Nucleic Acids Research, 1995, Vol. 23, No. 17 3365–3372

The *Sry*-related HMG box-containing gene *Sox6* is expressed in the adult testis and developing nervous system of the mouse

Frances Connor, Edwina Wright[1], Paul Denny[+], Peter Koopman[1,2] and Alan Ashworth*

CRC Centre for Cell and Molecular Biology, Chester Beatty Laboratories, The Institute of Cancer Research, Fulham Road, London SW3 6JB, UK and [1]Centre for Molecular and Cellular Biology and [2]Department of Anatomical Sciences, The University of Queensland, Brisbane, Queensland 4072, Australia

Received July 3, 1995; Revised and Accepted August 2, 1995

GenBank accession no. U32614

ABSTRACT

We have cloned and sequenced a full-length cDNA for the HMG box-containing, *SRY*-related gene *Sox6* from mouse. The deduced protein sequence of Sox6 has considerable homology with that of the previously determined Sox5 sequence. It seems likely that these genes have diverged more recently than other members of the *SOX* gene family, although the two genes map to different chromosomes in the mouse. In common with *Sox5*, *Sox6* is highly expressed in the adult mouse testis and the HMG domains of both proteins bind to the sequence 5′-AACAAT-3′. This suggests that the two genes may have overlapping functions in the regulation of gene expression during spermatogenesis in the adult mouse. However, *Sox6* may have an additional role in the mouse embryo, where it is specifically expressed in the developing nervous system.

INTRODUCTION

DNA–protein interactions are involved in many of the fundamental processes that occur inside cells, including replication of the genome, repair of damaged DNA and transcription of active genes. Several families of DNA binding proteins have been implicated in the developmental control of gene expression (1). The Y chromosome gene *SRY* is the developmental switch gene responsible for inducing testis differentiation in mammals (2,3). A distinctive feature of the SRY protein is the presence of an HMG domain, a structural protein motif shared by a growing family of diverse eukaryotic DNA binding proteins (4). The SRY HMG domain has been shown to bind to specific DNA sequences (5,6) and mutations in this domain that abolish its ability to bind DNA have been found associated with cases of sex-reversal in humans (7). The importance of this domain is further confirmed by comparisons of the SRY sequences from several primates and marsupials: the HMG domain appears to be the only conserved region of the protein (8,9). SRY is thought to function by controlling the transcription of genes, as yet unidentified, which act downstream in the sex determining pathway. However, the apparent absence of conserved sequences additional to the HMG domain suggests that SRY is not a 'classical' transcription factor with a separable transactivation domain. Interaction of SRY with specific DNA sequences induces a significant bend in the bound DNA (10,11); moreover, one case of sex reversal associated with an SRY mutation appears to derive from the anomalous DNA bending activity of mutant protein (12). These observations suggest that SRY belongs to a growing class of transcription factors that perform architectural roles in the assembly of stereospecific nucleoprotein complexes which are essential for correct gene expression (13).

SRY is the founder member of a family of genes related by sequence homology within the HMG domain; *SOX* ('*SRY* box') genes encode proteins with >60% similarity to the SRY HMG domain. At least 40 different *SOX* loci have been reported in mammals, birds, reptiles, amphibians and insects (14 and references therein). However, only a few of the *SOX* genes have been characterized in any detail. These studies suggest that *SOX* genes, like *SRY*, have regulatory roles in developmental pathways; for example, in the developing nervous system (*Sox-1*, *-2*, *-3*; 15); in T-cell differentiation (*Sox4*; 16) and in bone formation and gonadogenesis (*SOX9*; 17–19).

Understanding the roles of the *SRY*-related genes may provide further insights into how *SRY* functions and into the molecular mechanisms underlying the initiation and maintenance of developmental decisions in general. We have previously described the *Sry*-related gene *Sox5* (20,21). *Sox5* is expressed exclusively in post-meiotic germ cells of adult mouse testis, predominantly in the nucleus. Consistent with this location, Sox5 is a DNA binding protein with a moderately high affinity for the sequence 5′-AACAAT-3′. The function of *Sox5 in vivo* is still unclear, but these results are consistent with a role in regulating gene expression during the post-meiotic phase of spermatogenesis. Sox5 is apparently unable to activate transcription from a minimal promoter linked to single or multimerized Sox5 binding sites but, in common with SRY, Sox5 induces a large bend in the

* To whom correspondence should be addressed

[+]Present address: Nuffield Department of Surgery, John Radcliffe Hospital, Headington, Oxford OX3 9DU, UK

double helix upon DNA binding, due primarily to the presence of the HMG domain (20,21). These observations imply that Sox5 may function as an architectural transcription factor.

In addition to *Sox5*, a second distinct *Sry*-related gene, *Sox6*, was isolated from adult mouse testis (22). The original *Sox6* clone was restricted to the HMG box and over this region *Sox5* and *Sox6* are closely related (91% amino acid identity). Here we report the isolation of the full length cDNA sequence of *Sox6* and show that the homology with *Sox5* extends beyond the HMG domain. We show that both genes are highly expressed in the adult mouse testis and bind similar DNA sequences. However the *Sox6* gene is also expressed in the developing nervous system, suggesting an additional function for this gene in the mouse.

MATERIALS AND METHODS

Screening cDNA libraries and DNA sequencing

A *Sox6* cDNA clone, pD602, was recovered from a screen of an adult mouse testis cDNA library with a mixed probe consisting of *Sox5* and *Sox6* HMG box fragments (22). This partial *Sox6* cDNA fragment was used to screen a second amplified adult mouse testis cDNA library in λZAP (Stratagene) (a gift of K. Willison). One full length *Sox6* cDNA clone was isolated and the insert sequenced using either the Sequenase DNA Sequencing kit (Version 2.0; USB Corporation) or the *Taq* DyeDeoxy™ Terminator Cycle Sequencing kit and the 373A DNA Sequencer (Applied Biosystems). A combination of specific oligonucleotide primers and various deletion templates generated by subcloning specific restriction fragments was used to obtain the sequence of both strands.

Genetic mapping

The European Collaborative Interspecific Backcross (EUCIB; 23) was used to map the *Sox6* gene in the mouse genome. The segregation of the C57BL/6 and *Mus spretus* (EUCIB parental species) *Sox6* alleles was followed by single stranded conformation polymorphism (SSCP) gel analysis of PCR products. Genomic DNAs from the backcross parental species and from 46 randomly chosen samples from the backcross were amplified by PCR using the *Sox6* specific primers 800 (5′-AATCTGTCTCACACGTTA-CG-3′) and 801 (5′-CCTGTCAGTGCCAACTGAGG-3′). Oligonucleotide 801 was 5′ end labelled using T4 polynucleotide kinase and [γ-^{32}P]ATP. Each PCR reaction contained 50 ng of genomic DNA, 2 μl of 10 × PCR buffer, 0.4 μl of primer 800 (10 pmol/μl), 0.325 μl of primer 801 (10 pmol/μl), 0.5 μl of 5′ end labelled primer 801 (~1.5 pmol/μl), 0.5 μl of dNTPs (10 mM), 0.2 μl of *Taq* DNA polymerase (Cetus, 5 U/μl) and water to a total volume of 20 μl. The PCR samples were subjected to 30 cycles of 94°C for 0.5 min, 50°C for 0.5 min and 72°C for 1 min, followed by a final incubation at 72°C for 5 min. 2.5 μl of each PCR was mixed with 7.5 μl of formamide dye, heated at 100°C for 5 min and placed on ice for 5 min. This was then loaded on a 0.4 mm non-denaturing acrylamide based gel [0.25 × MDE (AT Biochem) in 0.6 × TBE] run in 0.6 × TBE at 8 W for 12 h. The wet gel was exposed directly to Kodak XAR-5 film. The segregation pattern was analysed using the Mbx database (23) at the HGMP Resource Centre, UK.

Northern hybridization

Poly(A)$^+$ RNAs were prepared from adult male MF1 mouse tissues using the FastTrack mRNA Isolation kit (Invitrogen). RNA samples of 1 μg were separated on a 1% agarose/1 × MOPS gel containing 6% formaldehyde. RNA markers (0.24–9.5 kb RNA ladder; BRL) were loaded alongside. RNA was then capillary blotted and UV-immobilized onto Hybond N$^+$ membrane (Amersham), according to manufacturer's instructions. A mouse *Sox6* cDNA probe was prepared from p8304 by ^{32}P-labelling the 1.5 kb *Acc*I restriction fragment (nucleotides 460–1992), which comprises sequence at the 5′ end of the cDNA but excluding the HMG box, using the Prime It II Random Primer Labelling kit (Stratagene). The Northern blot was hybridized to this probe using Rapid-Hyb buffer (Amersham), according to manufacturer's instructions.

In situ hybridization

Antisense and sense RNA probes were prepared from subclones of *Sox6* cDNA 3′ to the HMG box but not containing any HMG box, polyadenine or *Sox5* homologous sequences (nucleotides 2319–2906, GenBank/EMBL accession no. U32614). Hybridizations were carried out essentially as described by Wilkinson and Nieto (24) with minor modifications (25).

Expression of *Sox6* in *E.coli*

Nucleotides 2007–2318 of the mouse *Sox6* cDNA, encoding amino acid residues 612–715 including the HMG domain, were excised as a *Sac*II/*Nru*I restriction fragment, blunt ended and subcloned into the *Sma*I site of the expression vector pGEX-3T (26). GST-fusion protein was expressed in BL21(DE3)pLysS host bacteria (27) and purified using glutathione-sepharose beads (26). The protein was >90% pure, as determined by Coomassie blue staining of SDS–polyacrylamide gels.

Electrophoretic mobility shift assays

Electrophoretic mobility shift assays (EMSA) were performed essentially as described previously (20). Briefly, 1 ng of purified GST-fusion protein was pre-incubated in a final 15 μl volume of EMSA binding buffer (10 mM HEPES pH 7.9, 60 mM KCl, 1 mM DDT, 1 mM EDTA, 0.3 mg/ml BSA, 12% glycerol) with appropriate competitor DNA for 5 min at room temperature. Then 10 fmol (~ 35 000 c.p.m.) of ^{32}P-labelled, annealed oligonucleotides were added and incubated for a further 20 min at room temperature. The reactions were electrophoresed on non-denaturing 4% polyacrylamide gels in 0.5 × TBE buffer at room temperature at 10 V/cm for 90 min. Gels were dried prior to autoradiography.

The sequences of the probes and competitors ('top' strand of double-stranded oligonucleotides) are: BS12 5′-TCGAGCACTAAAACAATTCAAGCCCGGGG-3′ (20); MW56 5′-GGGAGACTGAGAACAAAGCGCTCTCACAC-3′ (28); NF1 5′-CGCATTTGGCAGCTTGCCAAGGATCCTTGGCAGCTTGCCAAG-3′ (gift of R. Nicolas). The BS12 double stranded oligonucleotide probe was labelled using the Klenow fragment of DNA polymerase I and [^{32}P]dCTP.

RESULTS

Isolation of cDNAs encoding *Sox6*, a novel *Sry*-related gene

An RT–PCR fragment derived from the HMG box of mouse *Sox6* (22) was used to screen an adult mouse testis cDNA library and several clones were isolated. The longest cDNA was sequenced on both strands and the sequence has been deposited in the GenBank/EMBL database (accession no. U32614). The *Sox6* cDNA contains an open reading frame that encodes an 827 amino acid protein. An in frame termination codon is present nine nucleotides upstream of the putative initiating methionine (data not shown) providing strong evidence that we have isolated a full-length cDNA clone for *Sox6*. The deduced amino acid sequence of Sox6 is shown in Figure 1A. The predicted molecular weight of this protein is 91.8 kDa, an estimate that is supported by the production of an *in vitro* translated protein from the *Sox6* cDNA of ~105 kDa (data not shown). The end of the coding sequence is followed by 251 bp of 3′ untranslated sequence and then a stretch of adenosine residues. A canonical polyadenylation signal (AATAAA) was not identified, however, an alternative (ATTAAA) is present 20 bp upstream of the polyadenosine tract (data not shown).

Sox6 contains a single HMG box and no other obvious DNA binding motifs. There is one nucleotide difference within the HMG box between this cDNA and the original PCR clone of *Sox6* (22) which gives rise to a conservative difference at amino acid residue 632, from a lysine to an arginine. This difference is probably the result of a polymorphism between mouse strains. Comparison of the predicted amino acid sequence of the *Sox6* HMG domain with other HMG domains clearly shows that *Sox6* is a member of the *Sry*-related family (Fig. 1B). On the basis of HMG domain sequence, the mouse *Sox* genes can be classified into subgroups (29); *Sox6* is part of the distinct subgroup D containing *Sox5* and *Sox13*. Another notable feature of the Sox6 protein is a region (residues 184–205) to the N-terminal side of the HMG domain in which a leucine (indicated by asterisks in Fig. 1A) is repeated every seven residues. Such 'heptad repeats' are found in the leucine zipper dimerization motif of the bZIP family of transcription factors (30). Further features of the predicted Sox6 protein include two regions (residues 190–261 and 493–517), also on the N-terminal side of the HMG domain, that are rich in glutamine residues (33 and 32% respectively). The sequence between these glutamine-rich regions contains two short polyalanine stretches of 6 and 5 residues.

The complete coding sequences for a few mouse *Sox* genes are known. Outside the HMG domain, Sox6 does not show any significant similarity at amino acid level to Sry, Sox-1, -2, -3, -4, or -9 (16,19,31; R. Lovell-Badge, pers. comm.). However, a comparison of the translated full-length cDNA sequences of Sox5 and Sox6 reveals a similarity extending beyond the HMG domain (Fig. 1A). This strongly suggests that *Sox5* and *Sox6* have diverged more recently than the other known members of the *Sox* gene family. Database searches using *Sox6* excluding the HMG box detect no other closely related sequences, at either the nucleotide (EMBL and GenBank databases) or protein level (Swiss-Prot database).

Sox6 and *Sox5* map to different mouse chromosomes

The high degree of homology between *Sox5* and *Sox6* strongly suggests that they have descended from a single ancestral gene through a duplication event. It was, therefore, of interest to compare the chromosomal location of *Sox6* with that of *Sox5*. The *Sox6* gene was mapped by interspecific mouse backcross pedigree analysis using the EUCIB (23). PCR–SSCP analysis defined a variant in the *Sox6* gene between the backcross parental species C57BL/6 and *Mus spretus*. This variant was used to follow the segregation of C57BL/6 and *Mus spretus Sox6* alleles in 46 randomly chosen EUCIB backcross mice. Analysis of the haplotypes placed the *Sox6* gene on mouse chromosome 7, linked to the anchor loci *D7Mit40* (one recombinant; 2.17 ± 2.15 cM) and *D7Mit15* (11 recombinants; 23.91 ± 6.29 cM; Fig. 2). The *Sox5* gene has been mapped previously to mouse chromosome 6 (P.D. and A.A., manuscript in preparation). Hybridization of a Southern blot of mouse genomic DNA with a probe corresponding to the 3′ untranslated region of mouse genomic *Sox6* gave a single band (not shown), suggesting that there is a single mouse gene closely related to the *Sox6* cDNA clone. *Sox6* does not map close to any mutations mapped to chromosome 7 for which it is an obvious candidate.

Sox6 mRNA is highly expressed in adult mouse testis

Expression of the *Sox6* gene in the adult mouse was analysed by hybridization of a *Sox6* probe to a Northern blot of poly(A)$^+$ RNAs isolated from several mouse tissues. This demonstrated that among the adult mouse tissues analysed *Sox6* expression, like *Sox5*, appears to be restricted to the testis. A single transcript of ~3.2 kb was detected in adult testis (Fig. 3). Prolonged exposure of the autoradiogram failed to reveal hybridizing species of the same size as in testis, although in some tissues some faint hybridization to high molecular weight transcripts was seen (data not shown). Integrity of the mRNA was checked by re-hybridization of the blot with a mouse GAPDH probe (Fig. 3).

Sox6 mRNA is expressed in the developing nervous system of the mouse

Wholemount *in situ* hybridization was used to study the expression of *Sox6* during mouse embryo development. A profile of *Sox6* expression was observed that was specific to the development of the nervous system (Fig. 4). At 9.5 days post coitum (dpc), expression was seen in the central nervous system (CNS), with highest levels in the forebrain gradually decreasing in a rostrocaudal gradient to a region posterior to the forelimb bud, beyond which expression ceased. This pattern of expression was maintained at 10.5 dpc, but the caudal boundary of expression had regressed anteriorly to a point in the upper cervical region. At both stages expression was confined to the ventricular cells of the brain and spinal cord, and extended into the optic evaginations. No dorsoventral gradient of expression was visible. No expression was seen in the surface ectoderm or mesoderm overlying these CNS components. In addition to expression in the CNS, *Sox6* expression in the neural component of the developing ears was observed at 9.5 and 10.5 dpc. No sites of expression were visible elsewhere in the embryo. By 12.5 dpc, expression was extinguished in the brain and only weak expression remained in the spinal cord (data not shown).

Sequence specific DNA binding by the Sox6 HMG domain

Where tested, the HMG domain of Sox proteins has been shown to confer the ability to bind DNA sequence specifically. *In vitro*

A

```
  1  MSSKQATSPFACTADGEEAMTQDLTSREKEEGSDQHPASHLPLHPIMHNKPHSEELPTLV
 61  STIQQDADWDSVLSSQQRMESENNKLCSLYSFRNTSTSPHKPDEGSREREIMNSVTFGTP
121  ERRKGSLADVVDTLKQKKLEEMTRTEQEDSSCMEKLLSKDWKEKMERLNTSELLGEIKGT
181  PESLAEKERQLSTMITQLISLREQLLAAHDEQKKLAASQIEKQRQQMDLARQQQEQIARQ
241  QQQLLQQQHKINLLQQQIQVQGHMPPLMIPIFPHDQRTLAAAAAAQQGFLFPPGITYKPG
301  DNYPVQFIPSTMAAAAA--SGLSPLQLQKGHVSHPQINPRLKGISDRFGRNLDPSEHGGG
                       MAAAAAATPGLGPLQLQ----------------------------------  17
359  HSYNHRQIEQLYAAQLASMQVSPGAKMPSTPQPPNSAGAVSPTGIKNEKRGTSPVTQVKD
     ------------------------------------------------------------
419  ETTAQPLNLSSRPKTAEPVKSPTSPTQNLFPASKTSPVNLPNKSSIPSPIGGSLGRGSSL
     DEVAQPLNLSAKPKTSDG-KSPASPTSPHMPALRINSGAGPLKASVPAALASPSARVSTI  76
479  DILSSLNSPALFGDQDTVMKAIQEARKMREQIQREQQQQPHGVDGKLSSMNNMGLSNCRT
     GYL---------NDHDAVTKAIQEARQMKEQLRREQQ----ALDGKVAVVNSIGLSNCRT  123
539  EKERTRFENLGPQLTGKSSEDGKLGPGVIDLTRPEDAEGSKAMNGSAAKLQQYYCWPTGG
     EKEKTTLESLTQQLAVKQNEEGKFSHGMMDFNMSGDSDGS--------------------  163
599  ATVAEARVYRDARGRASSEPHIKRPMNAFMVWAKDERRKILQAFPDMHNSNISKILGSRW
     AGVSESRIYRESRGRGSNEPHIKRPMNAFMVWAKDERRKILQAFPDMHNSNISKILGSRW  223
659  KSMSNQEKQPYYEEQARLSKIHLEKYPNYKYKPRPKRTCIVDGKKLRIGEYKQLMRSRRQ
     KAMTNLEKQPYYEEQARLSKQHLEKYPDYKYKPRPKRTCLVDGKKLRIGEYKAIMRNRRQ  283
719  EMRQFFTVGQQPQMPITTGTGVVYPGAITMATTTPSPQMTSDCSSTSASPEPSLPVIQST
     EMRQYFNVGQQAQIPIATA-GVVYPGAIAMA-GMPSPHLPSEHSSVSSSPEPGMPVIQST  341
779  YGMKMDGASLAGN---DMINGEDEMEAYDDYEDDPKSDYSSENEAPEPVSAN
     YGAKGEEPHIKEEIQAEDINGE-IYEEYDEEEEDPDVDYGSDSENHIAGQAN          392
```

B

			% IDENTITY
Sox6		EPHIKRPMNA FMVWAKDERR KILQAFPDMH NSNISKILGS RWKSMSNQEK QPYYEEQARL SKIHLEKYPN YKYKPRPKRT	100
Sox5	(D)	EPHIKRPMNA FMVWAKDERR KILQAFPDMH NSNISKILGS RWKAMTNLEK QPYYEEQARL SKQHLEKYPD YKYKPRPKRT	93
Sox13	(D)	MVWAKDERR KILQAFPDMH NSSISKILGS RWKSMTNQEK QPYYEEQARL SRQHLEK	93
mSry	(A)	EGHVKRPMNA FMVWSRGERH KLAQQNPSMQ NTEISKQLGC RWKSLTEAEK RPFFQEAQRL KILHREKYPN YKYQPHRRAK	58
Sox4	(C)	SGHIKRPMNA FMVWSQIERR KIMEQSPDMH NAEISKRLGK RWKLLKDSDK IPFIQEAERL RLKHMADYPD YKYRPRKKVK	55
Sox1	(B)	QDRVKRPMNA FMVWSRGQRR KMAQENPKMH NSEISKRLGA EWKVMSEAEK RPFIDEAKRL RALHMKEHPD YKYRPRRKTK	52
hSRY		QDRVKRPMNA FIVWSRDQRR KMALENPRMR NSEISKQLGY QWKMLTEAEK WPFFQEAQKL QAMHREKYPN YKYRPRRKAK	51
Sox9	(E)	KPHVKRPMNA FMVWAQAARR KLADQYPHLH NAELSKTLGK LWRLLNESEK RPFVEEAERL RVQHKKDHPD YKYQPRRRKS	48
Sox7	(F)	AKDERK RLAVQNPDLH NAELSKMLGK SWKALTLSQK RPYVDEAERL RLQHMQDY	43
LEF1		RPHIKKPLNA FMLYMKEMRA NVVAECTLKE SAAINQILGR RWHALSREEQ AKYYELARKE RQLHMQLYPG WSARDNYGKK	35
Mat-Mc		TERTPRPPNA FILYRKEKHA TLLKSNPSIN NSQVSKLVGE MWRNESKEVR MRYFKMSEFY KAQHQKMYPG YKYQPRKNKV	28
HMG1		PNAPKRPPSA FFLFCSEYRP KIKGEHPGLS IGDVAKKLGE MWNNTAADDK QPYEKKAAKL KEKYEKDIAA YRAKGKPDAA	28
UBF.1		PDFPKKPLTP YFRFFMEKRA KYAKLHPEMS NLDLTKILSK KYKELPEKKK MKYIQDFQRE KQEFERNLAR FREDHPDLIQ	20

```
                  *******  *******        *************          ************************
                  Helix 1  Helix 1'         Helix 2                   Helix 3          Basic
```

Figure 1. (**A**) Amino acid sequence of mouse Sox6. The amino acid sequence of Sox6 (upper sequence), predicted from the cDNA sequence of the single *Sox6* clone p8304, is shown compared with that of Sox5 (lower sequence) (20). Numbers on the right refer to Sox6, those on the left to Sox5. The optimal alignment of the Sox5 and Sox6 sequences was found by inserting the minimum number of gaps to maximize the number of matches, according to the algorithm described by Needleman and Wunsch (47). The alignment was generated by the GAP program (GCG package, Daresbury Laboratory, Daresbury, Warrington, UK). Vertical lines and double dots denote amino acid identities and conservative changes, respectively. Asterisks mark the leucine residues potentially forming part of a 'heptad repeat'. (**B**) Comparison of Sox6 HMG domain with other HMG domain proteins. Amino acid sequences of various HMG domain proteins are aligned according to the algorithm described by Read and coworkers (48). The sequences represent the 71 residue 'minimum' HMG domain followed by the adjacent, often basic, C-terminal sequences. Representative members of the six subgroups (A–F) of the mouse *Sox* gene family are shown. Examples of the more distantly related HMG domains, mouse LEF1, *Schizosaccharomyces pombe* Mat-Mc, domain 1 of human UBF and B-domain of hamster HMG1, are towards the bottom of the figure. Sequences are taken from refs 4 and 14. The percentages of amino acid residues that are identical between the 'minimum' HMG domains of these proteins and Sox6 have been calculated (% identity column). The position of α-helices shown at the bottom is derived from NMR analysis of the B-domain of HMG1 (49).

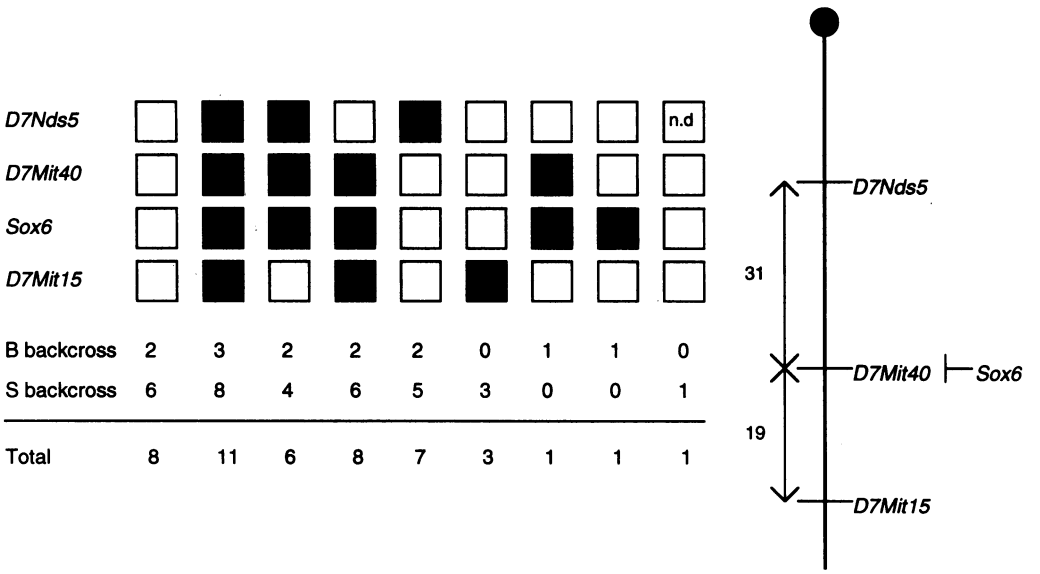

Figure 2. The *Sox6* gene maps to mouse chromosome 7. The types and numbers of recombinants observed in 46 EUCIB mice typed for all four markers are tabulated. The data for the backcrosses to *M.m.domesticus* (C57BL/6) and to *M.spretus* are shown separately. The filled boxes represent the presence of the C57BL/6 allele, the open boxes the *M.spretus* allele (n.d, not determined). Primary data are stored on the Mbx database (23) at the HGMP Resource Centre, UK. A partial mouse chromosome 7 linkage map showing the location of *Sox6* in relation to the anchor markers *D7Nds5*, *D7Mit40* and *D7Mit15* is shown to the right of the table. Recombination distances between loci are expressed in cM and are taken from ref. 23 and the Mbx database.

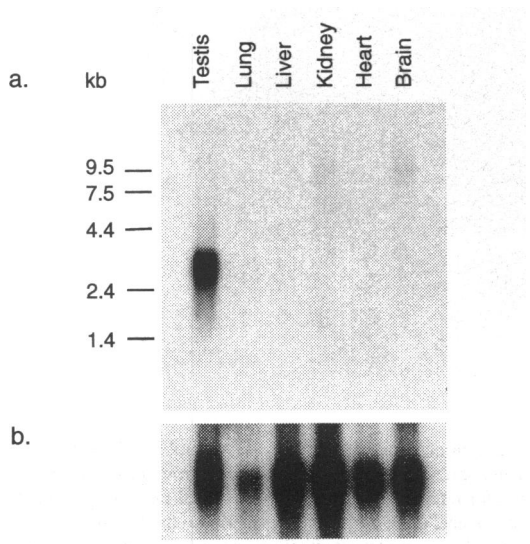

Figure 3. *Sox6* is expressed specifically in the testis of the adult mouse. (**a**) One microgram of poly(A)$^+$ RNA extracted from various adult male mouse tissues was electrophoresed, transferred to Hybond N$^+$ (Amersham) and hybridized with a mouse *Sox6* cDNA probe (the 1.5 kb *Acc*I fragment of p8304 comprising coding sequence 5′ to and excluding the HMG box). Exposure was for 16 h. (**b**) The same membrane was subsequently rehybridized with a mouse GAPDH probe. Exposure was for 30 min.

site selection assays have been used to define the consensus motif 5′ AACAAT 3′ as a preferred binding site for Sox5 (20) and SRY (6). Sox-1, -2 and -3 will also bind to this site in EMSAs (R. Lovell-Badge, pers. comm.). Sox4 interacts with high affinity to a similar element, 5′-AACAAAG-3′ (16).

As a preliminary study of the interaction of Sox6 with DNA, Sox6 protein was analysed for its ability to bind to the 5′-AACAAT-3′ motif. The isolated Sox6 HMG domain was expressed in *E.coli* as a fusion protein with glutathione-*S*-transferase. The fusion protein was used in EMSAs with the radiolabelled double-stranded oligonucleotide BS12. The Sox6 HMG domain indeed bound to this 5′-AACAAT-3′-containing probe (Fig. 5). The specificity of the interaction was tested by comparing the effect of adding various unlabelled oligonucleotides to the binding reaction. Complexes formed with the BS12 probe were competed efficiently by BS12 but poorly by an irrelevant control oligonucleotide, NF1. The MW56 oligonucleotide, which contains the 5′-AACAAAG-3′ sequence, showed competition characteristics similar to those obtained using the 5′-AACAAT-3′-containing oligonucleotide BS12. These results mirror those seen using recombinant Sox5 protein (20), suggesting that the two HMG domains have similar binding specificities.

DISCUSSION

Since the discovery of the mammalian testis determining gene *SRY*, an expanding family of genes that encode proteins with related HMG domains have been identified. We have characterized a novel member of this gene family, *Sox6*. On the basis of the sequence of its HMG box, *Sox6* can be classified, together with *Sox5* and *Sox13*, as a member of the D subgroup of mouse *Sry*-related genes (29). Full-length cDNA sequences are now known for both *Sox6* and *Sox5*. Comparison of the translated sequences reveals considerable homology between the two extending beyond the HMG domain (Fig. 1A). Similarly, alignment of sequences of the Sox proteins belonging to the B subfamily has revealed conserved domains outside the HMG domain (32). The extensive sequence homology between Sox5 and Sox6 strongly implies that they have diverged more recently

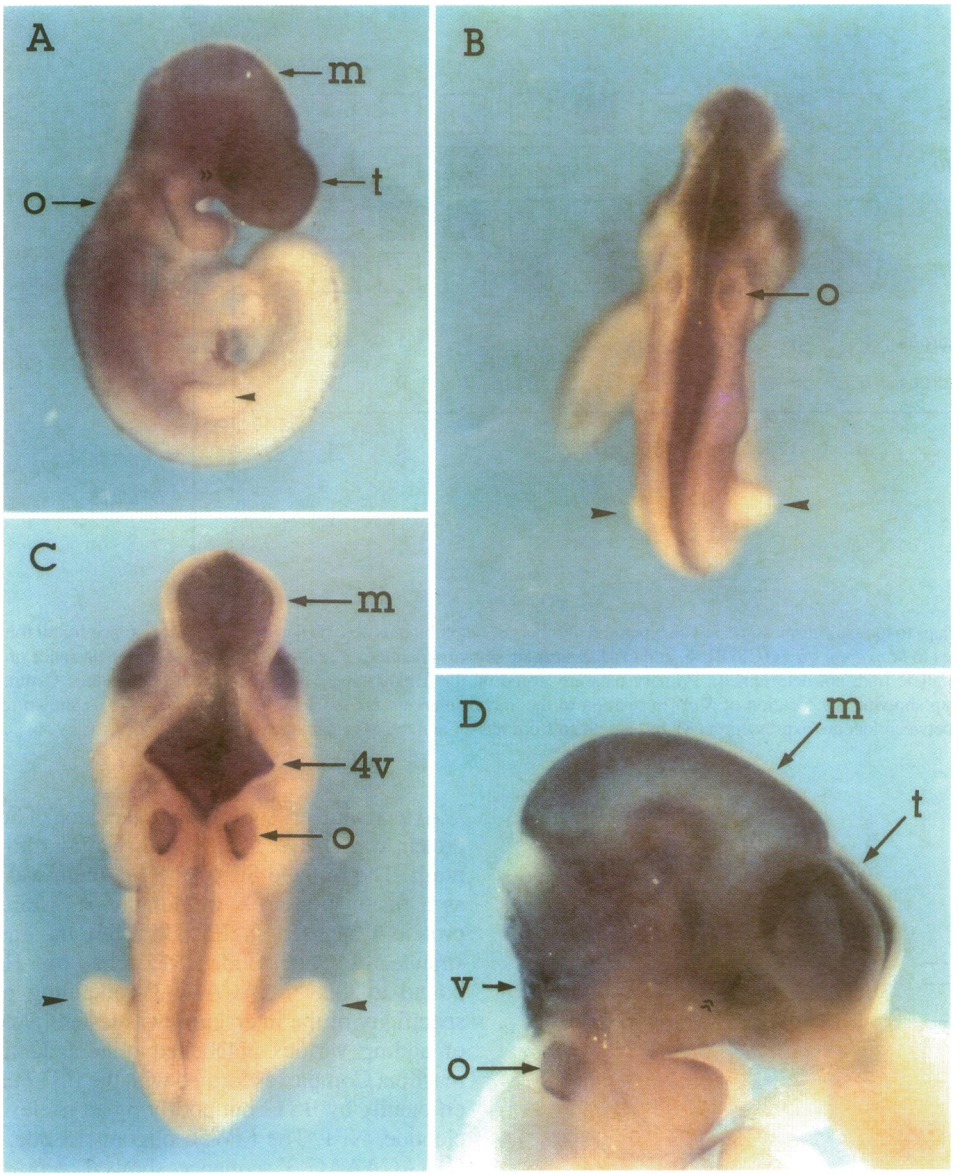

Figure 4. Analysis of *Sox6* expression in developing mouse embryos by wholemount *in situ* hybridizations. (**A**) 9.5 dpc whole embryo showing expression in the anterior CNS and the otic placode (o). Most intense staining is seen in the telencephalon (t) and optic evagination (>>) emanating from the diencephalon, decreasing caudally through the mesencephalon (m) and hindbrain regions. The forelimb bud is shown for reference (arrowhead). (**B**) Dorsal view of same embryo showing expression in neuroepithelial cells of the ventricular zone of the midbrain, hindbrain and upper spinal cord, diminishing posterior to the forelimb buds (arrowheads). o, otic placode. (**C**) Dorsal view of a 10.5 dpc showing expression in the midbrain, hindbrain, spinal cord and neuroepithelium of the developing ear. m, mesencephalon; 4v, fourth ventricle; o, otic vesicle; arrowheads, forelimb buds. (**D**) Head of same embryo illustrating the specificity of staining to the neuroepithelial lining of the telencephalic hemispheres (t); mesencephalon (m); fourth ventricle (v); and expression in the optic stalk (>>); and otic vesicle (o). Surface ectoderm and mesoderm overlying the brain are negative, as are all other tissues. The posterior limit of staining in A, B and C are indicated by arrows. Optical sections of the midbrain shown in B and C indicate dorsoventral uniformity of *Sox6* staining. Equivalent embryos hybridized with sense *Sox6* probe showed no staining.

than the other members of the *SOX* gene family. However, genetic mapping places the genes on different mouse chromosomes: *Sox6* is on chromosome 7, while *Sox5* is located on chromosome 6.

One major difference between *Sox5* and *Sox6* is the size of their predicted encoded proteins; the Sox6 protein is predicted to be approximately twice the size of Sox5. Endogenous Sox5 protein detected by immunoblot analysis of protein extracted from mouse testis cells using anti-Sox5 peptide antibody is 52 kDa, similar to the size estimated from its cDNA sequence (20). The *Sox6* start codon was chosen on the basis of the largest open reading frame and the use of the first methionine encoded by the *Sox6* cDNA sequence. This estimate is supported by the production of an *in vitro* translated protein from the *Sox6* cDNA of ~105 kDa (data not shown). Nevertheless, confirmation of the functional initiator codon awaits analysis of the Sox6 protein using specific antibody reagents.

The predicted Sox6 protein shows several hallmarks of a regulatory transcription factor. These include the DNA binding

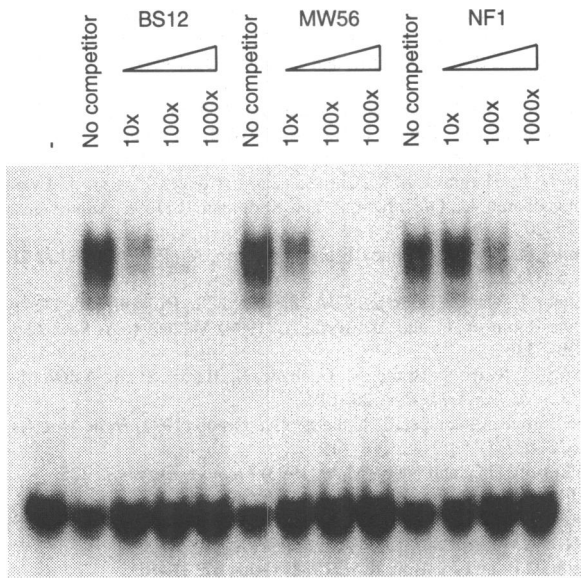

Figure 5. Electrophoretic mobility shift assays of oligonucleotide binding by GST-Sox6 HMG domain. ^{32}P-labelled BS12 oligonucleotide probe was incubated with 1 ng of purified GST-Sox6 HMG domain protein. The protein–DNA complexes were separated from unbound free DNA probe by electrophoresis through a native 4% polyacrylamide gel and visualised by autoradiography. For competition with protein binding, unlabelled oligonucleotides were included in some binding reactions at the molar excess over probe oligonucleotide indicated in the panel. A control reaction performed in the absence of GST-Sox6 HMG domain protein is indicated with a dash (–).

HMG domain. Immediately C-terminal to the 71 amino acid residue 'minimal' HMG domain is a short stretch of primarily basic residues. Such regions can be recognized in many of the TCF/MATA/SOX subfamily of HMG domain proteins. In particular, these basic residues have been shown to be required for high affinity binding of human LEF to its specific site (33). Sox6 also possesses regions that are similar in character to transcriptional activation domains defined in RNA polymerase II transcription factors. These include glutamine-rich, proline-rich and acidic stretches; such domains can act as transactivation domains (e.g. 34–36). Proline- and acidic-rich regions are also found in Sox5 in similar positions relative to the HMG domain. In common with Sox1 and Sox3, Sox6 contains short stretches of alanine residues and alanine-rich regions have been noted in homeodomain proteins that act as transcriptional repressors (e.g. 37).

A potentially interesting feature of the predicted Sox6 protein is the presence of a sequence motif associated with a dimerization domain. A 21 amino acid region to the N-terminal side of the HMG domain contains a leucine repeated every seven residues. Such 'heptad repeats' are reminiscent of the leucine zipper motif of the bZIP family of transcription factors (30). These proteins bind to DNA as homo- or hetero-dimers between compatible family members. The leucine zipper region forms the dimerization interface that serves to juxtapose the adjacent basic DNA binding regions of each protein. *Bona fide* leucine zipper dimerization domains form a parallel, two-stranded, α-helical coiled-coil structure (38). The leucine heptad repeat in Sox6 does not appear (data not shown) to be similarly composed of an α-helix that can form a coiled–coil, as shown by analysis of the Sox6 protein using an α-helical coiled-coil prediction program ('stripe', a gift of A. Knight), based on the algorithm described by Lupas and coworkers (39). Further work is required to determine whether the potential leucine zipper sequnce in Sox6 is functional.

Sox5 and *Sox6* are not only similar in terms of sequence but also appear to have other properties in common. Their HMG domains show identical DNA binding characteristics *in vitro*. In the EMSA, the HMG domain of Sox6 binds specifically to oligonucleotides containing the sequence 5′-AACAAT-3′. This sequence was defined as the preferred binding site of Sox5 in *in vitro* oligonucleotide binding assays (20). Similar to the behaviour of Sox5 in competitive EMSAs, the Sox6 HMG domain will also bind to oligonucleotides containing the related sequence 5′-AACAAAG-3′. This is not unexpected as, where tested, HMG domain proteins belonging to the TCF/MATA/SOX subfamily all appear to bind *in vitro* to similar DNA sequences containing a central CA/TG dinucleotide pair often flanked by A/T rich sequence. This is in spite of only a low level of sequence conservation between members (4).

High levels of transcripts from both the *Sox5* and *Sox6* genes have been detected in the adult mouse testis, as shown by Northern blot hybridization (20; Fig. 3). Immunoblot analysis has revealed that Sox5 protein is restricted to the post-meiotic germ cells, being found at highest levels in the round spermatids (20). The presence of Sox5 protein suggests that this expression is not simply promiscuous gene transcription which is seen for many genes in the testis. Further work is necessary to determine whether *Sox6* RNA is present in the same cell type as that of *Sox5* and to confirm the presence of Sox6 protein. However, if this is the case it is possible that either the proteins have degenerate functions or that they act synergistically or cooperatively.

Aside from any function in the adult testis, *Sox6* appears to have a role in the developing CNS, where a striking pattern of expression in anterior but not posterior structures was observed. Further, a stage-specific role in anterior CNS development is indicated by the extinction of *Sox6* expression by 12.5 dpc and the lack of *Sox6* expression in the adult brain. The expression profile, combined with the ability of Sox6 protein to bind to specific DNA motifs and its potential to act as a transcriptional activator/repressor, suggests that Sox6 may act as a regulator of pattern formation in the CNS. In this regard there are obvious contrasts with the expression of *Hox*, *Pax*, *Evx-1* and engrailed genes, which are expressed caudally from specific points in the developing CNS (40). *Hox* gene expression is initiated at the posterior end of the neural tube at 8.0–8.5 dpc, and extends forward so that the anterior limits of expression correspond with rhombomere boundaries in the hindbrain at 9.5 dpc (41,42). Different *Hox* genes have different anterior boundaries of expression, which has led to speculation that the specific combination of *Hox* genes expressed at any given point along the rostrocaudal axis is responsible for providing positional signalling that leads to segment identity in the hindbrain region of the CNS (43). This theory is supported by observations that induced gain- and loss-of function mutations involving mouse *Hox* genes lead to altered segment identity phenotypes in neural crest derivatives (44,45). Given the complementary patterns of expression of *Sox6* and *Hox* genes in the CNS, it is possible that *Sox6* regulates or is regulated by one or more *Hox* genes. Alternatively, *Sox6* may be independently regulated and add further complexity to the so-called *Hox* code in generating positional information in the CNS. It will be of interest to examine the expression of *Sox6*

expression at earlier stages than those examined in the present study. Interestingly, some *Hox* genes (46) are also expressed in the developing nervous system and later in the adult testis. Previous work (P.D. and R. Lovell-Badge, pers. comm.) suggested that, in contrast to *Sox6*, *Sox5* is expressed specifically in the adult testis. This suggests that embryonic CNS expression has been acquired by the *Sox6* gene or lost by the *Sox5* gene and it will be interesting to compare the structures of the promoters of the two genes.

ACKNOWLEDGEMENTS

We thank Emma Tarttelin at the HGMP Resource Centre, UK, for the DNA samples of backcross mice and computer analysis. We are also grateful to A. Knight for providing the 'stripe' program, to Keith Willison and Rob Nicholas for gifts of reagents, to Patrick Tam for help with interpretation of the *in situ* results and to Robin Lovell-Badge for his comments on the manuscript. This work was supported by the Cancer Research Campaign (A.A. and F.C.), the Australian Research Council (P.K. and E.W.) and by a Wellcome Trust Prize Studentship to F.C.

REFERENCES

1 Pabo, C.O. and Sauer, R.T. (1992) *Annu. Rev. Biochem.*, **61**, 1053–1095.
2 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) *Nature*, **346**, 240–244.
3 Koopman, P., Gubbay, J., Vivian, N., Goodfellow, P. and Lovell-Badge, R. (1991) *Nature*, **351**, 117–121.
4 Baxevanis, A.D. and Landsman, D. (1995) *Nucleic Acids Res.*, **23**, 1604–1613.
5 Nasrin, N., Buggs, C., Kong, X.F., Carnazza, J., Goebl, M. and Alexander-Bridges, M. (1991) *Nature*, **354**, 317–320.
6 Harley, V.R., Lovell-Badge, R. and Goodfellow, P.N. (1994) *Nucleic Acids Res.*, **22**, 1500–1501.
7 Harley, V.R., Jackson, D.I., Hextall, P.J., Hawkins, J.R., Berkovitz, G.D., Sockanathan, S., Lovell-Badge, R. and Goodfellow, P.N. (1992) *Science*, **255**, 453–456.
8 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) *Nature*, **359**, 531–533.
9 Whitfield, L.S., Lovell-Badge, R. and Goodfellow, P.N. (1993) *Nature*, **364**, 713–715.
10 Ferrari, S., Harley, V.R., Pontiggia, A., Goodfellow, P.N., Lovell-Badge, R. and Bianchi, M.E. (1992) *EMBO J.*, **11**, 4497–4506.
11 Giese, K., Cox, J. and Grosschedl, R. (1992) *Cell*, **69**, 185–195.
12 Pontiggia, A., Rimini, R., Harley, V.R., Goodfellow, P.N., Lovell-Badge, R. and Bianchi, M.E. (1994) *EMBO J.*, **13**, 6115–6124.
13 Tjian, R. and Maniatis, T. (1994) *Cell*, **77**, 5–8.
14 Laudet, V., Stehelin, D. and Clevers, H. (1993) *Nucleic Acids Res.*, **21**, 2493–2501.
15 Sockanathan, S., Cohen-Tannoudji, M., Collignon, J. and Lovell-Badge, R. (1993) *Genet. Res. Camb.*, **61**, 149.
16 van de Wetering, M., Oosterwegel, M., van Norren, K. and Clevers, H. (1993) *EMBO J.*, **12**, 3847–3854.
17 Foster, J.W., Dominguez-Steglich, M.A., Guioli, S., Kowk, G., Weller, P.A., Stevanovic, M., Weissenbach, J., Mansour, S., Young, I.D., Goodfellow, P.N., Brook, J.D. and Schafer, A.J. (1994) *Nature*, **372**, 525–530.
18 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) *Cell*, **79**, 1111–1120.
19 Wright, E., Hargrave, M.R., Christiansen, J., Cooper, L., Kun, J., Evans, T., Gangadharan, U., Greenfield, A. and Koopman, P. (1995) *Nature Genet.*, **9**, 15–20.
20 Denny, P., Swift, S., Connor, F. and Ashworth, A. (1992) *EMBO J.*, **11**, 3705–3712.
21 Connor, F., Cary, P.D., Read, C.M., Preston, N.S., Driscoll, P.C., Denny, P., Crane-Robinson, C. and Ashworth, A. (1994) *Nucleic Acids Res.*, **22**, 3339–3346.
22 Denny, P., Swift, S., Brand, N., Dabhade, N., Barton, P. and Ashworth, A. (1992) *Nucleic Acids Res.*, **20**, 2887.
23 The European Interspecific Collaborative Group (1994) *Hum. Mol. Genet.*, **3**, 621–627.
24 Wilkinson, D.G. and Nieto, M.A. (1993) *Methods Enzymol.*, **225**, 361–373.
25 Christiansen, J.H., Dennis, C.L., Wicking, C.A., Monkley, S.J., Wilkinson, D.G. and Wainwright, B.J. (1995) *Mech. Dev.*, in press.
26 Smith, D.B. and Johnson, K.S. (1988) *Gene*, **67**, 31–40.
27 Studier, F.W. (1991) *J. Mol. Biol.*, **219**, 37–44.
28 van de Wetering, M., Oosterwegel, M., Dooijes, D. and Clevers, H. (1991) *EMBO J.*, **10**, 123–132.
29 Wright, E.M., Snopek, B. and Koopman, P. (1993) *Nucleic Acids Res.*, **21**, 744.
30 Ellenberger, T. (1994) *Curr. Opin. Struct. Biol.*, **4**, 12–21.
31 Gubbay, J., Collignon, J., Koopman, P., Capel, B., Economou, A., Munsterberg, A., Vivian, N., Goodfellow, P. and Lovell-Badge, R. (1990) *Nature*, **346**, 245–250.
32 Vriz, S. and Lovell-Badge, R. (1995) *Gene*, **153**, 275–276.
33 Carlsson, P., Waterman, M.L. and Jones, K.A. (1993) *Genes Dev.*, **7**, 2418–2430.
34 Courey, A.J. and Tjian, R. (1988) *Cell*, **55**, 887–898.
35 Mermod, N., O'Neill, E.A., Kelly, T.J. and Tjian, R. (1989) *Cell*, **58**, 741–753.
36 Ptashne, M. (1988) *Nature*, **335**, 683–689.
37 Licht, J.D., Grossel, M.J., Figge, J. and Hansen, U.M. (1990) *Nature*, **346**, 76–79.
38 O'Shea, E.K., Klemm, J.D., Kim, P.S. and Alber, T. (1991) *Science*, **254**, 539–544.
39 Lupas, A., van-Dyke, M. and Stock, J. (1991) *Science*, **252**, 1162–1164.
40 Kessel, M. and Gruss, P. (1990) *Science*, **249**, 374–379.
41 Graham, A., Papalopulu, N. and Krumlauf, R. (1989) *Cell*, **57**, 367–378.
42 Wilkinson, D.G., Bhatt, S., Cook, M., Boncinelli, E. and Krumlauf, R. (1989) *Nature*, **341**, 405–409.
43 Hunt, P., Gulisano, M., Cook, M., Sham, M.H., Faiella, A., Wilkinson, D., Boncinelli, E. and Krumlauf, R. (1991) *Nature*, **353**, 861–864.
44 Marshall, H., Nonchev, S., Sham, M.H., Muchamore, I., Lumsden, A. and Krumlauf, R. (1992) *Nature*, **360**, 737–741.
45 Gendron-Maguire, M., Mallo, M., Zhang, M. and Gridley, T. (1993) *Cell*, **75**, 1317–1331.
46 Wolgemuth, D.J., Viviano, C.M., Gizang-Ginsberg, E., Frohman, M.A., Joyner, A.L. and Martin, G. (1987) *Proc. Natl. Acad. Sci. USA*, **84**, 5813–5817.
47 Needleman, S.B. and Wunsch, C.D. (1970) *J. Mol. Biol.*, **48**, 443–453.
48 Read, C.M., Cary, P.D., Crane-Robinson, C., Driscoll, P.C., Olga, M., Carrillo, M. and Norman, D.G. (1995) *Nucleic Acids Mol. Biol.*, in press.
49 Read, C.M., Cary, P.D., Crane-Robinson, C., Driscoll, P.C. and Norman, D.G. (1993) *Nucleic Acids Res.*, **21**, 3427–3436.