Sox10. *Mol. Cell. Biol.* **20,** 3198–3209.

Pennisi, D., Gardner, J., Chambers, D., Hosking, B., Peters, J., Muscat, G., Abbott, C., and Koopman, P. (2000). Mutations in Sox18 underlie cardiovascular and hair follicle defects in ragged mice. *Nat. Genet.* **24,** 434–437.

Pevny, L. H., and Lovell-Badge, R. (1997). SOX genes find their feet. *Curr. Opin. Genet. Dev.* **7,** 338–344.

Pevny, L. H., Sockanathan, S., Placzek, M., and Lovell-Badge, R. (1998). A role for SOX1 in neural determination. *Development* **125,** 1967–1978.

Pfeifer, D., Poulat, F., Holinski-Feder, E., Kooy, F., and Scherer, G. (2000). The *SOX8* gene is located within 700 kb of the tip of chromosome 16p and is deleted in a patient with ATR-16 syndrome. *Genomics* **63,** 108–116.

Pingault, V., Bondurand, N., Kuhlbrodt, K., Goerich, D. E., Prehu, M. O., Puliti, A., Herbarth, B., Hermans-Borgmeyer, I., Legius, E., Matthijs, G., Amiel, J., Lyonnet, S., Ceccherini, I., Romeo, G., Smith, J. C., Read, A. P., Wegner, M., and Goossens, M. (1998). SOX10 mutations in patients with Waardenburg–Hirschsprung disease. *Nat. Genet.* **18,** 171–173.

Poulat, F., de Santa Barbara, P., Desclozeaux, M., Soullier, S., Moniot, B., Bonneaud, N., Boizet, B., and Berta, P. (1997). The human testis determining factor SRY binds a nuclear factor containing PDZ protein interaction domains. *J. Biol. Chem.* **272,** 7167–7172.

Pusch, C., Hustert, E., Pfeifer, D., Sudbeck, P., Kist, R., Roe, B., Wang, Z., Balling, R., Blin, N., and Scherer, G. (1998). The SOX10/Sox10 gene from human and mouse: Sequence, expression, and transactivation by the encoded HMG domain transcription factor. *Hum. Genet.* **103,** 115–123.

Rex, M., Uwanogho, D. A., Orme, A., Scotting, P. J., and Sharpe, P. T. (1997). cSox21 exhibits a complex and dynamic pattern of transcription during embryonic development of the chicken central nervous system. *Mech. Dev.* **66,** 39–53.

Rimini, R., Beltrame, M., Argenton, F., Szymczak, D., Cotelli, F., and Bianchi, M. E. (1999). Expression patterns of zebrafish *sox11A*, *sox11B* and *sox21*. *Mech. Dev.* **89,** 167–171.

Roose, J., Korver, W., de Boer, R., Kuipers, J., Hurenkamp, J., and Clevers, H. (1999). The Sox-13 gene: Structure, promoter characterization, and chromosomal localization. *Genomics* **57,** 301–305.

Roose, J., Korver, W., Oving, E., Wilson, A., Wagenaar, G., Markman, M., Lamers, W., and Clevers, H. (1998). High expression of the HMG box factor sox-13 in arterial walls during embryonic development. *Nucleic Acids Res.* **26,** 469–476.

Sanchez-Soriano, N., and Russell, S. (1998). The *Drosophila* SOX-domain protein Dichaete is required for the development of the central nervous system midline. *Development* **125,** 3989–3996.

Schepers, G., Bullejos, M., Hoskings, B., and Koopman, P. (2000). Cloning and characterization of the *Sry*-related transcription factor gene, *Sox8*. *Nucleic Acids Res.* **28,** 1473–1480.

Schilham, M. W., Oosterwegel, M. A., Moerer, P., Ya, J., de Boer, P. A., van de Wetering, M., Verbeek, S., Lamers, W. H., Kruisbeek, A. M., Cumano, A., and Clevers, H. (1996). Defects in cardiac outflow tract formation and pro-B-lymphocyte expansion in mice lacking Sox-4. *Nature* **380,** 711–714.

Sekiya, I., Koopman, P., Watanabe, H., Ezura, Y., Yamada, Y., and Noda, M. (1997). SOX9 enhances aggrecan gene expression via the promoter region containing a single HMG box sequence in a chondrogenic cell line, TC6. *J. Bone Miner. Res.* **12,** P222.

Soullier, S., Jay, P., Poulat, F., Vanacker, J.-M., Berta, P., and Laudet, V. (1999). Diversification pattern of the HMG and SOX family members during evolution. *J. Mol. Evol.* **48,** 517–527.

Southard-Smith, E. M., Kos, L., and Pavan, W. J. (1998). Sox10 mutation disrupts neural crest development in Dom Hirschsprung mouse model. *Nat. Genet.* **18,** 60–64.

Tani, M., Shindo-Okada, N., Hashimoto, Y., Shiroishi, T., Takenoshita, S., Nagamachi, Y., and Yokota, J. (1997). Isolation of a novel Sry-related gene that is expressed in high-metastatic K-1735 murine melanoma cells. *Genomics* **39,** 30–37.

Taniguchi, K., Hiraoka, Y., Ogawa, M., Sakai, Y., Kido, S., and Aiso, S. (1999). Isolation and characterization of a mouse SRY-related cDNA, mSox7. *Biochim. Biophys. Acta* **1445,** 225–231.

Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994). CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22,** 4673–4680.

Tucker, P. K., and Lundrigan, B. L. (1993). Rapid evolution of the sex determining locus in Old World mice and rats. *Nature* **364,** 715–717.

Uchikawa, M., Kamachi, Y., and Kondoh, H. (1999). Two distinct subgroups of Group B Sox genes for transcriptional activators and repressors: Their expression during embryonic organogenesis of the chicken. *Mech. Dev.* **84,** 103–120.

van de Wetering, M., and Clevers, H. (1993). Sox 15, a novel member of the murine Sox family of HMG box transcription factors. *Nucleic Acids Res.* **21,** 1669.

van de Wetering, M., Oosterwegel, M., van Norren, K., and Clevers, H. (1993). Sox-4, an Sry-like HMG box protein, is a transcriptional activator in lymphocytes. *EMBO J.* **12,** 3847–3854.

Vriz, S., Joly, C., Boulekbache, H., and Condamine, H. (1996). Zygotic expression of the zebrafish *Sox-19*, an HMG box-containing gene, suggests an involvement in central nervous system development. *Mol. Brain Res.* **40,** 221–228.

Vriz, S., and Lovell-Badge, R. (1995). The zebrafish Zf-Sox 19 protein—A novel member of the Sox family which reveals conserved motifs outside of the DNA-binding domain. *Gene* **153,** 275–276.

Wagner, T., Wirth, J., Meyer, J., Zabel, B., Held, M., Zimmer, J., Pasantes, J., Bricarelli, F. D., Keutel, J., Hustert, E., Wolf, U., Tommerup, N., Schempp, W., and Scherer, G. (1994). Autosomal sex reversal and campomelic dysplasia are caused by mutations in and around the *SRY*-related gene *SOX9*. *Cell* **79,** 1111–1120.

Wegner, M. (1999). From head to toes: The multiple facets of Sox proteins. *Nucleic Acids Res.* **27,** 1409–1420.

Whitfield, L. S., Lovell-Badge, R., and Goodfellow, P. N. (1993). Rapid sequence evolution of the mammalian sex determining gene *SRY*. *Nature* **364,** 713–715.

Wright, E., Hargrave, M. R., Christiansen, J., Cooper, L., Kun, J., Evans, T., Gangadharan, U., Greenfield, A., and Koopman, P. (1995). The *Sry*-related gene *Sox-9* is expressed during chondrogenesis in mouse embryos. *Nature Genet.* **9,** 15–20.

Wright, E. M., Snopek, B., and Koopman, P. (1993). Seven new members of the *Sox* gene family expressed during mouse development. *Nucleic Acids Res.* **21,** 744.

Wunderle, V. M., Critcher, R., Ashworth, A., and Goodfellow, P. N. (1996). Cloning and characterization of SOX5, a new member of the human SOX gene family. *Genomics* **36,** 354–358.

Yamashita, A., Suzuki, S., Fujitani, K., Kojima, M., Kanda, H., Ito, M., Takamatsu, N., Yamashita, S., and Shiba, T. (1998). cDNA cloning of a novel rainbow trout SRY-type HMG box protein, rtSox23 and its functional analysis. *Gene* **209,** 193–200.

Yuan, H. B., Corbi, N., Basilico, C., and Dailey, L. (1995). Developmental-specific activity of the FGF-4 enhancer requires the synergistic action of Sox2 and Oct-3. *Genes Dev.* **9,** 2635–2645.

Received for publication June 20, 2000
Revised August 7, 2000
Accepted August 7, 2000
Published online October 13, 2000