Short Communications

Mammalian Genome

Mammalian Genome 5, 640–642 (1994).

The cDNA sequence and chromosomal location of the human SOX2 gene

M. Stevanovic,[1,*] O. Zuffardi,[2] J. Collignon,[3,] R. Lovell-Badge,[3] P. Goodfellow[1]**

[1]Department of Genetics, University of Cambridge, Downing Street, Cambridge, CB2 3EH, UK
[2]Biologia General E Genetica, Medica Via Forlanini 14, 27100 Pavia, Italy
[3]Medical Research Council, National Institute for Medical Research, The Ridgeway, Mill Hill, London, NW7 1AA, UK

Received: 21 April 1994 / Accepted: 1 June 1994

We have characterized a cDNA containing the coding region of a human SOX gene expressed in fetal brain. The cDNA is 1085 bp long, contains an open reading frame of 317 amino acids, and displays a high degree of similarity with the mouse *Sox-2* gene. Human SOX2 has been localized to the long arm of Chromosome (Chr) 3 in the region q26.3-27.

The mammalian sex-determining gene, SRY, encodes a protein that includes a sequence motif known as the HMG-box (Sinclair et al. 1990; Goodfellow and Lovell-Badge 1993). This motif is responsible for the different DNA-binding activities of the SRY protein, and mutation analysis suggests that a functional HMG-box is required for sex determination (Harley et al. 1992). The HMG-box motif is found widely in proteins that bind to DNA including transcription factors, proteins that bind to DNA with limited sequence specificity, and non-sequence-specific chromatin proteins (Ner 1992).

The mammalian genome contains a family of genes that are related to SRY in the region that encodes the HMG-box. These genes have been called SOX genes (SRY-related HMG-box genes). In the original report, four genes were identified as members of the mouse *Sox* gene family (Gubbay et al. 1990). These have been named *Sox-1, Sox-2, Sox-3,* and *Sox-4* (previously referred to as *a1, a2, a3,* and *a4*). All these genes are expressed during embryonic development as well as in some adult tissues; *Sox-1, Sox-2,* and *Sox-3* are expressed at the highest levels in the developing nervous system (Collignon 1993). Subsequently, SOX genes have been identified in a wide variety of phylogenetically diverse organisms including mammals, birds, and insects. In most cases, the sequence information available is limited to the HMG-box (Griffiths 1991; Denny et al. 1992a; Goze et al. 1993; Wright et al. 1993; van de Wetering and Clevers 1993; Chardard et al. 1993), and only a few SOX genes have been characterized in more detail including *Sox-5* (Denny et al. 1992b), IRE-ABP (Nasrin et al. 1992), *Sox-4* (Schilham et al. 1993) SOX4 (Farr et al. 1993), and SOX3 (Stevanovic et al. 1993).

Southern analysis and hybridization with different SOX-box-containing probes indicate that the SOX gene superfamily is large and, based on sequence similarities, SOX genes can be divided into subfamilies (Wright et al. 1993).

The neuronal expression of *Sox-1, Sox-2,* and *Sox-3* and the high level of expression of SOX3 in fetal brain and spinal cord (Collignon 1993; Stevanovic et al. 1993) prompted us to screen a fetal brain cDNA library for additional SOX gene clones. A library, constructed with mRNA derived from the brains of 14- to 16-week-old embryos (a gift from H. Lehrach), was screened with a 250-bp DNA fragment containing the SOX-box of the human SOXA gene (see Stevanovic et al. 1993). 20,000 clones were screened, and four positive clones were identified, isolated, and sequenced. These clones fall into two classes and encode two different human SOX genes. One cDNA is derived from the SOX4 gene, while the other three clones, with inserts varying from 0.9 to 1.15 kb in size, have overlapping sequences. The 1085-bp nucleotide sequence, obtained from the longest cDNA clone, is presented in Fig. 1. The open reading frame includes the expected HMG-box. The amino acid sequence of the HMG-box is identical to that encoded by murine *Sox-2* (Gubbay et al. 1990). The high level of sequence similarity extends outside the box region (Collignon 1993), indicating that we have cloned sequences corresponding to the human SOX2 gene. This conclusion is strengthened by Southern blot analysis of human DNA; under conditions of high stringency, probes derived from the SOX2 cDNA recognize a single locus in the human genome (data not shown).

The sequence of the SOX2 cDNA contains an open reading frame with three potential in-frame start codons (Fig. 1). The unfavorable sequence context of the first ATG codon suggests that this may not be the translation initiation site. The second ATG fulfills the requirements of Kozak's rules (Kozak 1987). Although the cDNA sequence reveals a short poly A tract at the 3′ end, no obvious

Present address: Genetic Engineering, Vojvode Stepe 283, PO Box 794, 11001 Belgrade, Yugoslavia.

**The Biological Laboratories, Harvard University, 16 Divinity Avenue, Cambridge, MA 02138, USA

Correspondence to: M. Stevanovic

```
CACAGCGCCCGC ATG TAC AAC ATG ATG GAG ACG GAG CTG AAG CCG CCG GGC CCG CAG CAA ACT TCG GGG GGC GGC GGC GGC AAC TCC ACC GCG GCG GCG   99
             M   Y   N   M   M   E   T   E   L   K   P   P   G   P   Q   Q   T   S   G   G   G   G   G   N   S   T   A   A   A

GCC GGC GGC AAC CAG AAA AAC AGC CCG GAC CGC GTC AAG CGG CCC ATG AAT GCC TTC ATG GTG TGG TCC CGC GGG CAG CGG CGC AAG ATG GCC CAG  195
 A   G   G   N   Q   K   N   S   P   D   R   V   K   R   P   M   N   A   F   M   V   W   S   R   G   Q   R   R   K   M   A   Q

GAG AAC CCC AAG ATG CAC AAC TCG GAG ATC AGC AAG CGC CTG GGC GCC GAG TGG AAA CTT TTG TCG GAG ACG GAG AAG CGG CCG TTC ATC GAC GAG  292
 E   N   P   K   M   H   N   S   E   I   S   K   R   L   G   A   E   W   K   L   L   S   E   T   E   K   R   P   F   I   D   E

GCT AAG CGG CTG CGA GCG CTG CAC ATG AAG GAG CAC CCG GAT TAT AAA TAC CGG CCC CGG CGG AAA ACC AAG ACG CTC ATG AAG AAG GAT AAG TAC  387
 A   K   R   L   R   A   L   H   M   K   E   H   P   D   Y   K   Y   R   P   R   R   K   T   K   T   L   M   K   K   D   K   Y

ACG CTG CCC GGC GGG CTG CTG GCC CCC GGC GGC AAT AGC ATG GCG AGC GGG GTC GGG GTG GGC GCC GGC CTG GGC GCG GGC GTG AAC CAG CGC ATG  483
 T   L   P   G   G   L   L   A   P   G   G   N   S   M   A   S   G   V   G   V   G   A   G   L   G   A   G   V   N   Q   R   M

GAC AGT TAC GCG CAC ATG AAC GGC TGG AGC AAC GGC AGC TAC AGC ATG ATG CAG GAC CAG CTG GGC TAC CCG CAG CAC CCG GGC CTC AAT GCG CAC  579
 D   S   Y   A   H   M   N   G   W   S   N   G   S   Y   S   M   M   Q   D   Q   L   G   Y   P   Q   H   P   G   L   N   A   H

GGC GCA GCG CAG ATG CAG CCC ATG CAC CGC TAC GAC GTG AGC GCC CTG CAG TAC AAC TCC ATG ACC AGC TCG CAG ACC TAC ATG AAC GGC TCG CCC  675
 G   A   A   Q   M   Q   P   M   H   R   Y   D   V   S   A   L   Q   Y   N   S   M   T   S   S   Q   T   Y   M   N   G   S   P

ACC TAC AGC ATG TCC TAC TCG CAG CAG GGC ACC CCT GGC ATG GCT CTT GGC TCC ATG GGT TCG GTG GTC AAG TCC GAG GCC AGC TCC AGC CCC CCT  772
 T   Y   S   M   S   Y   S   Q   Q   G   T   P   G   M   A   L   G   S   M   G   S   V   V   K   S   E   A   S   S   S   P   P

GTG GTT ACC TCT TCC TCC CAC TCC AGG GCG CCC TGC CAG GCC GGG GAC CTC CGG GAC ATG ATC AGC ATG TAT CTC CCC GGC GCC GAG GTG CCG GAA  867
 V   V   T   S   S   S   H   S   R   A   P   C   Q   A   G   D   L   R   D   M   I   S   M   Y   L   P   G   A   E   V   P   E

CCC GCC GCC CCC AGC AGA CTT CAC ATG TCC CAG CAC TAC CAG AGC GGC CCG GTG CCC GGC ACG GCC ATT AAC GGC ACA CTG CCC CTC TCA CAC ATG  963
 P   A   A   P   S   R   L   H   M   S   Q   H   Y   Q   S   G   P   V   P   G   T   A   I   N   G   T   L   P   L   S   H   M

TGAGGGCCGGACAGCGAACTGGAGGGGGGAGAAATTTTCAAAGAAAAACGAGGGAAATGGGAGGGGTGCAAAAGAGGAGAGTAAGAAACAGCATGGAGAAAACCCGGTACGCTCAAAAAAAA  1086
```

Fig. 1. DNA sequence and deduced amino acid sequence of the human SOX2 cDNA. The cDNA was sequenced according to the Sanger dideoxy DNA sequencing procedure (Sanger et al. 1977). The HMG-box is boxed. The accession no. for this sequence is Z31560.

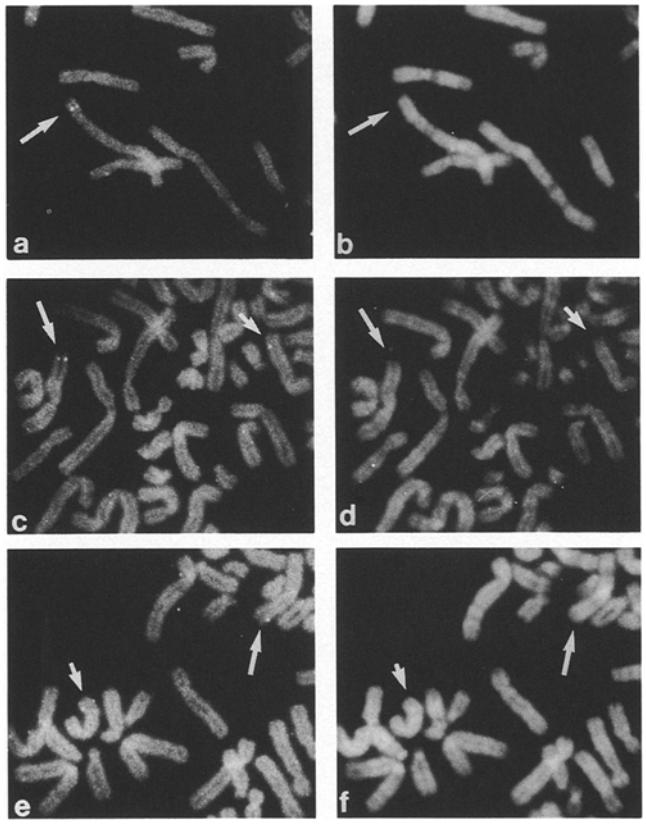

Fig. 2. Fluorescence in situ hybridization of SOX2 cDNA. Detection was made according to the ONCOR detection kit with three amplification steps. Chromosomes were counterstained with propidium iodide (1 μg/ml), banded with diamidinophenylenindole (DAPI; Schweizer 1981), and mounted in antifading solution. a,c,e: partial metaphases with hybridization signals at 3q26.3-q27 (long arrows) and at 8q24.1 (short arrows). In b,d,f, DAPI banding of the same metaphases.

polyadenylation signal precedes this run of As. Preliminary Northern blot analysis suggests that the size of the SOX2 mRNA is approximately 3.5 kb in a human teratocarcinoma cell line (NTera2ND1, data not shown) implying that noncoding sequences are missing from the cDNA.

Chromosomal assignment of the SOX2 gene was initially determined by Southern blot hybridization of the cDNA probe to the DNA from a panel of 15 human–rodent somatic cell hybrids previously characterized for retention of human chromosomes (see Farr et al. 1993). There was cross-hybridization of the human SOX2 probe with the rodent DNA; however, *Eco*R1 digestion distinguished between human (5.0 kb), hamster (6.3 kb), and mouse (approximately 20 kb) derived fragments. Perfect concordance was observed between the presence of the human SOX2 hybridizing band and human Chr 3.

Regional mapping of SOX2 was performed with fluorescent in situ hybridization to the metaphase chromosomes from a normal 46XY male (Fig. 2). The plasmid containing the complete 1.08-kb SOX2 cDNA was labeled by nick translation with biotin-16dUTP and hybridized to chromosomes as described by Rossi and coworkers (1993). In 50 metaphases with no more than ten signals, 91 hybridization spots were associated with chromosomes: 57 (62.6%) were located at 3q26.3-27, and 20 (22%) at 8q24. The distribution of the remaining 14 signals was random. This result confirms the somatic cell hybrid studies and indicates that the SOX2 gene resides in the q26.3-27 region of Chr 3; the signal at 8q24 may be due to either a pseudogene or another member of the SOX gene family.

We have cloned sequences corresponding to SOX2, another member of the human SOX gene family expressed in fetal brain. The discovery of humans mutated for SOX2 or the construction of mice mutant for *Sox2* should allow a direct test of the function of SOX2 during neuronal development.

Acknowledgment. We would like to thank Dr. H. Lehrach for providing the fetal brain cDNA library.

References

Chardard, D., Chesnel, A., Goze, C., Dournon, C., Berta, P. (1993). PwSOX-1: the first member of the Sox gene family in Urodeles. Nucleic Acids Res. 21, 3576.

Collignon, J. (1993). Study of a new family of genes related to the mammalian testis determining gene. PhD Thesis, London University.

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

Denny, P., Swift, S., Connor, F., Ashworth, A. (1992b). An SRY-related gene expressed during spermatogenesis in the mouse encodes a sequence-specific DNA-binding protein. EMBO J. 11, 3705–3712.

Farr, C.J., Easty, D.J., Ragoussis, J., Collignon, J., Lovell-Badge, R., Goodfellow, P.N. (1993). Characterisation and mapping of the human SOX4 gene. Mamm. Genome 4, 577–584.

Goodfellow, P.N., Lovell-Badge, R. (1993). SRY and sex determination in mammals. Annu. Rev. Genet. 27, 71–92.

Goze, C., Poulat, F., Berta, P. (1993). Partial cloning of SOX11 and SOX12, two new human SOX genes. Nucleic Acids Res. 21, 2943.

Griffiths, R. (1991). The isolation of conserved DNA sequences related to the human sex-determining region Y gene from the lesser black-backed gull (*Larus fuscus*) Proc. R. Soc. Lond. [Biol.] 244, 123–128.

Gubbay, J., Collignon, J., Koopman, P., Capel, B., Economou, A., Münsterberg, A., Vivian, N., Goodfellow, P., 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.

Harley, V.R., Jackson, D.I., Hextall, P.J., Hawkins, J.R., Berkovitz, G.D., Sockanathan, S., Lovell-Badge, R., Goodfellow, P.N. (1992). DNA binding activity of recombinant SRY derived from normal males and XY females. Science 255, 453–457.

Kozak, M. (1987). An analysis of 5′ noncoding sequences from 699 vertebrate messenger RNAs. Nucleic Acids Res. 15, 8125–8148.

Nasrin, N., Buggs, C., Kong, X.F., Carnazza, J., Goebl, M., Alexander-Bridges, M. (1991). DNA-binding properties of the product of the testis-determining gene and a related protein. Nature 354, 317–320.

Ner, S.S. (1992). HMGs everywhere. Current Biol. 2, 208–210.

Rossi, E., Zarrilla, R., Zuffardi, O. (1993). Regional assignment of the gene coding for the human Grave's disease autoantigen to 10q21.3-q22.1. Hum. Genet. 90, 653–654.

Sanger, F., Nicklen, S., Coulson, A.R. (1977). DNA sequencing with chain-termination inhibitors. Proc. Natl. Acad. Sci. USA 74, 5463–5467.

Schilham, M.W., van-Eijk, M., van der Wetering, M., Clevers, H.C. (1993). The murine Sox-4 protein is encoded on a single exon. Nucleic Acids Res. 21, 2009.

Schweizer, D. (1981). Counterstain-enhanced chromosome banding. Hum. Genet. 57, 1–14.

Sinclair, A.H., Berta, P., Palmer, M.S., Hawkins, J.R., Griffiths, B.L., Smith, M.J., Foster, J.W., Frischauf, A-M., Lovell-Badge, R., Goodfellow, P.N. (1990). A gene from the sex determining region encoding a protein with homology to a conserved DNA-binding motif. Nature 346, 240–244.

Stevanovic, M., Lovell-Badge, R., Collignon, J., Goodfellow, P.N. (1993). *SOX3* is an X-linked gene related to *SRY*. Hum. Mol. Genet. 2, 2013–2018.

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

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