Developmental Cell, Vol. 3, 167–170, August, 2002, Copyright ©2002 by Cell Press

Letter to the Editor

Twenty Pairs of *Sox*: Extent, Homology, and Nomenclature of the Mouse and Human *Sox* Transcription Factor Gene Families

The genomics era is characterized by the rapid identification of genes, gene fragments, and gene paralogs within species, and orthologs between species. The highly conserved HMG box that defines the *Sox* family of developmental transcription factor genes (Bowles et al., 2000) has been exploited in many laboratories to identify approximately 30 vertebrate and over a dozen invertebrate *Sox* genes or gene fragments. However, the actual number of *Sox* genes in the mouse and human genomes has remained unknown. With the availability of complete drafts of these genome sequences, we can now determine the precise number of *Sox* genes, assign names, and identify orthologs. This in turn provides a basis for similar efforts in other model organisms as sequence data become available.

In this analysis, we examined all published *Sox* sequences, and recent releases of the human and mouse genome sequence from the relevant public sequencing consortia (Mouse Genome Assembly v3, 2 May 2002, http://www.ensembl.org and Human Genome Assembly build 29, 5 April 2002, http://www.ncbi.nlm.nih.gov/genome/seq) and from Celera Genomics (Celera Discovery System, indexed 2 May 2002, http://www.celera.com) (Lander et al., 2001; Venter et al., 2001). SOX proteins other than SRY were defined by the presence of the HMG domain signature sequence RPMNAFMVW (Bowles et al., 2000). Orthologous *Sox* genes were identified by sequence similarity and chromosomal location within regions of conserved synteny, determined by comparison of gene order.

The mouse and human genomes were found to contain 20 orthologous pairs of *Sox* genes (Table 1). The paired *Sox* genes show identical genomic organization, with the exception of *Sox6* and *Sox13*, which varied between mouse and human by the loss or gain of an intron in the untranslated region. No novel *Sox* genes were identified.

We and others have previously noted that in *Drosophila melanogaster* and *Caenorhabditis elegans*, the number of *Sox* genes is relatively small (five and eight, respectively; Bowles et al., 2000; Crémazy et al., 2001) and that a single gene in these organisms typically corresponds to a group or subgroup of *Sox* genes in vertebrates. Further, it is conspicuous that 9 of the 20 human/mouse *Sox* genes are single exons and that these are distributed evenly throughout the genome in both species. These properties likely reflect expansion of this ancient gene family via nontandem duplication and retroposition (Ohno, 1970). We have found evidence of tandem duplication in only two cases, where fragments similar to parts of human *SOX17* and *-30* lie adjacent to these genes (see below).

Several human and mouse genes predicted in previous studies from partial PCR amplification of the HMG box are absent from the genome. These sequences probably represent amplification or sequencing errors and most likely correspond to some of the 20 bona fide *Sox* genes (Table 2). One of these fragments, originally designated human *SOX29*, shows significant sequence similarity to *SOX5*, but has a 2 base pair deletion in the HMG box, suggesting that it may correspond to a *SOX5* pseudogene (Wunderle et al., 1996; Crémazy et al., 1998). We find that this gene lacks ESTs and introns, confirming that it is a pseudogene, which we name *SOX5P* (Table 2). Other fragments corresponding to parts of *SOX2*, *-17*, *-20*, and *-30* can be found in the human genome, but these are short, lack an HMG box, and contain gaps, insertions, or in-frame stop codons, indicating that they are not segments of functional *Sox* genes (Table 2). No pseudogenes or pseudogene fragments were found in the mouse genome.

Sequence similarity between mouse *Sox12* and human *SOX22* has been reported previously (Jay et al., 1997; Bowles et al., 2000). The availability of the complete coding sequence reveals extensive non-HMG box sequence homology between these two genes. This homology, and the chromosomal location of both genes within regions of conserved synteny, confirm that *Sox12* and *SOX22* are orthologs. Similar observations indicate that *SOX20* and *Sox15* (Bowles et al., 2000; Hiraoka et al., 1998) also are orthologs. We therefore rename human *SOX22* as *SOX12* and human *SOX20* as *SOX15* (Table 1).

Our analysis suggests that no further nomenclature changes or additions will be required for the mouse and human *Sox* family. The current system of nomenclature, loosely based on the order of gene discovery, is firmly entrenched in the literature, and the likely confusion and noncompliance associated with a more systematic nomenclature revision in our view outweigh the potential benefits. Our recommendations have been endorsed by the HUGO Gene Nomenclature Committee (http://www.gene.ucl.ac.uk/nomenclature).

Genomic idiosyncracies—notably pseudotetraploidy in *Xenopus laevis* and genome duplication in teleost fish—have hampered clear identification of *Sox* orthologs in some model organisms. Contentiously assigned full-length *Sox* genes isolated in such organisms are listed in Table 3, together with their closest mouse/human *Sox* homologs. Definitive nomenclature assignments are impossible in any species for which whole genome sequence has not been determined. We suggest that novel *Sox* genes identified in vertebrates be provisionally assigned the lowest available *Sox* number (currently 33), unless or until they can be confirmed as orthologs of existing mammalian genes.

In summary, our genomic analysis defines the extent of the *Sox* family of transcription factor genes in humans and mice, confirms gene homologies based on sequence, genomic organization, and chromosomal locations, and streamlines the nomenclature for vertebrate

Table 1. Pairing of Mouse and Human *Sox* Genes

Gene	*Sox* Group[a]	Major Known (or Deduced) Functions[b]	Species	Accession Number	Number of Exons	Chromosomal Location
Sry	A	Testis determination	Mouse	NM_0011564	1	Y (3cM)
			Human	NM_003140		Yp11.3
Sox1	B1	Lens development, (neural determination)	Mouse	NM_009233	1	8 (4cM)
			Human	NM_005986		13q34
Sox2	B1	Neural induction, (lens induction, pluripotency)	Mouse	NM_011443	1	3 (15cM)
			Human	BC013923		3q26.3
Sox3	B1	(Neural determination, lens induction)	Mouse	NM_009237	1	X (24.3cM)
			Human	NM_005634		Xq27
Sox4	C	Heart, lymphocyte, thymocyte development	Mouse	NM_009238	1	13 (20cM)
			Human	NM_003107		6q22.3
Sox5	D	Chondrogenesis	Mouse	NM_011444	15[c]	6 (69.5cM)
			Human	NM_006940		12p11.1
Sox6	D	Chondrogenesis, (cardiac myogenesis)	Mouse	NM_011445	17	7 (55cM)
			Human	NM_033326	16	11p15.3
Sox7	F	(Development of vascular and many other tissues)	Mouse	NM_011446	2	14 (28cM)[d]
			Human	NM_031439		8p22
Sox8	E	(Development of many tissues)	Mouse	AF191325	3	17 (8cM)
			Human	NM_014587		16p13.3
Sox9	E	Chondrogenesis, sex determination	Mouse	BC024958	3	11 (69.5cM)
			Human	NM_000346		17q25
Sox10	E	Neural crest specification	Mouse	AF047043	3	15 (46.5cM)
			Human	NM_006941		22q13
Sox11	C	(Neuronal, glial maturation)	Mouse	NM_009234	1	12 (11cM)[d]
			Human	NM_003108		2p25
Sox12[e]	C	(Development of many tissues)	Mouse	BF714412[f]	1	2 (86cM)[d]
			Human	NM_006943		20p13
Sox13	D	(Development of arterial walls, pancreatic islets)	Mouse	AB006329	13	1 (70cM)[d]
			Human	NM_005686	14	1q31
Sox14	B2	(Interneuron specification, limb development)	Mouse	AF193437	1	9 (53cM)
			Human	NM_004189		3q22
Sox15[g]	G	(Myogenesis)	Mouse	AB014474	2	11 (39cM)
			Human	NM_006942		17p13
Sox17	F	Endoderm specification	Mouse	NM_011441	3	1 (7cM)[d]
			Human	NM_022454		8q11.2
Sox18	F	Vascular and hair follicle development	Mouse	NM_009236	2	2 (96cM)[d]
			Human	NM_018419		20p13.3
Sox21	B2	(CNS patterning)	Mouse	BE647677[f]	1	14 (50cM)[d]
			Human	NM_007084		13q32
Sox30	H	(Male germ cell maturation)	Mouse	AV255326	5[c]	11 (20cM)[d]
			Human	NM_007017		5q35

[a] *Sox* groupings as determined by Bowles et al., 2000.

[b] Functions demonstrated by human mutant or mouse knockout phenotype; other possible functions (in parentheses) deduced from expression, cell transfection, or other studies. See Bowles et al., 2000; Wegner, 1999, and references therein; Cohen-Barak et al., 2001; Hosking et al., 2001; Katoh, 2002; and Takash et al., 2001. Also, see Uwanogho, 2001 (GenBank accession number AY069926).

[c] Gene subject to alternative splicing; value given indicates total number of utilized exons.

[d] Chromosomal location determined by comparison with the closest mapped gene.

[e] Human ortholog previously named *SOX22* (see text).

[f] Partially characterized gene that may contain additional exons.

[g] Human ortholog previously named *SOX20* (see text).

Sox genes. We hope that this will provide a useful framework for comparative and functional studies in a range of developmental model systems.

Goslik E. Schepers,[1] Rohan D. Teasdale,[1]
and Peter Koopman[2]
Institute for Molecular Bioscience and
ARC Special Research Centre for Functional
and Applied Genomics
The University of Queensland
Brisbane, Queensland 4072
Australia

[1] These authors contributed equally to this work.

[2] Correspondence: p.koopman@imb.uq.edu.au

Acknowledgments

We thank Dr. Elspeth Bruford, HUGO Nomenclature Committee, for comments on the manuscript and helpful discussions. We apologize

Table 2. Illegitimate Mouse and Human *Sox* Gene Fragments and Pseudogenes

Recorded Fragment	Accession Number	Species	Likely Identity	Notes
PCR-Derived HMG Box Sequences Submitted to GenBank[a]				
Sox16	L29084	Mouse	*Sox15*	
SOX25	AF032449	Human	*SOX21*	
SOX26	AF032450	Human	*SOX20*	
SOX27	AF032452	Human	*SOX20*	
SOX28	AF032452	Human	*SOX14*	
SOX29	AF032454	Human	*SOX5P*	
SOX-Related Genomic Fragments[b]				
SOX29	NT_008046 (LOC138007)	Human	*SOX5P*	Pseudogene at 8q21.1
Novel	NT_023726 (LOC206736)	Human	*SOX2*-related	474 bp non-HMG-box fragment at 8q24.13, with in-frame stop codons and gaps
Novel	NT_008101 (LOC137755)	Human	*SOX17*-related	234 bp non-HMG-box fragment adjacent to *SOX17* (8q11.22), with gaps
Novel	NT_033899 (LOC220283)	Human	*SOX20*-related	210 bp non-HMG-box fragment at 11q24.2, with gaps
Novel	NT_006788 (LOC206350)	Human	*SOX30*-related	668 bp non-HMG-box fragment adjacent to *SOX30* (5q34), with in-frame stop codons and insertions

[a] Wunderle et al., 1996; Crémazy et al., 1998. See also Layfield et al., 1994 (GenBank accession number L29084).

[b] Genomic fragments analyzed using NCBI LocusLink (http://www.ncbi.nlm.nih.gov/LocusLink/index.html).

Table 3. Contentiously Assigned Vertebrate *Sox* Genes

Species[a]	Published Gene Name[b]	Accession Number	*SOX* Group	Closest Mammalian Homolog[c]
Human	*HAF-1*	(deleted)	F	*SOX17*
Human	*HAF-2*	(deleted)	F	*SOX18*
Mouse	*SoxM/Sox21*	U66141	E	*Sox10*
Frog	*SoxD*	BAA32249	I[d]	*Sox31*[d]
Frog	*SoxB1*	(deleted)	B1	*Sox3*
Frog	*Sox12*	BAA09119	D	*Sox13*
Zebrafish	*Sox19*	X79821	B1	*Sox3*
Zebrafish	*Sox31*	AJ404687	B1	*Sox3*
Zebrafish	*Sox25/Sox30*	AF101266	B2	*Sox21*
Zebrafish	*Sox32/226D7*	NM_131851/AB071895	(non-Sox)	*Casanova*
Trout	*SoxLZ*	D61688	D	*Sox6*
Trout	*SoxP1*	D83256	E	*Sox8*
Trout	*Sox23*	BAA24402	D	*Sox13*
Trout	*Sox24*	BAA24575	C	*Sox11*

[a] Frog species, *Xenopus laevis*; trout species, *Oncorhynchus mykiss*.

[b] See GenBank entries and Stevens et al., 1996; Sakai et al., 1997; Bowles et al., 2000; Kikuchi et al., 2001; Sakaguchi et al., 2001; Hosking et al., 2001.

[c] Determined by BLAST and CLUSTALW analysis as being the closest mouse/human homolog.

[d] *Xenopus Sox31* does not correspond to any of the 20 mouse/human *Sox* genes and is in a group (I: Bowles et al., 2000) that is not represented in these species.

to colleagues whose work was not cited directly due to space constraints. P.K. is an Australian Research Council Professorial Research Fellow.

References

Bowles, J., Schepers, G., and Koopman, P. (2000). Phylogeny of the SOX family of developmental transcription factors based on sequence and structural indicators. Dev. Biol. *227*, 239–255.

Cohen-Barak, G., Hagiwara, N., Arlt, M., Horton, J., and Brilliant, M. (2001). Cloning, characterization and chromosome mapping of the human SOX6 gene. Gene *265*, 157–164.

Crémazy, F., Soullier, S., Berta, P., and Jay, P. (1998). Further complexity of the human *SOX* gene family revealed by the combined use of highly degenerate primers and nested PCR. FEBS Lett. *438*, 311–314.

Crémazy, F., Berta, P., and Girard, F. (2001). Genome-wide analysis of *Sox* genes in *Drosophila melanogaster*. Mech. Dev. *109*, 371–375.

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

Hosking, B., Wyeth, J., Pennisi, D., Wang, S., Koopman, P., and Muscat, G. (2001). Cloning and functional analysis of the Sry-related HMG box gene, Sox18. Gene *262*, 239–247.

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.

Katoh, M. (2002). Molecular cloning and characterization of human *SOX17*. Int. J. Mol. Med. *9*, 153–157.

Kikuchi, Y., Agathon, A., Alexander, J., Thisse, C., Waldron, S., Yelon, D., Thisse, B., and Stainier, D.Y. (2001). casanova encodes a novel Sox-related protein necessary and sufficient for early endoderm formation in zebrafish. Genes Dev. *15*, 1493–1505.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921.

Ohno, S. (1970). Evolution by Gene Duplication (Berlin: Springer-Verlag).

Sakaguchi, T., Kuroiwa, A., and Takeda, H. (2001). A novel sox gene, 226D7, acts downstream of Nodal signaling to specify endoderm precursors in zebrafish. Mech. Dev. *107*, 25–38.

Sakai, Y., Hiraoka, Y., Konishi, M., Ogawa, M., and Aiso, S. (1997). Isolation and characterization of *Xenopus laevis Xsox-b1* cDNA. Arch. Biochem. Biophys. *346*, 1–6.

Stevens, S., Ordentlich, P., Sen, R., and Kadesch, T. (1996). HMG box-activating factors 1 and *2*, two HMG box transcription factors that bind the human Ig heavy chain enhancer. J. Immunol. *157*, 3491–3498.

Takash, W., Canizares, J., Bonneaud, N., Poulat, F., Mattei, M.G., Jay, P., and Berta, P. (2001). SOX7 transcription factor: sequence, chromosomal localisation, expression, transactivation and interference with Wnt signalling. Nucleic Acids Res. *29*, 4274–4283.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., et al. (2001). The sequence of the human genome. Science *291*, 1304–1351.

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

Wunderle, V.M., Critcher, R., Ashworth, A., and Goodfellow, P.N. (1996). Cloning and characterization of *S0X5*, a new member of the human *SOX* gene family. Genomics *36*, 354–358.