Development 122, 509-520 (1996)
Printed in Great Britain © The Company of Biologists Limited 1996
DEV3213

A comparison of the properties of *Sox-3* with *Sry* and two related genes, *Sox-1* and *Sox-2*

Jérôme Collignon[1,†,‡], Shanthini Sockanathan[1,†], Adam Hacker[1,†], Michel Cohen-Tannoudji[1,§], Dominic Norris[2,¶], Sohaila Rastan[2,¶], Milena Stevanovic[3], Peter N. Goodfellow[3] and Robin Lovell-Badge[1,*]

[1]Laboratory of Developmental Genetics, MRC National Institute for Medical Research, The Ridgeway, Mill Hill, London NW7 1AA, UK
[2]Section of Comparative Biology, MRC Clinical Research Centre, Harrow, HA1 3UJ, UK
[3]Department of Genetics, University of Cambridge, Downing Street, Cambridge CB2 3EH, UK

*Author for correspondence
†These authors made an equal contribution to the work presented
‡Present addresses: The Biological Laboratories, Harvard University, Cambridge MA, USA
§Present address: Institut Pasteur, 28 rue du Dr Roux, 75724 Paris Cedex 15, France
¶MRC Clinical Sciences Centre, RPMS, Hammersmith Hospital, Du Cane Road, London W12 0NN, UK

SUMMARY

The *Sox* gene family consists of a large number of embryonically expressed genes related via the possession of a 79-amino-acid DNA-binding domain known as the HMG box. Partial clones for the first three *Sox* genes (a1-a3) were isolated by homology to the HMG box of the testis-determining gene *Sry* and are now termed *Sox-1*, *Sox-2* and *Sox-3*. *Sox-3* is highly conserved amongst mammalian species and is located on the X chromosome. This has led to the proposal that *Sry* evolved from *Sox-3*. We present the cloning and sequencing of *Sox-1*, *Sox-2* and *Sox-3* from the mouse and show that *Sox-3* is most closely related to *Sry*. We also confirm that mouse *Sox-3* is located on the X chromosome between *Hprt* and *Dmd*. Analysis of the distribution of *Sox-3* RNA shows that its main site of expression is in the developing central nervous system, suggesting a role for *Sox-3* in neural development. Moreover, we demonstrate that *Sox-3*, as well as *Sox-1* and *Sox-2*, are expressed in the urogenital ridge and that their protein products are able to bind the same DNA sequence motif as *Sry* in vitro, but with different affinities. These observations prompt discussion of an evolutionary link between the genes and support the model that *Sry* has evolved from *Sox-3*. However our findings imply that if this is true, then *Sry* has undergone concomitant changes resulting in loss of CNS expression and altered DNA-binding properties.

Key words: *Sox* genes, HMG Box, transcription factor, *Sry*, sex determination

INTRODUCTION

The mammalian testis-determining gene *Sry* is expressed for a brief period in the indifferent gonad and causes a switch in developmental fate from the default ovarian pathway to that of the testis (Koopman et al., 1991; Hacker et al., 1995). The gene is located on the Y chromosome and encodes a transcription factor containing a 79-amino-acid DNA-binding domain known as the HMG box (Capel and Lovell-Badge, 1993). Many genes have now been cloned on the basis of their homology to the HMG box of *Sry* and comprise the *Sox* gene family, for *Sry*-related HMG box-containing genes. Approximately 20 *Sox* genes have been found in mammals, with more being identified in other vertebrates and invertebrates such as *Drosophila* and *C. elegans* (Denny et al., 1992a; Wright et al., 1993; Laudet et al., 1993, Coriat et al., 1993; our unpublished results). The homology exhibited within the HMG box varies between members of the family but allows them to be grouped into different subfamilies. For example, the first *Sox* genes to be cloned, a1-a3 (now termed *Sox1-3*), are in the subfamily containing *Sry*, due to their high similarity (approximately 90%) (Gubbay et al., 1990; Denny et al., 1992a).

Recently, it has been proposed that *Sry* originated from the *Sox-3* gene. This is based upon sequence information from human and marsupial *Sox-3* homologues, showing it to be highly conserved in mammalian evolution and to be the *Sox* gene most closely related to *Sry*, despite the latter's divergence (Stevanovic et al., 1993; Foster and Graves, 1994). In addition, *Sox-3* has been mapped to the X chromosome in both these mammals. This is relevant because the X and the Y chromosomes are thought to have arisen from a common ancestor 'autosome' in the lineage that gave rise to mammals. However, given that there are at least 20 mammalian *Sox* genes (e.g. Wright et al., 1993) and probably many more that have not yet been identified, the location of *Sox-3* on the X chromosome may be fortuitous. *Sry* could have originated from any other *Sox* gene by gene duplication and translocation or by retroposition. It would therefore be more informative to compare the properties exhibited by *Sry* with *Sox-3* and other closely related members of the *Sox* gene family in order to determine how similar they are.

The expression of *Sry* in the mouse is tightly regulated, allowing expression in the indifferent gonad between 10.5 and 12.5 dpc to occur in order for the testis determination pathway to be initiated (Koopman et al., 1991; Hacker et al., 1995). The function of SRY protein relies critically on its ability to bind DNA in a sequence-specific manner (Berta et al., 1990; Harley et al., 1992; Pontiggia et al., 1994). When it does so, it causes the DNA to bend through a dramatic angle, which itself is thought to be crucial functionally. The ability of the protein to interact with DNA is dependent on its DNA-binding domain, the HMG box. Mutations within this domain can interfere with both DNA-binding as well as bending and result in sex reversal to give XY females (Pontiggia et al., 1994). The function of the rest of the SRY protein is unclear. It has diverged very rapidly during evolution such that the mouse and human genes show no homology outside the HMG box region (Whitfield et al., 1993; Tucker and Lundrigan, 1993). Also, no de novo mutations known to lead to sex reversal have been found outside the HMG box. These data suggest that the HMG box is the only functional part of SRY and that it works only through interaction with specific target sequences.

In short, the properties of *Sry* known to be required for its function are its expression within the genital ridge and the sequence-specific interaction of the protein with DNA. We have chosen to examine if the mouse *Sox-3* gene and its product share these properties with *Sry*. We have also compared two other closely related genes, *Sox-1* and *Sox-2*, to determine if any similarities observed are peculiar to *Sry* and *Sox-3*, or indeed, are representative of other members of the subfamily. In this report we describe some properties of *Sox-1*, *Sox-2* and *Sox-3* in the mouse and confirm that only *Sox-3* is X-linked. We show that SOX-2 and SRY show similar affinities in binding to the AACAAT consensus motif but surprisingly weaker binding is effected by SOX-3 and SOX-1. The main site of expression of *Sox-3* is, like *Sox-1* and *Sox-2*, within the developing CNS, although all three genes are also expressed within the genital ridge. In addition, we show that *Sox-3* is likely to be expressed in the same cell type as *Sry* in the indifferent gonad. Our findings do not eliminate the possibility that *Sry* could have evolved from any *Sox* gene, but *Sox-3* is the best candidate due to its chromosomal position, its degree of homology to *Sry* and its genital ridge expression. Our results suggest that if *Sox-3* is the evolutionary ancestor of *Sry*, then *Sry* must have undergone concomitant changes, resulting in loss of CNS expression and altered DNA-binding properties.

MATERIALS AND METHODS

DNA analysis

DNA manipulations were carried out according to Sambrook et al. (1989). Genomic DNA was isolated from adult spleens of male and female mice, as described by Lovell-Badge (1987). Southern blots were washed in 0.1× SSC, 0.1% SDS at 70°C for 1 hour and then exposed to X-ray film for 3 days.

Inserts were subcloned into pBluescript (Stratagene) for sequencing. Sequencing, by the dideoxy method (Sanger et al., 1977), was performed using the T7 Sequencing Kit (Pharmacia) according to the manufacturer's instructions. *Sox-3* was sequenced on both strands of subclones derived from the cDNA clone 7d and a *Sox-3* genomic clone. Gaps were filled using the following primers:

0-30, GCCAACGGCGCCTACTC;
0-31, TGCGTGTACGTGTCCAG;
Sox-3.3: CAGCCTGCTGGAGACTGAACT.

Sequence analysis was performed using software designed by the Genetics Computer Group at the University of Wisconsin (Devereux et al., 1984).

Preparation of purified GST fusion proteins

The HMG boxes of mouse SRY, SOX-1, SOX-2 and SOX-3 were subcloned into the *Bam*HI-*Eco*RI sites of the bacterial expression vector pGEX3T by PCR. The fusion proteins were induced by addition of IPTG for 3 hours and purified by adsorption to glutathione agarose beads, as described by Smith and Johnson (1988). Purity was assessed by staining SDS-polyacrylamide gels with Coomassie blue.

Gel retardation assays

Probes were prepared by annealing the complementary oligonucleotides prior to labelling by filling in with Klenow and [α-^{32}P]dCTP. These were then gel-purified by electrophoresis through a 12% non-denaturing polyacrylamide gel. DNA-binding assays were performed essentially as described by Harley et al. (1992). However, 1.5 μg of sonicated salmon sperm DNA was used as non-specific competitor in the binding reactions instead of poly(dIdC). The sequence of the target oligonucleotide used is 5′ CGGAGAAACTTCTAG**AACAAT**GGTGAAGACGGGGATC 3′.

Northern analysis

Total RNA was isolated according to the method described by Auffray and Rougeon (1980). 10 μg of total RNA was electrophoresed in a 1% agarose gel containing 7% (v/v) formaldehyde in 1× MOPS buffer and transferred onto Genescreen membrane. ^{32}P-labelled single-stranded antisense RNA probes were produced using a 440-bp *Xho*I-*Xmn*I fragment from the *Sox-3* genomic clone (200 bp 5′ from start of cDNA to position 223 of cDNA, Fig. 1), a 460-bp *Eco*RI-*Apa*I fragment from the *Sox-2* cDNA (position 1-460 on cDNA, Fig. 1) and a 340-bp *Stu*I-*Xho*I fragment from the *Sox-1* cDNA (position 1694-2046 on cDNA, Fig. 1). Hybridisation was performed at 65°C for 16 hours in 60% formamide, 5× SSC, 1× Denhardts, 20 mM sodium phosphate (pH 6.8), 10% dextran sulfate, 1% SDS, 100 μg/ml tRNA and 100 μg/ml sheared denatured herring testis DNA. Filters were washed in 0.1× SSC, 0.5% SDS at 70°C for 2 hours. Background signal was eliminated by RNase treatment of the filters (0.2 μg/ml RNase A in 2× SSC for 20 minutes, followed by 2× SSC, 0.5% SDS for 30 minutes at 50°C).

In situ hybridisation

Two probes from the 5′ region of the *Sox-3* gene were used: a 100-bp *Eco*RI-*Pst*I fragment from the cDNA (position 1-108 on cDNA, Fig. 1) and a 440-bp *Xho*I-*Xmn*I fragment from the genomic clone (see above). Transcription reactions were performed in vitro using T7 RNA polymerase (Promega) or T3 RNA polymerase (Pharmacia) to generate sense (control) and antisense RNA probes labelled with ^{35}S-UTP. The complete procedure is as described in Wilkinson and Green (1990). After hybridisation the slides were subjected twice to a high stringency wash in 50% formamide, 2× SSC, 10 mM DTT at 65°C for 30 minutes. Slides were exposed for 6-10 days. No signal was detected with the sense probe.

RNase protection

Total RNA was prepared from adult tissue samples of Parkes inbred mice as described by Capel et al. (1993), or from embryonic tissue by homogenising in 300 μl NETS (100 mM NaCl, 50 mM Tris HCl, pH 8.0, 0.5% SDS, 5 mM EDTA), followed by phenol extraction and phenol-chloroform extraction before precipitating with 10 μg yeast tRNA. RNase protection assays were carried out using 10 μg adult tissue RNA or RNA prepared from four pairs of genital ridge, as

described by Capel et al. (1993). Anti-sense labelled probes were derived from the 440-bp *Xho*I-*Xmn*I *Sox-3* genomic fragment, as used in northern analyses and in situ hybridisations (see above), subcloned into *Xho*I-*Hin*cII pBluescript KS (Stratagene), linearised at *Dde*I and transcribed with T7 RNA polymerase; and from a 363-bp *Bgl*II-*Pst*I genomic fragment of *Sry* subcloned into *Bam*HI-*Pst*I pBluescript KS, linearised at *Not*I in the polylinker and transcribed with T3 RNA polymerase. The *Sox-2* probe used is as described in Fig. 1 (200 bp 5′ to cDNA start to position 257 on cDNA), whereas the probe used for *Sox-1* consisted of a 227-bp *Sma*I-*Stu*I fragment (position 1467-1694 on cDNA, Fig. 1). Embryos from matings of W^e heterozygous mice were genotyped as described previously (Koopman et al., 1990). The W^e mutation is maintained on a C3H background, where homozygotes show essentially no germ cells in the genital ridge at 11.5 dpc (Buehr et al., 1993).

RESULTS

Cloning and sequencing of *Sox-1*, *Sox-2* and *Sox-3*

An 8.5-dpc whole mouse embryo cDNA library (Fahrner et al., 1987) was screened with a human *SRY* probe in order to identify genes closely related to the testis-determining gene (Gubbay et al., 1990). 12 clones were identified, seven of which correspond to *Sox-2*, two to *Sox-1*, one to *Sox-3* and two weakly hybridizing clones, which were found to be *Sox-4*. The full-length sequences for *Sox-1*, *-2* and *-3* are presented in Fig. 1. The sequence obtained for *Sox-1* includes the open reading frame, but at 2.4 kb, it is shorter than the specific 4.1-kb transcript seen on northern blots (Fig. 4). No polyadenylation signal or poly(A) tail was identified at the 3′ end of the cDNA. This is missing in conjunction with 5′ untranslated sequences. The *Sox-2* sequence shown is likely to correspond to the full-length transcript as its size corresponds to the specific 2.4-kb band detected in northern analyses (Fig. 4). A poly(A) tail can be seen downstream of a canonical polyadenylation signal. Moreover, potential start and termination codons can be identified, respectively preceded or followed by several in-frame stop codons. The single cDNA obtained for *Sox-3* was found to be incomplete as no translational termination codon was present. Therefore, a genomic clone was isolated from a 129/Sv mouse strain library (Gubbay et al., 1990) to enable the sequence to be completed. The transcriptional start site of *Sox-3* has been confirmed by primer extension and RNase protection assays and is in an equivalent position to that of human *SOX3* (data not shown; Hacker, 1995). Sequence and restriction analysis comparisons between the cDNA and genomic clones, as well as RNase protection studies (M. Parsons and R. Lovell-Badge, unpublished data), indicated that *Sox-3* coding sequences are located within a single exon. This is also true for *Sox-1*, *Sox-2* and mouse and human *Sry/SRY* (Clépet et al., 1993; Hacker et al., 1995).

Sequence comparisons and conservation

The three proteins obtained after conceptual translation of the *Sox-1*, *Sox-2* and *Sox-3* cDNAs show that they share sequence homology outside the HMG box region. A schematic view of this is shown in Fig. 2. The proteins start with a short (40-65 amino acids) N-terminal domain rich in hydrophobic residues, followed by the 79-amino-acid HMG box prior to a large C-terminal domain varying in length between the genes (262, 199 and 229 residues for SOX-1, SOX-2 and SOX-3 respectively). Fig. 2B presents the sequence comparisons between the C-terminal domains of the three proteins. Four poly(alanine) stretches are present in SOX-1 and SOX-3 but not in SOX-2, two of which are at similar positions in both proteins. Furthermore, a PRD-type repeat (His-Pro) (Frigerio et al., 1986) is present only in SOX-1. Thus although some similarities are evident between the three proteins, each one has a unique combination of structural motifs.

All three genes show no homology outside the HMG box with human or mouse *SRY/Sry*, or with sequences obtained from other *Sox* genes, for example *Sox-4* (Gubbay et al., 1990; Sinclair et al., 1990; Farr et al., 1993; van de Wettering et al., 1993) or *Sox-5* (Denny et al., 1992b). *Sox-1*, *-2* and *-3* can therefore be grouped into a subfamily distinct from other *Sox* genes. Moreover, the similarity of their structure as well as the sequence of these genes suggest that they may have arisen early on by a gene duplication event from a single ancestral *Sox* gene of this type. From comparisons within the HMG box (Fig. 2C), *Sry* can be considered a member of this subfamily; as human, marsupial and mouse *Sry* show high degrees of homology to each of *Sox-1*, *Sox-2* and *Sox-3* (Stevanovic et al., 1993; Foster and Graves, 1994). However, comparisons between the three genes show that *Sox-3* is most closely related to *Sry*.

Chromosomal localisation of *Sox-3*

Southern analysis of *Bgl*II-digested mouse genomic DNA, with both *Sox-3* and *Sox-1* cDNA probes, showed the 9.5-kb *Sox-3* cognate band to be twice as intense in the female track compared to the male, consistent with it being X-linked (Fig. 3A). In order to confirm the linkage and to determine the position of *Sox-3* on the chromosome, interspecific backcross pedigree analysis was utilized. A *Pst*I fragment corresponding to the 5′ end of the *Sox-3* cDNA was used to define a *Hin*dIII restriction fragment length variant (RFLV) between *Mus musculus domesticus* (allele size 6.1 kb) and *Mus spretus* (allele size 5.3 kb). The segregation of this RFLV was analysed through a panel of interspecific backcross animals with previously characterised breakpoints spanning the X chromosome (Kay et al., 1991). Fig. 3B shows the haplotypes of the recombinant X chromosome of 15 representative individuals out of a total of 19 selected backcross animals analysed. This analysis locates *Sox-3* between the anchor loci *Hprt* and *Dmd*, at a genetic distance of 7.4 ± centiMorgans distal to *Hprt*. No known mutation corresponds to this position in mice, but our analysis places *Sox-3* close to *Fmr-1* (Laval et al., 1992). Consistent with this result, mapping data obtained for the human *SOX3* gene places it in the conserved syntenic region of the X chromosome, close to fragile X syndrome (Stevanovic et al., 1993). In comparison, *Sry/SRY* is located on the short arm of both the mouse and human Y chromosome, whereas in contrast, the two other subfamily members, *Sox-1* and *Sox-2*, are located on different autosomes (data not shown).

Sizing of the *Sox-1*, *Sox-2* and *Sox-3* transcripts

Northern analyses were carried out to ascertain the sizes of the transcripts of *Sox-1*, *Sox-2* and *Sox-3* and to determine if the cDNA clones obtained were complete. As shown in Fig. 4A, *Sox-1* and *Sox-2* probes detected bands at 4.0 kb and 2.4 kb, respectively, in RNA samples from embryos at 9.5 dpc to 11.5 dpc. A less abundant transcript that extends more 5′ to the *Sox-*

Sox-1

```
   1 ACAGGAACGGAGACTTCGAGCCGAGAAGAGGAGGCAGCGAACCCTGCGTCGGGCCCAGGGGCACCGCTTCAGACCCAGAAAGTGGAGCCTCAACTTGGCC  100
 101 ACGACTGCACCTGTTTGCACAGTTCAGCCCTGAGTGACCGGACGGCAGCAATCAACCTGGCCATCGGCCTCTTTGGCAAGTGGTTTGTGCACCGGGAGAA  200
 201 ACTTTCCACCTGCGAGCTGGACCCGCGCTAAGTGCGTGTGCTTTTGCCTCTTTTTTGTTGTTGTTGTTGTTGTGGCCTCCACCCAACCCCCTTCTCTCCG  300
 301 CTAGGCACCCACCGCACACACACCCCCCCCCCCAGTCTCTCTGGGCTGATCCTCTCTCCACCCACCCACCCCCACCCGGCCGTCTATGCTCCAGGCCCTC  400
 401 TCTTTGCGGTACCGGTGAACCCGCTAGCCGCCCAGATGTACAGCATGATGATGGAGACCGACCTGCACTCGCCCGGCGGCGCCCAGGCGCCCACGAACCT  500
                                     M  Y  S  M  M  M  E  T  D  L  H  S  P  G  G  A  Q  A  P  T  N  L
 501 CTCGGGCCCGGCCGGGGCGCGCGGGGGCGGCGGTGGGGGCGGGGGCGGCGGCGGCGGCGGGGGCACCAAGGCCAACCAGGATCGGGTCAAGCGGCCCATG  600
       S  G  P  A  G  A  R  G  G  G  G  G  G  G  G  G  G  G  G  G  G  T  K  A  N  Q  D  R  V  K  R  P  M
 601 AACGCCTTCATGGTGTGGTCCCGCGGACAGCGGCGCAAGATGGCCCAGGAAAACCCCAAGATGCACAACTCGGAGATCAGCAAGCGCCTCGGGGCCGAGT  700
       N  A  F  M  V  W  S  R  G  Q  R  R  K  M  A  Q  E  N  P  K  M  H  N  S  E  I  S  K  R  L  G  A  E  W
 701 GGAAGGTCATGTCCGAGGCCGAGAAGCGGCCGTTCATCGACGAGGCCAAGAGACTGCGCGCGCTGCACATGAAGGAACACCCGGATTACAAGTACCGGCC  800
       K  V  M  S  E  A  E  K  R  P  F  I  D  E  A  K  R  L  R  A  L  H  M  K  E  H  P  D  Y  K  Y  R  P
 801 GCGCCGCAAGACCAAGACGCTGCTCAAGAAGGACAAGTACTCGCTGGCCGGCGGGCTGCTAGCGGCCGGCGCGGGTGGCGGCGGCGCGGCCGTGGCCATG  900
       R  R  K  T  K  T  L  L  K  K  D  K  Y  S  L  A  G  G  L  L  A  A  G  A  G  G  G  G  A  A  V  A  M
 901 GGTGTGGGCGTGGGCGTCGGGGCGGCGGCGGTGGGCCAGCGCTTGGAGAGCCCAGGCGGCGCGGCGGGAGGAGGCTACGCGCATGTCAACGGCTGGGCTA 1000
       G  V  G  V  G  V  G  A  A  A  V  G  Q  R  L  E  S  P  G  G  A  A  G  G  G  Y  A  H  V  N  G  W  A  N
1001 ACGGCGCCTACCCCGGCTCGGTGGCCGCGGCGGCTGCGGCCGCGGCCATGATGCAGGAGGCACAGCTGGCCTACGGGCAGCACCCAGGCGCGGGCGGCCG 1100
         G  A  Y  P  G  S  V  A  A  A  A  A  A  A  A  M  M  Q  E  A  Q  L  A  Y  G  Q  H  P  G  A  G  G  R
1101 GCACCCGCACGCACACCCGGCGCACCCGCACCCGCACCATCCGCACGCGCATCCTCACAACCCGCAGCCCATGCACCGCTACGACATGGGCGCGCTGCAG 1200
       H  P  H  A  H  P  A  H  P  H  P  H  P  H  H  P  H  A  H  P  H  N  P  Q  P  M  H  R  Y  D  M  G  A  L  Q
1201 TACAGCCCCATCTCCAACTCTCAGGGCTACATGAGCGCGTCGCCTTCGGGCTACGGCGGCATCCCTTACGGCGCCGCGGCCGCCGCCGCCGCTGCGG 1300
       Y  S  P  I  S  N  S  Q  G  Y  M  S  A  S  P  S  G  Y  G  G  I  P  Y  G  A  A  A  A  A  A  A  A  G
1301 GCGGCGCGCACCAGAACTCGGCGGTGGCGGCAGCGGCAGCCGCGGCAGCCGCGTCGTCGGGGGCCCTGGGCGCCCTCGGATCTCTGGTCAAGTCGGAGCC 1400
         G  A  H  Q  N  S  A  V  A  A  A  A  A  A  A  A  A  S  S  G  A  L  G  A  L  G  S  L  V  K  S  E  P
1401 GAGCGGCAGTCCGCCGGCCCCGGCTCACTCACGGGCACCGTGTCCCGGGGACCTGCGCGAGATGATCAGCATGTACCTGCCGGCCGGCGAGGGTGGCGAC 1500
       S  G  S  P  P  A  P  A  H  S  R  A  P  C  P  G  D  L  R  E  M  I  S  M  Y  L  P  A  G  E  G  G  D
1501 CCGGCGGCGGCAGCGGCTGCGGCGGCCCAAAGCCGGCTGCACTCGCTGCCACAGCACTACCAGGGCGCGGGCGCGGGCGTCAACGGCACGGTGCCCCTGA 1600
       P  A  A  A  A  A  A  A  A  Q  S  R  L  H  S  L  P  Q  H  Y  Q  G  A  G  A  G  V  N  G  T  V  P  L  T
1601 CGCACATCTAGCGCCGCGGGGACGCCGGGGACACTGCGGCTTAAGGCCGGCGCCCCGGCGACGAAGAGCGAGGCCTGCGCCCCAGCCTCCAGAGCCCGAC 1700
         H  I  *
1701 TTTGTACCGAGGTCCCCGCGCTCTCGATAAAAGGCCGCTCTGGAGAGCCGAGCGCCAGGTGACATCTGCCCCCATCACCTTCCCCAGGACTCCGAGGCGC 1800
1801 TGACACCAGACTGGCCTCTTAGACTGAACTTTGGTGTTTTCATGAGACCTTTTGTACAGTATTTATCGTCCGCAGAGGAGGCACACAGCGTTTTCTCGGC 1900
1901 TTCGGAGGACAAAAGACAAAAACCCAGCGAGGCGATGCCAACTTTTGTATGACTGCCGGCTCTGTAACTTTTTCCGGGGTTTACTTCCCGCCAGCTCTTC 2000
2001 TGCCTGAGGCCGAGTGACGGACCTCGAGCCCTTCTCACTTGTTATAAATCTAAGTAAGGCAGATCCAAACATTTACAAGTTTTTTGTAGTTGTTACCGCT 2100
2101 CTTTTGGGTTGGTTGGTTAATTTATACCGCAATCCCCTCTCAGACGGTGGAGTTATATTCTGGGTTTTGTAAATCTCTGTATCCGAGCATTTCCAATTTT 2200
2201 TTGTTTTGTTTTGTATTATTTCTTGTAAATGCGTTGTGACATTTTTATTTTAGGCGTTGCGATACGGGGGGAAGAGGAGTCGGATGTTGTACATAGCCTG 2300
2301 CAAGTCTTTCATCTAAAAGCAAAAACAAAGAGAGATACCCCCAAAATGCATCAAATTTGAACAATACATTTAAGAG  2376
```

Sox-2

```
   1 CTTGTTCAAAAAAGTATCAGGAGTTGTCAAGGCAGAGAAGAGAGTGTTTGCAAAAAGGGAAAAGTACTTTGCTGCCTCTTTAAGACTAGGGCTGGGAGAA  100
 101 AGAAGAGGAGAGAGAAAGAAAGGAGAGAAGTTTGGAGCCCGAGGCTTAAGCCTTTCCAAAAACTAATCACAACAATCGCGGCGGCCCGAGGAGGAGAGCG  200
 201 CCTGTTTTTTCATCCCAATTGCACTTCGCCCGTCTCGAGCTCCGCTTCCCCCCAACTATTCTCCGCCAGATCTCCGCGCAGGGCCGTGCACGCCGAGGCC  300
 301 CCCGCCCGCGGCCCCTGCATCCCGGCCCCCGAGCGCGGCCCCCACAGTCCTGGCCGGGCCGAGGGTTGGCGGCCGCCGGCGGGCCGCGCCCCGCCCCAGC  400
 401 GCGCGCATGTATAACATGATGGAGACGGAGCTGAAGCCGCCGGGCCCGCAGCAAGCTTCGGGGGGCGGCGGCGGAGGAGGCAACGCCACGGCGGCGGCGA  500
           M  Y  N  M  M  E  T  E  L  K  P  P  G  P  Q  Q  A  S  G  G  G  G  G  G  N  A  T  A  A  A  T
 501 CCGGCGGCAACCAGAAGAACAGCCCGGACCGCGTCAAGAGGCCCATGAACGCCTTCATGGTATGGTCCCGGGGGCAGCGGCGTAAGATGGCCCAGGAGAA  600
       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  E  N
 601 CCCCAAGATGCACAACTCGGAGATCAGCAAGCGCCTGGGCGCGGAGTGGAAACTTTGTCCGAGACCGAGAAGCGGCCGTTCATCGACGAGGCCAAGCGG  700
       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  A  K  R
 701 CTGCGCGCTCTGCACATGAAGGAGCACCCGGATTATAAATACCGGCCGCGGCGCAAAACCAAGACGCTCATGAAGAAGGATAAGTACACGCTTCCCGGAG  800
       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  T  L  P  G  G
 801 GCTTGCTGGCCCCCGGCGGGAACAGCATGGCGAGCGGGGTTGGGGTGGGCGCCGGCCTGGGTGCGGGCGTGAACCAGCGCATGGACAGCTACGCGCACAT  900
         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  D  S  Y  A  H  M
 901 GAACGGCTGGAGCAACGGCAGCTACAGCATGATGCAGGAGCAGCTGGGCTACCCGCAGCACCCGGGCCTCAACGCTCACGGCGCGGCACAGATGCAACCG  1000
       N  G  W  S  N  G  S  Y  S  M  M  Q  E  Q  L  G  Y  P  Q  H  P  G  L  N  A  H  G  A  A  Q  M  Q  P
1001 ATGCACCGCTACGACGTCAGCGCCCTGCAGTACAACTCCATGACCAGCTCGCAGACCTACATGAACGGCTCGCCCACCTACAGCATGTCCTACTCGCAGC  1100
       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  T  Y  S  M  S  Y  S  Q  Q
1101 AGGGCACCCCCGGTATGGCGCTGGGCTCCATGGGCTCTGTGGTCAAGTCCGAGGCCAGCTCCAGCCCCCCCGTGGTTACCTCTTCCTCCCACTCCAGGGC  1200
         G  T  P  G  M  A  L  G  S  M  G  S  V  V  K  S  E  A  S  S  S  P  P  V  V  T  S  S  S  H  S  R  A
1201 GCCCTGCCAGGCCGGGGACCTCCGGGACATGATCAGCATGTACCTCCCCGGCGCCGAGGTGCCGGAGCCCGCTGCGCCCAGTAGACTGCACATGGCCCAG  1300
       P  C  Q  A  G  D  L  R  D  M  I  S  M  Y  L  P  G  A  E  V  P  E  P  A  A  P  S  R  L  H  M  A  Q
1301 CACTACCAGAGCGGCCCGGTGCCCGGCACGGCCATTAACGGCACACTGCCCCTGTCGCACATGTGAGGGCTGGACTGCGAACTGGAGAAGGGGAGAGATT  1400
       H  Y  Q  S  G  P  V  P  G  T  A  I  N  G  T  L  P  L  S  H  M  *
1401 TTCAAAGAGATACAAGGGAATTGGGAGGGGTGCAAAAAGAGGAGAGTAGGAAAAATCTGATAATGCTCAAAAAGAAAAAAACCACCAATCCCATCCAAAT  1500
1501 TAACGCAAAAACCGTGATGCCGACTAGAAAACTTTTATGAGAGATCTTGGGACTTCTTTTTGGGGGACTATTTTTGTACAGAGAAAACCTGAGGGCGGCG  1600
1601 GGGAGGGCGGGGGGAATCGGACCATGTATAGATCTGGAGGAAAAAAAACTACGCAAAACTTTTTTTTAAAGTTCTAGTGGTACGTTAGGCGCTTCGCAGGGA  1700
1701 GTTCGCAAAAGTCTTTACCAGTAATATTTAGAGCTAGACTCCGGGCGATGAAAAAAAAAGTTTTAATATTTGCAAGCAACTTTTGTACAGTATTTATCGA  1800
1801 GATAAACATGGCAATCAAATGTCCATTGTTTATAAGCTGAGAATTTGCCAATATTTTTCGAGGAAAGGGTTCTTGCTGGGTTTGATCTGCAGCTTAAATT  1900
1901 TAGGACCGTTACAAACAAGGAAGGAGTTTATTCGGATTTGAACATTTTAGTTTTAAAATTGTACAAAAGGAAAACATGAGAGCAAGTACTGGCAAGACCG  2000
2001 TTTTCGTGGTCTTGTTTAAGGCAAACGTTCTAGATTGTACTAAATTTTTAACTTACTGTTAAAGGCAAAAAAAAAATGTCCATGCAGGTTGATATCGTTG  2100
2101 GTAATTTATAATAGCTTTTGTTCAATCCTACCCTTTCATTTTGTTCACATAAAAAATATGGAATTACTGTGTTTGAAATATTTTCTTATGGTTTGTAATA  2200
2201 TTTCTGTAAATTGTGATATTTTAAGGTTTTTCCCCCCTTTTATTTTCCGTAGTTGTATTTTAAAAGATTCGGCTCTGTTATTGGAATCAGGCTCCGAGAA  2300
2301 TCCATGTATATATTTGAACTAATACCATCCTTATAACAGCTACATTTTCAACTTAAGTTTTTACTCCATTATGCACAGTTTGAGATAAATAAATTTTTGA  2400
2401 AAAAAAAAAAAAAAAAA  2418
```

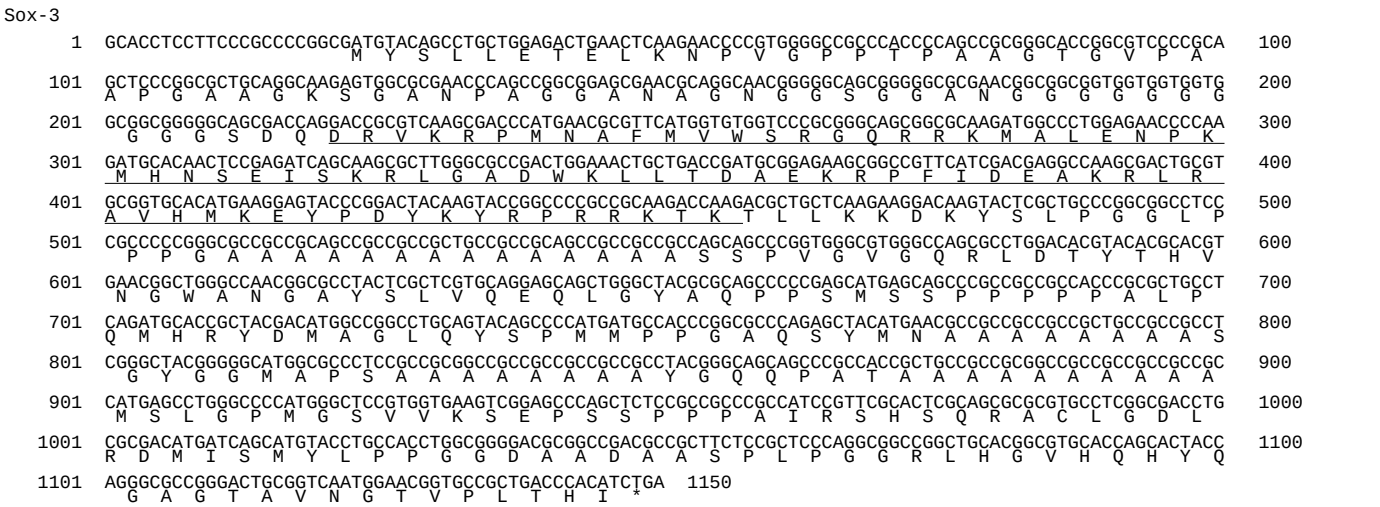


Fig. 1. Sequence of *Sox-1*, *Sox-2* and *Sox-3*. Nucleotide sequences have been obtained from cDNAs and genomic clones for *Sox-1*, *Sox-2* and *Sox-3*. The predicted amino-acid sequence is shown under the nucleotide sequence in single letter code. The characteristic *Sry*-related HMG box is underlined in each case. Accession numbers: *Sox-1*, X94126; *Sox-2*, X94127; *Sox-3*, X94125.

2 transcript identified here can also be detected in the embryo (S. Sockanathan, unpublished data). The *Sox-3* transcript in embryonic head RNA at 10.5 dpc and 11.5 dpc was found to be 2.3 kb in length. No expression was detected in other adult tissues examined except for testis (Fig. 4B).

In vitro DNA-binding properties of SOX-1, SOX-2 and SOX-3

It has been shown that SRY protein binds the consensus sequence AACAAT in vitro (Giese et al., 1992; Harley et al., 1994). Given the high degree of similarity between the HMG boxes of SRY, SOX-1, SOX-2 and SOX-3 and their overlapping patterns of expression (see below), it was of interest to determine if these proteins could also bind the AACAAT sequence with the same relative affinities. Identical size HMG box-GST fusion proteins were purified for all four proteins and gel-retardation assays were carried out using increasing amounts of protein (Fig. 5). SOX-3, as well as SOX-1 and SOX-2, are able to form specific complexes with the AACAAT motif. However, although SOX-3 shows the greatest sequence homology to SRY within the HMG box, it is clear that much weaker binding is exhibited by SOX-3 compared to SRY. Poor binding is also shown by SOX-1 but, in contrast, SOX-2 and SRY were found to bind the target oligonucleotide at similar affinities.

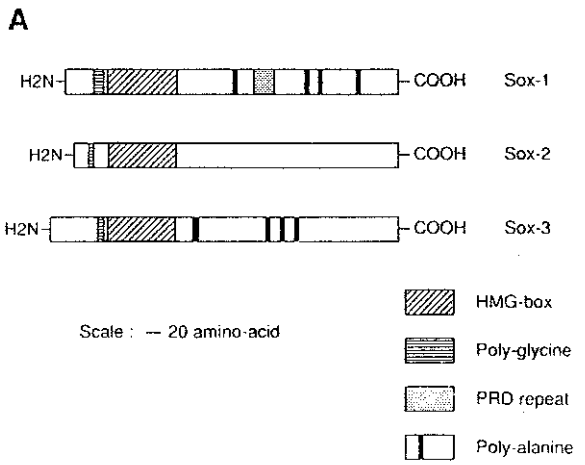
Expression of *Sox-3* during development

In situ hybridisation

A detailed analysis of the expression patterns of *Sox-1*, *Sox-2* and *Sox-3* was carried out by in situ hybridisation. The data obtained for *Sox-1* and *Sox-2* will be presented elsewhere; however both genes are expressed in the CNS, with *Sox-2* showing additional expression in the sensory placodes, the PNS and gut endoderm. The in situ hybridisation analysis for *Sox-3* shown in Fig. 6 covers mouse embryonic stages between 8.5 and 13.5 dpc. *Sox-3* shows overlapping domains of expression with *Sox-1* and *Sox-2* within the CNS and the sensory placodes throughout this period. In 8.5-dpc embryos, expression is detected throughout the neuroectoderm (Fig. 6A,B). This pattern is unchanged at 9.5 dpc, with the exception of the optic vesicles, which show a gradient of expression with the lowest level distally (Fig. 6C-E). Some expression is also evident in the olfactory placode (Fig. 6C). Transverse sections of a 10.5-dpc embryo show that *Sox-3* continues to be expressed along the full length of the developing CNS, including brain and spinal cord. However, frontal sections of a similar stage embryo reveal some discontinuity in the pattern at the level of the hindbrain, with areas of weaker and stronger expression apparently contained within the boundaries of specific rhombomeres (Fig. 6F). At 11.5 dpc, *Sox-3* expression is maintained throughout the fetal brain, but not in the optic cup, as seen in transverse sections through the head (Fig. 6G). However, by 13.5 dpc the expression has become restricted to the ependymal layer, where undifferentiated neural progenitor cells are still actively dividing (data not shown).

RNase protection studies

Our results by in situ hybridisation demonstrate that the main sites of expression of *Sox-3* are in the developing CNS, like *Sox-1* and *Sox-2*. In addition to this, some overlapping sites of expression are also observed between the genes elsewhere; for example, for *Sox-2* and *Sox-3* in the sensory placodes and, for *Sox-1* and *Sox-2*, in the developing lens (Collignon, 1992). In all cases, no expression was observed in the urogenital ridge where *Sry* is expressed. However, it has been shown that sensitive techniques such as RT-PCR or RNase protection are more reliable ways of detecting *Sry* in the genital ridge due to its low level of expression (Koopman et al., 1991; Hacker et al., 1995). RNase protection assays were therefore employed to examine the expression of *Sox-3*, as well as *Sox-1* and *Sox-2*, in the urogenital ridge, as these are quantitative assays not prone to problems of signal to noise ratio. Fig. 7 compares the expression of *Sox-1*, *Sox-2* and *Sox-3* with *Sry* in urogenital ridges at 11.5 dpc. It can be seen that all four of the members of the subfamily are expressed in both male and female genital ridges, with the exception of *Sry*, which is male-specific. RNA from embryonic head was used as a positive control for *Sox-*



B

```
SOX-1   128 TLLKKDKYSLAGGLLAAGAGGGGAAVAMGVGVGVGAAAVGQRLESPGGAA 177
SOX-2   120 TLMKKDKYTLPGGLLAPGGNSMASGVGVGAGLG...AGVNQRMDS..... 161
SOX-3   146 TLLKKDKYSLPGGLPPPGAAAAAAAAAAAAAA.SSPVGVGQRLDT..... 188

        178 GGGYAHVNGWANGAYPGSVAAAAAAAAMMQEAQLAYGQHPGAGGRHPHAH 227
        162 ...YAHMNGWSNGSYS..........MMQE.QLGYPQHPGL.....NAH 191
        190 ...YTHVNGWANGAYS..........LVQE.QLGYAQPP..SMSSPPPP 222

        228 PAHPHPHHPHAHPHNPQPMHRYDMGALQYSPISNS..QGYMSASP..... 270
        192 GA..........AQMQPMHRYDVSALQYNSMTSS..QTYMNGSP..... 223
        223 PAL...........PQ.MHRYDMAGLQYSPMMPPGAQSYMNAAAAAAAA 259

        271 SGYGGIPYGAAAAAAAAAGGAHQNSAVAAAAAAAAASSGALGALGSL... 317
        224 .............TYSMSYSQQGTP.............GMALGSMGSV 245
        260 SGYGGMAPSAAAAAAAA..YGQQPATAAAAAAAAAA.....MSLGPMGSV 302

        318 VKSEPSGSPPAPA...HS.RAPC.PGDLREMISMYLPAGEGGDPAAAAAA 362
        246 VKSEASSSPPVVTSSSHS.RAPCQAGDLRDMISMYLPGAEVPEPAA.... 290
        303 VKSEPSSPPPAIRS..HSQRA.C.LGDLRDMISMYLPP..GGDAADAASP 346

        363 AAQSRLHSLPQHYQGAGAG...VNGTVPLTHI* 390
        291 ..PSRLHMA.QHYQSGPVPGTAINGTLPLSHM* 319
        347 LPGGRLHGVHQHYQGAG...TAVNGTVPLTHI* 375
```

C

Similarity to HMG box of	SOX-1	SOX-2	SOX-3	SOX-4	SOX-5
SOX-3	96%	95%	100%	78%	69%
S.m. SRY	83%	83%	83%	77%	61%
Wal. SRY	81%	81%	85%	77%	68%
Hum. SRY	86%	85%	90%	74%	74%
Mus SRY	81%		82%	77%	75%
SOX-4			78%	100%	
SOX-5			69%		100%

Fig. 2. Sequence comparisons between SOX-1, SOX-2, SOX-3 and SRY. (A) Schematic representation of the predicted amino-acid sequences of SOX-1, SOX-2 and SOX-3. Each protein has a unique combination of structural motifs. (B) Comparison of the sequences C-terminal to the HMG box of SOX-1, SOX-2 and SOX-3. The poly(alanine) stretches shared by SOX-1 and SOX-3 are in bold. The His-Pro repeats in SOX-1 are underlined. (C) SRY homologues are more closely related to Sox-3. Sequence similarity between SRY homologues of two marsupials, *Sminthopsis macroura* (S.m.) and *Macropus eugenii* (wallaby, Wal.) (Foster et al., 1992), mouse (Mus.) SRY, human SRY (Sinclair et al., 1990) and the type I HMG box of SOX-1, SOX-2 and SOX-3, the type II HMG box of SOX-4 and the type IV HMG box of SOX-5 (Denny et al., 1992a). Among type I sequences, SOX-3 appears to be most closely related to the SRY homologues.

1, Sox-2 and *Sox-3* expression whereas adult liver was used as a negative control. Testis RNA was used as a positive control for *Sry* expression and it can be seen that both *Sox-1* and *Sox-3* are also expressed in this tissue, unlike *Sox-2*.

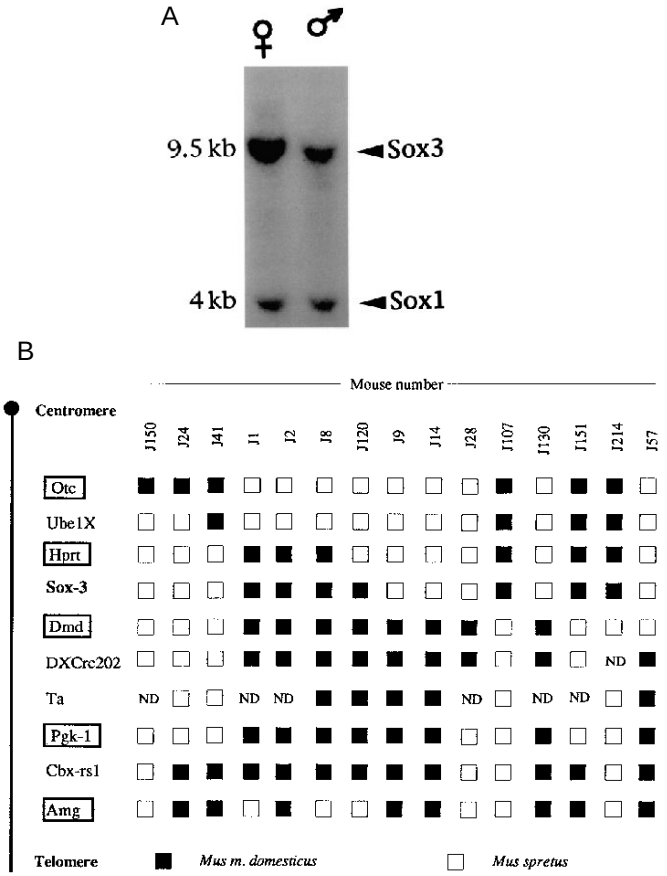


Fig. 3. Chromosomal localization of *Sox-3*. (A) Southern blot of *Bgl*II-digested mouse genomic DNA probed with *Sox-3* and *Sox-1* cDNA sequences. (B) *Sox-3* maps on the mouse X chromosome between *Hprt* and *Dmd*. A *Hin*DIII RFLV was used to distinguish the *Sox-3* allele in *Mus musculus domesticus* (6.1 kb) and *Mus spretus* (5.3 kb). A panel of 19 extensively analysed interspecific backcross mice with recombination breakpoints distributed throughout the X chromosome (Kay et al., 1991) was used to map the *Sox-3* gene with respect to X chromosome anchor loci. The haplotypes of the recombinant X chromosome of 15 representative backcross progeny are shown for each of the probes used. *Sox-3* maps between *Hprt* and *Dmd* at a genetic distance 7.4± centiMorgans distal to *Hprt*. Anchor loci are boxed.

We have shown so far that SOX-3 has similar properties to SRY in that they are both expressed in the urogenital ridge and are able to form complexes on the AACAAT motif in vitro. SOX-1 and SOX-2 also share these properties; indeed SOX-2 has a DNA-binding affinity more similar to SRY than SOX-3. However, SOX-2 is known to be expressed in the primordial germ cells of the genital ridge and not within the Sertoli cell precursors where *Sry* is found (S. Sockanathan, unpublished data). It is therefore important to identify in which of the two cell types *Sox-3* is expressed. To address this question, we used RNase protection assays on RNA from genital ridges of normal and W^e homozygous embryos. These mice show very low, if any, colonisation of the gonad by germ cells, resulting in the formation of genital ridges, which are composed primarily of somatic cell lineages (Buehr et al., 1993). The results presented in Fig. 7B show that *Sox-3* is expressed by somatic cells (W^e mutants) at similar levels to those seen in wild-type genital

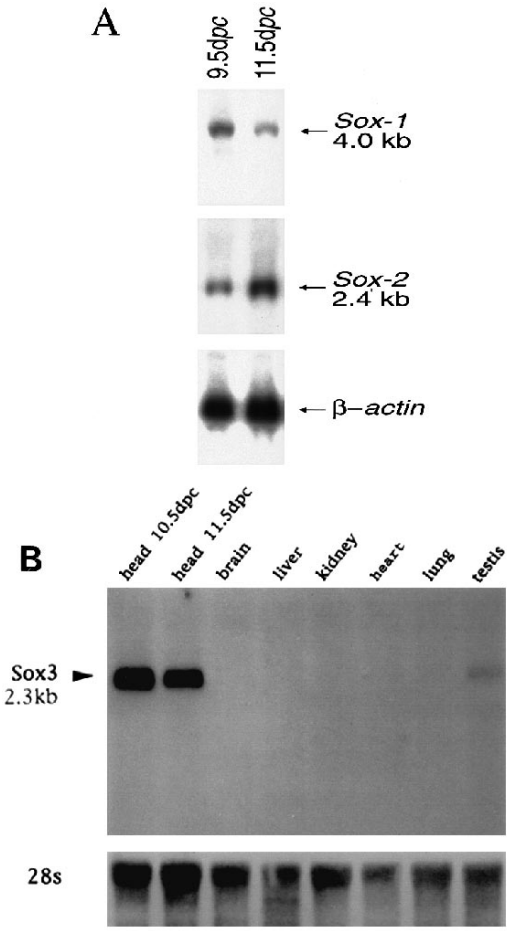


Fig. 4. Northern blot analysis of *Sox-1*, *Sox-2* and *Sox-3*. (A) A 2.4-kb *Sox-2* specific transcript and a 4.0-kb *Sox-1* transcript were detected in RNA prepared from 9.5-dpc and 10.5-dpc embryos. The same filter was subsequently hybridised with a *β-actin* probe to ensure integrity of the RNA samples. (B) A 2.3-kb *Sox-3*-specific transcript was detected in RNA from 10.5-dpc and 11.5-dpc embryonic heads and from adult testis (top). For each sample, 10 μg of total RNA was loaded and cross-hybridisation of the probe to 28S tRNA, seen before RNase treatment (bottom), shows that similar amounts of RNA were transferred to the filter in each lane.

ridges. This demonstrates that *Sox-3* is likely to be expressed in the same cell type as *Sry*. *Sox-3* transcripts were also found to be expressed by developing limb buds at this stage, but not in other tissues (Fig. 7 and data not shown). As the limb bud sample was much larger than the genital ridge samples, the signal is likely to represent a very low level of *Sox-3* expression.

DISCUSSION

In this report, we present a comparison of the mouse *Sox-1*, *Sox-2* and *Sox-3* genes, which together make up the members of the *Sox* family most related to the testis-determining gene *Sry*. We confirm that of the three genes, *Sox-3* bears the highest homology to *Sry* and that it is located on the X-chromosome, consistent with the proposal that *Sry* may have evolved from *Sox-3*. We have adressed this proposal further by examining

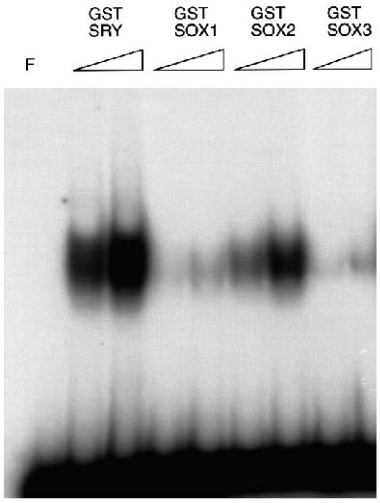


Fig. 5. HMG boxes from SRY, SOX-1, SOX-2 and SOX-3 show different affinities to target DNA. The open triangles indicate increasing amounts (0.2 μg to 1 μg) of purified GST-HMG box fusion proteins incubated with an oligonucleotide probe containing the binding site AACAAT. SRY and SOX-2 bind with similar affinities, whereas SOX-1 and SOX-3 bind poorly. F, free probe.

sites of expression and DNA-binding properties. The main site of expression of *Sox-3* is within the developing CNS, in a pattern overlapping that of *Sox-1* and *Sox-2*, suggesting a role for all these genes in neural development. But *Sox-1*, *Sox-2* and *Sox-3* transcripts are also detected within the developing urogenital ridge. Furthermore, we show that *Sox-3* is most likely to be expressed within the same cell type as *Sry*, suggesting that the two gene products could cooperate or compete. In gel mobility shift assays, SOX-1, SOX-2 and SOX-3 are able to bind to the same consensus motif as SRY; however SOX-3 shows much weaker affinity to the site in comparison to either SRY or SOX-2. Our findings suggest that *Sry* may have evolved from this subfamily of *Sox* genes and that of the three members, *Sox-3* is the probable ancestor of *Sry*, because of its sequence homology, chromosomal location and expression, but there are significant differences in the properties of the proteins.

SOX-1, SOX-2 and SOX-3 can be grouped into the same subfamily, along with SRY, due to the high homology (approximately 82%) exhibited within the HMG boxes of the proteins. For SOX-1, SOX-2 and SOX-3, sequence similarity also extends outside the HMG box, although each protein has its own characteristics. Thus, SOX-2 is devoid of poly(alanine) repeats such as those found in SOX-1 and SOX-3, while SOX-1 contains a PRD repeat (His-Pro) (Frigerio et al., 1986), which is absent in both SOX-2 and SOX-3. The similarities in structure and sequence suggest that these genes are evolutionarily related and may have arisen from duplications of a single ancestral gene. This must have been a fairly ancient event as the genes map to different chromosomes, but once duplicated they seem to have acquired important individual functions as there is extremely high sequence conservation amongst *Sox-2* genes and *Sox-3* genes between chicken, mouse and man (Stevanovic et al, 1993, 1994; Uwanogho et al., 1995; Kamachi et al., 1995).

As shown here, the main site of *Sox-3* expression is in the

developing CNS from early stages, which is also true for *Sox-1* and *Sox-2*. (see also J. Collignon et al., unpublished data; Collignon, 1992). All three of the genes show overlapping expression to some degree elsewhere in the embryo, e.g. in the genital ridge (see later); however, differences are also evident. For example, both *Sox-3* and *Sox-2* are expressed in the olfactory placode, but *Sox-2* shows additional expression in other placodal tissue as well as the gut endoderm. Within the

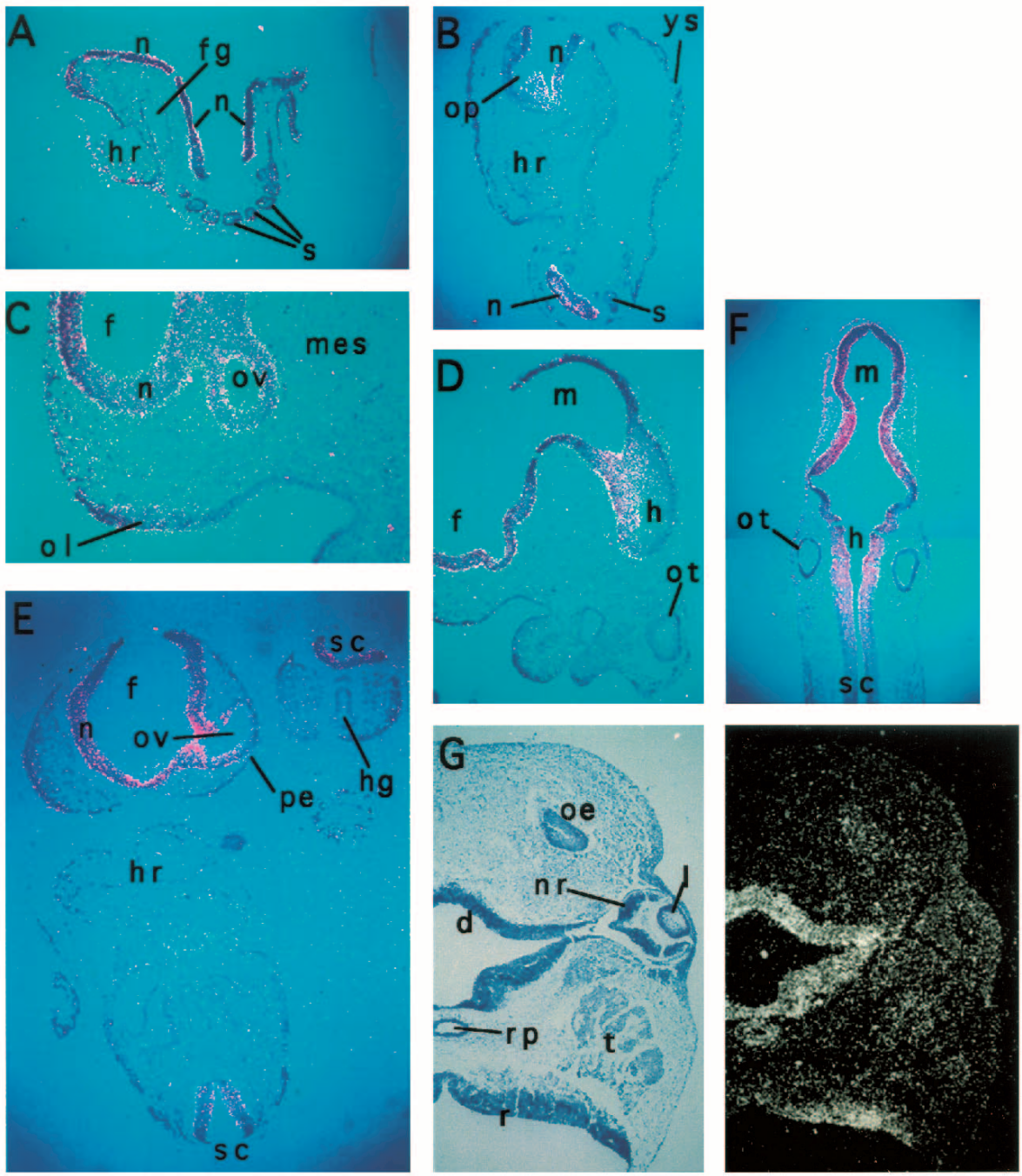


Fig. 6. In situ hybridisation of *Sox-3* expression between 8.5 dpc and 11.5 dpc. *Sox-3* expression is detected throughout the early developing central nervous system, with stronger expression anteriorly. The signal is weak in the most dorsal part of the neural tube and is much reduced or absent in the floorplate. *Sox-3* transcripts are seen in the olfactory placode, but no signal is seen in the developing eye and the otic vesicle. (A) Sagittal section of an 8.5-dpc embryo. (B) Frontal section of a 9.0-dpc embryo at eye level. (C) Sagittal section of a 9.5-dpc embryo showing olfactory placode. (D) Sagittal section of a 9.5-dpc embryo showing fore-, mid- and hindbrain and otic vesicle. (E) Frontal section of a 9.5-dpc embryo at eye level showing expression in forebrain and spinal cord. (F) Frontal section of a 10.5-dpc embryo showing CNS expression. (G) Transverse section through the region of the eye at 11.5 dpc in brightfield (left) and darkfield (right). n, neuroectoderm; s, somites; hr, heart; fg, foregut; op, optic pit; ys, yolk sac; ov, optic vesicle; ol, olfactory placode; mes, mesenchyme; f, forebrain; m, midbrain; h, hindbrain; ot, otic vesicle; sc, spinal cord; pe, lens placodal ectoderm; hg, hindgut; l, lens vesicle; nr, neural retina; oe, olfactory epithelium; d, wall of diencephalon; r, wall of rhombencephalon; rp, Rathke's pouch; t, trigeminal ganglion.

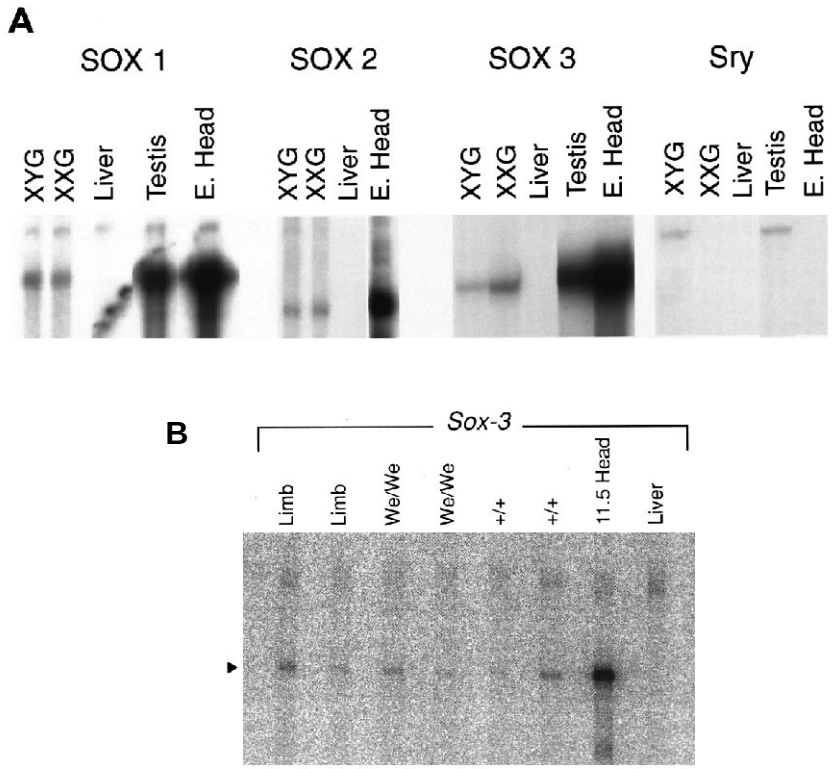


Fig. 7. RNase protection assay using anti-sense probes to *Sox-1*, *Sox-2*, *Sox-3* and *Sry*. (A) *Sox-1*, *Sox-2* and *Sox-3* transcripts are evident in both male and female genital ridges (XYG and XXG) whereas *Sry* transcripts are male-specific. *Sox-3* appears to be expressed more abundantly in female compared to male genital ridge. This was found to be reproducible in all six separate experiments carried out. Positive controls used for *Sox-1* and *Sox-3* were testis and embryonic head RNA, whereas only the latter was used for *Sox-2*, and the former for *Sry*. Adult liver RNA was used as a negative control in all cases. (B) *Sox-3* expression is somatic and not dependent on the presence of germ cells. RNA from genital ridges from either W^e homozygotes (We/We) or wild-type embryos (+/+), show *Sox-3*-protected fragments (352 nt), as do limb and head samples from 11.5-dpc embryos. These embryos were not sexed. Adult liver RNA is negative.

CNS there are also fine distinctions between the genes where, for example, *Sox-1* exhibits a sharp boundary of expression in the optic stalk and is not expressed at all in the optic vesicle; whereas *Sox-3* expression decreases gradually between the forebrain and the distal part of the vesicle. These differences are also found in adult tissues where *Sox-3* expression was not found in adult brain, in contrast to *Sox-1* and *Sox-2*, but is detected in testis like *Sox-1*, but not *Sox-2*.

The pattern of expression of *Sox-3* suggests that the gene functions mainly in the developing CNS. Stevanovic et al. (1993) described human patients carrying deletions of a region of the X chromosome including SOX3. These individuals show mental retardation, which is not apparent in patients deleted for similar regions of the X chromosome that do not include SOX3. Other mental retardation syndromes have also been mapped to the same region of the X chromosome, such as Borjeson-Forssman-Lehmann syndrome (Turner et al., 1989) and the X-linked centronuclear myotubular myopathy (Liechti-Gallati et al., 1991). SOX3 is therefore a candidate gene for one or more of these syndromes. The widespread expression of *Sox-3* seen throughout the CNS of early mouse embryos might suggest a more severe phenotype than that observed in the human patients deleted for the gene. However, since the domain of expression of *Sox-3* appears to be contained within those of both *Sox-1* and *Sox-2*, it is possible that either of these two genes could partially compensate for the loss of *Sox-3* function. Partial redundancy clearly explains the restricted phenotype of mice, with mutations in only one member of other gene families showing overlapping expression, such as with *En-1* or *En-2* (Millen et al., 1994; Wurst et al., 1994). Also, as discussed below, it is possible that SOX-3 competes inefficiently for target sites with the other SOX proteins in some parts of its expression domain, and is therefore critically important only in a subset of this domain.

We have shown that *Sox-1*, *Sox-2* and *Sox-3* are expressed in the urogenital ridge at the point of gonadal differentiation, similar to the founder member of the subfamily, *Sry*. This raises the question of whether any of these genes also play a role in sex determination. Recent work has provided a precedent for this. *Sox-9* falls into a distinct subfamily, less related to *Sry* (Wright et al., 1993). The gene maps to chromosome 11 in mice and 17 in humans (Foster et al., 1994; Wagner et al., 1994; Wright et al., 1995) and has been clearly implicated in development of cartilage, as it is abundantly expressed during chondrogenesis (Wright et al., 1995). However, mutations in the human gene can lead to XY female sex reversal as well as to a severe dwarfism syndrome, Campomelic dysplasia (Foster et al., 1994; Wagner et al., 1994). Its involvement in sex determination is not understood, although it is also expressed in the urogenital ridge throughout gonadal development (Wright et al., 1995 and our own unpublished observations).

At 11.5 dpc, when *Sry* is thought to act, the indifferent gonad consists of primordial germ cells that are not required for testis determination, and two bipotential somatic cell lineages that give rise to supporting and steroidigenic cell types (Buehr et al., 1993). *Sry* acts within the supporting cell precursors that give Sertoli cells in the testis and granulosa cells in the ovary. It is therefore important to identify if *Sox-1*, *Sox-2*, *Sox-3* or indeed *Sox-9* are expressed in the same cell type as *Sry*. It is not yet possible to distinguish supporting and steroidigenic cells at 11.5 dpc (indeed there could be a common precursor cell type), but it is possible to tell whether transcripts are due to germ cells or somatic cells. We have presented evidence that

SOX-2 is most similar in its DNA-binding activity to SRY when analysed in vitro. However, we have recently shown that SOX-2 is expressed within the germ cells of the genital ridge, thus excluding it from having a role in sex determination (S. Sockanathan, unpublished data). The cell-specific expression of *Sox-1* remains unknown; however, the analysis presented here of *Sox-3* expression in genital ridges of W^e/W^e homozygous mutant mice, which lack germ cells, reveals that *Sox-3* is expressed in the somatic cells of the genital ridge, at a level equivalent or greater than *Sry* (see also Koopman et al., 1990). Also, like *Sry*, the expression appears to be dynamic with no transcripts detectable by 12.5 dpc (A. Hacker, unpublished data). These findings raise two questions: what is the relationship between *Sox-3* and *Sry*, and could *Sox-3* also function in sex determination?

In humans, *SOX3* is not required for testis determination. Despite having small testes, patients deleted for the gene are clearly male (Stevanovic et al., 1993). However, the gene could have a role in ovarian determination. One simple hypothesis would be that the action of SOX-3 protein on its gene target(s) is a critical step in the normal genetic pathway leading to differentiation of an ovary, and that in a male, SRY protein competes for the same target site(s). We have shown here that lower concentrations of SRY are required to bind to target DNA sequences compared to SOX-3, so it is easy to imagine that the latter will be displaced by SRY. Alternatively, there may be other proteins or cofactors that interact with SOX-3 protein to modulate its sequence specificity or affinity in vivo, in which case it could act on a different set of target genes. Such interactions could involve the HMG box itself, or other domains of the protein. In contrast to SRY, the high degree of conservation throughout the entire length of the mouse, human, marsupial and chick SOX-3 proteins, including the poly(alanine) repeat regions, suggests that the N-and C-terminal domains have important functions. It is also possible that two (or more) SOX proteins cooperate or compete at the same cis-acting binding site, giving different types of response of the target gene (see, for example, Jaynes and O'Farrell, 1988; Han et al., 1989).

Sequence comparisons between *Sox-3* and *Sry* or their proteins show homology to be restricted to the HMG box domain. With 82% sequence identity between their HMG boxes, SOX-3 is the SOX protein most closely related to both the human and mouse SRY proteins. For the purpose of establishing evolutionary relationships with *Sry*, sequence homology outside the HMG box is irrelevant since these non-box regions, as well as regions outside the ORF, show little or no conservation when *Sry* genes from different species are compared with each other (Tucker and Lundrigan, 1993; Whitfield et al., 1993). Similar findings were presented by Foster and Graves (1994) on comparisons of marsupial and human SRY and SOX-3 sequences.

The many similarities between *Sox-3* and *Sry*, including some aspects of their expression, and the location of *Sox-3* on the X chromosome, are consistent with the notion that the two genes are evolutionarily related. Structurally distinct sex chromosomes are a common feature of genetic mechanisms of sex determination. Studies of XY pairs in various organisms led to the hypothesis that X and Y chromosomes were originally homologous, but became genetically distinct during the course of evolution because of specific constraints imparted by the sex-determining mechanism (Ohno, 1967; Charlesworth, 1991; Hodgkin, 1992). The evolution of a sex-determining locus would require the creation of a non-recombining region, which would then accumulate mutations by a process known as Muller's Ratchet (Muller, 1964; Felsenstein, 1974). With time, this mechanism would lead to the evolution of distinct sex chromosomes. In agreement with this hypothesis, several of the Y-linked genes cloned in human or in mouse (for example *ZFY*, *RPS4Y* or *Ube1-y1*) have been found to have X-linked related sequences (for a review, see Graves and Schmidt, 1992). Furthermore, recent experimental evidence obtained in *Drosophila* now supports this model by demonstrating direct causality between the presence of a sex-determining gene and the evolution of an heterologous pair of sex chromosomes (Charlesworth, 1992; Rice, 1992, 1994).

Comparison of the relative rates of divergence of *Sry* and its gene homologues led Griffiths (1991) to suggest that the evolution of *Sry* as a sex-determining gene is a recent event that occurred subsequent to the radiation of the different *Sox* gene subfamilies. Our results suggest that *Sry* evolved from the *Sox-1*, *Sox-2* and *Sox-3* subfamily. *Sox-3* is the most obvious candidate for a gene from which *Sry* evolved, but how could this have happened? It is conceivable that a mutation occurred in *Sox-3* such that the mutant gene product interfered with a pre-existing sex-determination mechanism, leading to a dominant male-determining mechanism. There is an experimental example of a similar situation in *Drosophila*, where mis-expression of the pair-rule gene *hairy*, which encodes an HLH protein, interferes with the process of sex-determination, despite playing no part in this process during normal development (Parkhurst et al., 1990). By taking over a sex-determining function, the mutant *Sox-3* allele may have been placed under evolutionary constraints very different from those it was facing previously. These new constraints may be the reason why the HMG box is the only conserved part of SRY (although adaptive selection for change in the other domains is also feasible). However, at present, it is equally plausible that *Sry* originated from another HMG-box-containing gene, perhaps even *Sox-9*, but came to resemble *Sox-3* through convergent evolution. Further examination of *Sox* gene sequences in lower vertebrates, and of *Sox-3* and *Sry* in prototheria (monotremes) and metatheria (marsupials), will be necessary to establish the evolutionary origins of *Sry*.

We are indebted to John Gubbay, who isolated the genomic clones for *Sox-3*, and to Vasso Episkopou for help in subcloning. We would also like to thank other members of the laboratories involved for their critical comments and encouragement. The work was supported by the Medical Research Council.

REFERENCES

Auffray, C. and Rougeon, F. (1980). Purification of mouse immunoglobulin heavy-chain messenger RNAs from total myeloma tumor RNA. *Eur. J. Biochem.* **107**, 303-324.

Berta, P., Hawkins, J. R., Sinclair, A. H., Taylor, A., Griffiths, B. L., Goodfellow, P. N. and Fellous, M. (1990). Genetic evidence equating SRY and the male sex determining gene. *Nature* **348**, 248-250.

Buehr, M., McLaren, A. and Darling, S. (1993). Proliferation and migration of primordial germ cells in *We/We* mouse embryos. *Developmental Dynamics* **198**, 182-189.

Capel, B. and Lovell-Badge, R. (1993). The Sry gene and sex determination in

mammals. In *Advances in Developmental Biology*, vol. 2 (ed. P. Wassarman), pp. 1-35. JAI Press Inc.

Capel, B., Swain, A., Nicolis, S., Hacker, A., Walter, M., Koopman, P., Goodfellow, P. and Lovell-Badge, R. (1993). Circular transcripts of the testis-determining gene Sry in adult mouse testis. *Cell* **73**, 1019-1030.

Charlesworth (1992). Why have Y chromosomes? *Current Biology* **2**, 515-516.

Charlesworth, B. (1991). The Evolution of Sex Chromosomes. *Science* **251**, 1030-1033.

Clépet, C., Schafer, A. J., Sinclair, A. H., Palmer, M. S., Lovell-Badge, R. and Goodfellow, P. N. (1993). The human *SRY* transcript. *Human Molecular Genetics* **2**, 2007-2012.

Collignon, J. (1992). *Study of a New Family of Genes Related to the Mammalian Testis Determining Gene*. PhD thesis, CNAA London.

Coriat, A. M., Muller, U., Harry, J. L., Uwanogho, D. & Sharpe, P. T. (1993). PCR amplification of *Sry*-related gene sequences reveals evolutionary conservation of the SRY-box motif. *PCR Methods and Applications* **2**, 218-222.

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. and 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.

Devereux, J., Haeberli, P. and Smithies, O. (1984). A comprehensive set of sequence analysis programs for the VAX. *Nucleic Acids Res.* **12**, 387-395.

Fahrner, K., Hogan, B. L. M. and Flavell, R. (1987). Transcription of *H-2* and *Qa* genes in embryonic and adult mice. *EMBO J.* **6**, 1265-1271.

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

Felsenstein, J. (1974). The evolutionary advantage of recombination. *Genetics* **78**, 737-756.

Foster, J. W. and Graves, J. A. (1994). An SRY-related sequence on the marsupial X chromosome: implications for the evolution of the mammalian testis determining gene. *Proceedings of the National Academy of Sciences,USA* **91**, 1927-1931.

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 Marshall Graves, J. A. (1992). Evolution of sex determination and the Y chromosome: SRY-related sequences in marsupials. *Nature* **359**, 531-533.

Foster, J. W., Dominguez-Steglich, M. A., Guioli, S., Kwok, C., Weller, P. A., Stevanovic, M., 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.

Frigerio,G., Burri, M., Bopp, D., Baumgartner, S. and Noll, M. (1986). Structure of the Segmentation gene Paired and the Drosophila PRD gene as part of a gene network. *Cell* **47**, 735-746.

Giese, K., Cox, J. and Grosschedl, R. (1992). The HMG domain of lymphoid enhancer factor 1 bends DNA and facilitates assembly of functional nucleoprotein structures. *Cell* **69**, 1-20.

Graves, J. A. and Schmidt, M. M. (1992). Mammalian sex chromosomes: design or accident? *Current Opinion in Genetics and Development* **2**, 890-901.

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.* **B 244**, 123-128.

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.

Hacker, A. (1995). Gene expression during testis determination in the mouse. PhD thesis (University of London).

Hacker, A., Capel, B., Goodfellow, P. and Lovell-Badge, R. (1995). Expression of *Sry*, the mouse sex determining gene. *Development*. **121**,1603-1614

Han, K., Levine, M. S. and Manley, J. L. (1989). Synergistic activation and repression of transcription by Drosophila homeobox proteins. *Cell* **56**, 573-583.

Harley, V. R., Jackson, D. I., Hextal, P. J., Hawkins, J. R., Berkovitz, G. D., Sockanathan, S., Lovell-Badge, R. and Goodfellow, P. N. (1992). DNA-binding activity of recombinant SRY from normal males and XY females. *Science* **255**, 453-456.

Harley, V. R., Lovell-Badge, R. & Goodfellow, P. N. (1994). Definition of a consensus DNA-binding site for SRY. *Nucleic Acids Res.* **22**, 1500-1501.

Hodgkin, J. (1992). Genetic sex determination, mechanisms and evolution. *Bioessays* **14**, 253-261.

Jaynes, J. B. and O'Farrell, P. H. (1988). Activation and repression of transcription by homeodomain-containing proteins that bind a common site. *Nature* **336**, 744-749.

Kamachi, Y., Sockanathan, S., Liu, Q., Breitman., M., Lovell-Badge, R. and Kondoh, H. (1995). Involvement of SOX proteins in lens specific activation of crystallin genes. *EMBO J.***14,** 3510-3519

Kay, G., Thakker, R. V. and Rastan, S. (1991). Determination of a molecular map position for Hyp using a new interspecific backcross produced by in vitro fertilization. *Genomics* **11**, 651-657.

Koopman, P., Münsterberg, A., Capel, B., Vivian, N. and Lovell-Badge, R. (1990). Expression of a candidate sex-determining gene during mouse testis differentiation. *Nature* **348**, 450-452.

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.

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

Laval, S. H., Blair, H. J., Hirst, M. C., Davies, K. E. and Boyd, Y. (1992). Mapping of FMR1, the gene implicated in fragile X-linked mental retardation, on the mouse X chromosome. *Genomics* **12**, 818-821.

Liechti-Gallati, S., Muller, B., Grimm, T., Kress, W., Muller, C., Boltshauser, E., Moser, H. and Braga, S. (1991). X-linked centronuclear myopathy: mapping the gene to Xq28. *Neuromuscul. Disord.* **1**, 239-245.

Lovell-Badge, R. H. (1987). Introduction of DNA into embryonic stem cells. In *Teratocarcinomas and Embryonic Stem Cells, a Practical Approach,; ed. E.J.* (ed. Robertson), pp. 153-182. Oxford: IRL Press.

Millen, K. J., Wurst, W., Herrup, K. and Joyner, A. L. (1994). Abnormal embryonic cerebellar development and patterning of postnatal foliation in two mouse engrailed-2 mutants. *Development* **120**, 695-706.

Muller, H. J. (1964). The relation of recombination to mutational advance. *Mutation Research* **1**, 2-9.

Ohno, S., ed. (1967). *Sex Chromosomes and Sex-linked Genes*. New-York: Springer-Verlag.

Parkhurst, S. M., Bopp, D. and Ish-Horowicz, D. (1990). X:A ratio, the primary sex-determining signal in drosophila, is transduced by helix-loop-helix proteins. *Cell* **63**, 1179-1191.

Pontiggia, A., Rimini, R., Harley, V.R., Goodfellow, P.N., Lovell-Badge, R. and Bianchi, M.E. (1994). Sex-reversing mutations affect the architecture of SRY-DNA complexes. *EMBO Journal* **13**, 6115-6124.

Rice, W. (1992). Sexually antagonistic genes: Experimental evidence. *Science* **256**, 1436-1439.

Rice, W. R. (1994). Degeneration of a non-recombining chromosome. *Science* **263**, 230-232.

Sambrook, J., Fritsch, E. F. and Maniatis, T., eds. (1989). *Molecular cloning: a laboratory manual*, 2nd edn. New York: Cold Spring Harbor Laboratory Press.

Sanger, F. S., Nicklen, S. and Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proc. natn. Acad. Sci. U.S.A.* **74**, 5463-5467.

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. and Goodfellow, P. N. (1990). A gene from the human sex-determining region encodes a protein with homology to a conserved DNA-binding motif. *Nature* **346**, 240-244.

Smith, D. B. and Johnson, K. S. (1988). Single-step purification of polypeptides expressed in *Escherichia coli* as fusions with glutathione-S-transferase. *Gene* **67**, 31-40.

Stevanovic, M., Lovell-Badge, R., Collignon, J. and Goodfellow, P. N. (1993). SOX3 is an X-linked gene related to SRY. *Human Molecular Genetics* **2**, 2007-2012.

Stevanovic, M., Zuffardi, O., Collignon, J., Lovell-Badge, R., and Goodfellow, P. N. (1994). The cDNA sequence and chromosomal location of the human *SOX2* gene. *Mammalian Genome* **5**, 640-642.

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.

Turner, G., Gedeon, A., Mulley, J., Sutherland, G., Rae, J., Power, K. and Arthur, I. (1989). Borjeson-Forssman-Lehmann syndrome: clinical manifestations and gene localisation to Xq26-27. *Am. J. Med. Genet.* **34**, 463-469.

Uwanogho, D., Rex, M., Cartwright, E.J., Pearl, G., Healy, C., Scotting, P.J. & Sharpe, P.T. (1995). Embryonic expression of the chicken Sox2, Sox3 and Sox11 genes suggests an interactive role in neuronal development. *Mechanisms of Development* **49**, 23-36.

van de Wettering, 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.

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.

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

Wilkinson, D. G. and Green, J. (1990). In situ hybridization and the three-dimensional reconstruction of serial sections. In *Postimplantation Mammalian Embryos. A Practical Approach* (ed. A. J. Copp and D. L. Cockroft), pp. 155-171. Oxford: Oxford University Press.

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.

Wright, E., Hargreave, 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 Genetics* **9**, 15-20.

Wurst, W, Auerbach, A. B. and Joyner, A. L. (1994). Multiple developmental defects in Engrailed-1 mutant mice: an early mid-hindbrain deletion and patterning defects in forelimbs and sternum. *Development* **120**, 2065-2075.

(Accepted 22 November 1995)