SHORT COMMUNICATION

Evidence That the SRY Protein Is Encoded by a Single Exon on the Human Y Chromosome

MARK A. BEHLKE,*,† JONATHAN S. BOGAN,* PEGGY BEER-ROMERO,* AND DAVID C. PAGE*

*Howard Hughes Research Laboratories at Whitehead Institute and Department of Biology, Massachusetts Institute of Technology, 9 Cambridge Center, Cambridge, Massachusetts 02142; and †Endocrine-Hypertension Division, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, 221 Longwood Avenue, Boston, Massachusetts 02115

Received January 1, 1993

To facilitate studies of the *SRY* gene, a 4741-bp portion of the sex-determining region of the human Y chromosome was sequenced and characterized. Two RNAs were found to hybridize to this genomic segment, one transcript deriving from *SRY* and the second cross-hybridizing to a pseudogene located 2.5 kb 5′ of the *SRY* open reading frame (ORF). Analysis of the *SRY* transcript using 3′ and 5′ rapid amplification and cloning of ends suggested that the entire *SRY* protein is encoded by a single exon. A 700-bp CpG island is located immediately 5′ of the pseudogene (and 2 kb 5′ of the *SRY* ORF). Within this CpG island lies the sequence CGCCCCCGC, a potential binding site for the EGR-1/WT1 family of transcription factors, some of which appear to function in gonadal development. © 1993 Academic Press, Inc.

During mammalian embryogenesis, the presence or absence of the *SRY* gene determines whether the bipotential gonads develop as testes or ovaries, in turn determining whether the remainder of the embryo develops as male or female (7, 8). In humans, transcription of *SRY* has been detected only in the adult testes, where a low abundance 1.1-kb transcript was found (17). The human *SRY* genomic locus is known to contain at least one large open reading frame (ORF), 471 bp of which have been reported (6, 17). If translated, this ORF would encode a protein containing an HMG1-like DNA-binding motif. The *SRY* transcription unit is otherwise uncharacterized.

Having previously cloned the sex-determining region of the Y chromosome (12), we began the present studies by sequencing a 4741-bp portion of distal Yp that should contain part or all of the *SRY* gene (Fig. 1A). The sequence contains a single long ORF, 669 bp in length, 471 bp of which are identical to the portion of the *SRY* gene previously reported (6, 17). We then attempted to isolate *SRY* cDNA clones from a library of 10^6 clones constructed using poly(A)$^+$ RNA from adult human testis. Plaque-hybridization screening with genomic DNA probes revealed no *SRY* clones, perhaps because of the low abundance of *SRY* transcripts.

To characterize the *SRY* transcription unit in the absence of conventional cDNA clones, we carried out 3′ and 5′ rapid amplification and cloning of ends (RACE) analysis (4) using human testis poly(A)$^+$ RNA as template and *SRY*-specific oligonucleotides as primers. For 3′ RACE, reverse transcription was primed using the oligonucleotide GAGGATCCGCGGCCGCGTCGACAGTTTTTTTTTTTTTTTT. Subsequent PCR reactions employed the 3′ primer GAGGATCCGCGGCCGCGTCGACAG and the 5′ *SRY* primers GAGTGAAGCGACCCATGA and CGTCGGAAGGCGAAGATG in a sequential, heminested fashion. The 3′ terminus of the gene was readily defined, with all six clones examined containing a poly(A) tract following nucleotide 749 (numbered as in Fig. 1), 133 bp after the termination codon that closes the ORF. A canonical polyadenylation signal (AATAAA, underlined in Fig. 1A) is located 21 bp upstream, at nucleotides 728–733, 112 bp 3′ of the termination codon. These results are at odds with a prior report that placed the polyadenylation signal 133 bp 3′ of the termination codon (17).

For 5′ RACE, reverse transcription was primed using the *SRY*-specific oligonucleotide CATCTTCGCCTTCCGAC. Subsequent PCR reactions employed reagents and 5′ primers from the GIBCO-BRL 5′ RACE system, using the 3′ *SRY*-specific primers TCGGGTATTTCTCTCTGTGC and TATCCCAGCTGCTTGCTG in a sequential, heminested fashion. In contrast to the 3′ results, 5′ RACE clones exhibited a diversity of 5′ ends, most of which fell into one of two clusters. The first cluster consisted of 5 clones (of 12 examined), all of which began at nucleotide -138. Clones composing the second cluster were somewhat shorter, beginning at nucleotides -81, -80, and -79. The remaining 4 clones were shorter still, beginning at nucleotides -65, -35, -2, and +38. Thus, 11 of 12 clones extended 5′ of the first ATG codon in the long ORF, and 9 of these 11 clones extended 5′ of the ORF. These results imply that *SRY* transcripts contain the entire genomic ORF and, in turn, that a single exon encompassing the ORF encodes the entire protein, predicted to contain 204 amino acid residues (Fig. 1A).

To test these inferences, *SRY* transcripts in adult hu-

GENOMICS 17, 736–739 (1993)
0888-7543/93 $5.00
Copyright © 1993 by Academic Press, Inc.
All rights of reproduction in any form reserved.

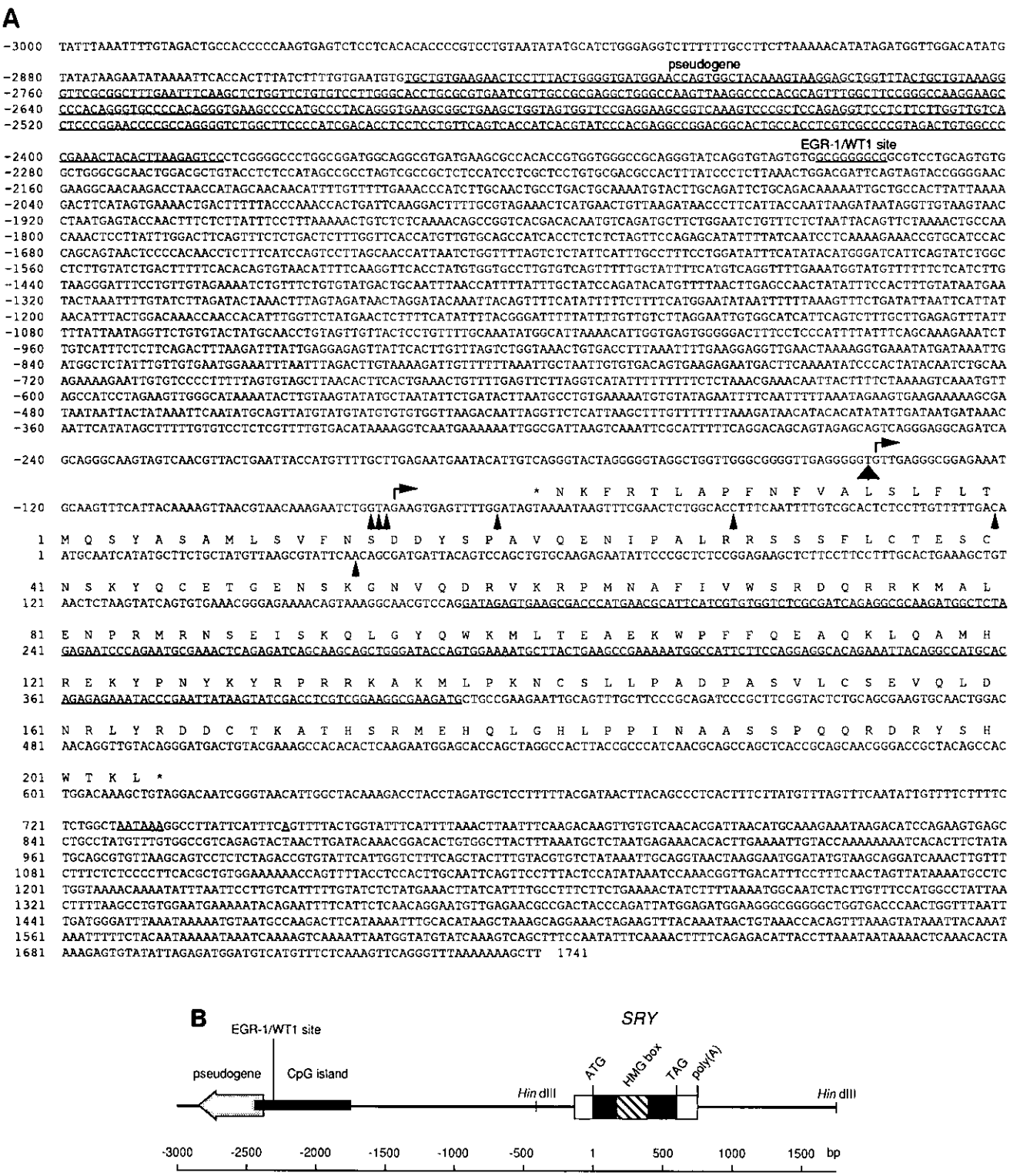

FIG. 1. The human *SRY* genomic locus and flanking regions. **(A)** Nucleotide sequence of a 4741-bp portion of the human Y chromosome (GenBank Accession No. L08063). Above the nucleotide sequence is the 223-amino-acid sequence corresponding to the entire *SRY* ORF; asterisks denote stop codons that bound the ORF. Nucleotides and amino acids are numbered with respect to the putative initiation codon, the first in-frame ATG. Nucleotides (172 to 411) encoding the HMG domain of the protein (amino acids 58 to 137) are underlined. Also shown are (1) a pseudogene (region of high nucleotide similarity to cDNA pDP1310 is underlined); (2) a potential binding site for transcription factors of the EGR-1/WT1 family; (3) the 5′ ends of *SRY* 5′ RACE clones (upward pointing arrowheads at −81, −80, −79, −65, −35, −2, and 38 represent single clones; heavy arrowhead at −138 represents five independent clones; (4) two *SRY* transcription initiation sites identified by Su and Lau (Ref. 18; horizontal arrows above sequence); (5) the polyadenylation signal, AATAAA (underlined; nucleotides 728 to 733); and the nucleotide (749) immediately preceding the poly(A) tract in 3′ RACE clones. **(B)** Schematic representation of the same 4741-bp region. Portions of the *SRY* transcription unit are indicated: 3′ and 5′ untranslated regions (open boxes), HMG-box-encoding region (hatched), and other coding sequences (solid). The positions of *Hin*dIII restriction sites are shown.

man testis poly(A)$^+$ RNA were further characterized using RT (reverse transcription) PCR. RT-PCR reactions were carried out using a single 3′ primer (CATCTTCGCCTTCCGAC, within the *SRY*-HMG box) in combination with various primers located at increasing distances 5′. Positive RT-PCR signals were obtained with ORF primers GAGTGAAGCGACCCATGA, GAATATTCCCGCTCTCCG, and AATAAGTTTCGAACTCTGGCA (the latter at the 5′ extreme end of the ORF), but no signal was obtained with ATTGTCAGGGTACTAGGGGG, a primer located 120 bp 5′ of the ORF (data not shown). These results confirm that *SRY* transcripts contain the entire genomic ORF. Accordingly, we have numbered nucleotides and amino acid residues with reference to the first in-frame ATG of the ORF (Fig. 1A).

Attempts to map the 5′ end of the transcription unit using S1 nuclease or RNase protection were unsuccessful. No *SRY*-specific signal could be obtained reproducibly, perhaps due to low expression of *SRY* in human adult testes. We were unable to confirm or refute the transcription initiation sites suggested by our 5′ RACE analysis.

Although a single exon defines the coding region of *SRY*, the gene might contain additional untranslated exons. In addition, genes other than *SRY* might be present in this portion of the genome. To pursue these possibilities, we hybridized a series of genomic probes spanning this 4.7-kb region to Northern blots of RNAs from various human tissues. We detected not only the 1.1-kb, testis-specific *SRY* transcript but also a 1.0-kb transcript that hybridized to genomic sequences located roughly 2.5 kb 5′ of the known *SRY* exon. This 1.0-kb transcript was present in all tissues examined, both male and female (data not shown), suggesting that it was not from *SRY* and might derive from an autosomal or X-linked gene. We isolated a corresponding cDNA clone (pDP1310) from a library constructed using RNA from a 46,XY lymphoblastoid cell line. Sequencing of the 540-bp insert of this partial cDNA (GenBank Accession No. L08647) revealed a 351-bp ORF at the 5′ end, a polyadenylation signal at nucleotide 482, and a poly(A) tract beginning at nucleotide 503. The cDNA was 86% identical in sequence to its Y genomic homolog (nucleotides −2380 through −2836) and was oriented opposite to *SRY* (Fig. 1). The Y genomic sequence differed from the cDNA by several frameshift, missense, and nonsense mutations, leading us to conclude that the Y-chromosomal homolog of the 1.0-kb transcript is probably a pseudogene. The sequence of cDNA pDP1310 was unrelated to any entry in GenBank; its function is not known. There was no obvious poly(A) tract at the 3′ end of the pseudogene. Lacking knowledge of the structure of the functional gene, we cannot conclude whether the pseudogene arose by gene duplication or by retroposition of an RNA transcript.

Computer analysis of the 4741-bp Y genomic sequence revealed no evidence of additional genes. We first searched for similarities to prior entries in GenBank using the BLAST (v1.5.3) program (1). Apart from the HMG box of *SRY*, no nucleotide or predicted amino acid sequence displayed sigificant similarity to previous entries. We then searched for potential coding exons using the GRAIL program (19), which identified two regions of interest, one corresponding to the known *SRY* exon and the other corresponding to the previously identified pseudogene.

Further analysis of the 4741-bp sequence revealed a 700-bp CpG-rich island approximately 2.0 to 2.7 kb 5′ of the *SRY* ORF. This CpG island overlaps the 5′ end of the pseudogene (Fig. 1B). Similar CpG islands function as transcriptional promoters for a number of genes (2, 5). In many cases, the transcription initiation site lies within the CpG island. Interestingly, a human XY female has been described in whom the *SRY* ORF is intact but whose Y chromosome bears a deletion that begins 1.7 kb 5′ of the *SRY* ORF and extends 25 to 50 kb farther 5′ (11). This deletion removes the 700-bp CpG island; it is possible, although entirely speculative, that loss of the CpG island is the cause of the patient's sex reversal. It will be of interest to determine whether *SRY* transcription—adult or embryonic—initiates within this CpG island. (If so, the transcript must be spliced within the 5′ untranslated region, since the transcript is shorter than the distance between the CpG island and the ORF.) Alternatively, the CpG island may simply be part of the pseudogene unit and play no role in *SRY* expression.

The nonamer CGCCCCCGC occurs near the 5′ end of the CpG island (Fig. 1). This sequence is the consensus binding site for a family of zinc-finger proteins that includes EGR-1 (also known as Zif268, Krox 24, NGF1-A, or TIS-8) and WT1 (3, 16). EGR-1 is known to be a transcriptional activator (9). Although a role for EGR-1 in sex determination has not been reported, it is possible that EGR-1 (or some related protein) might activate *SRY* transcription in the developing gonadal ridge or in the adult testis. More interestingly, deletions and point mutations in the human Wilms tumor gene *WT1*, when present in XY individuals, are often associated with abnormal development of the gonads and external genitalia (13, 14). The WT1 protein appears to be a transcriptional repressor (10), and, at least in mice, is expressed in the developing gonadal ridge (15). Perhaps one or more members of this family of zinc-finger proteins directly activate or repress transcription of the human *SRY* gene.

Addendum. Su and Lau recently reported (18) that mouse fibroblasts transfected with a human *SRY*-containing cosmid expressed *SRY* transcripts. These transcripts had poly(A) tracts beginning at the same site as in our 3′ RACE clones. Transcription was shown to initiate at two sites (indicated in Fig. 1A), each within one nucleotide of the two clusters of 5′ termini defined by our 5′ RACE analysis. Su and Lau reported 3817 bp of *SRY* genomic DNA sequence, all within the 4741 bp reported here. Our sequence differs from that of Su and Lau at 21 nucleotides, all within a 1.3-kb region. We sequenced this region again using new primers and were unable to reconcile these differences.

ACKNOWLEDGMENTS

We thank Laura Brown, Simon Foote, Adrienne Hilton, Tzielan Lee, George Mutter, and Anne Ridley for materials or experimental assistance and Andrew Zinn and Doug Vollrath for comments on the manuscript. This work was supported by National Institutes of Health and Searle Scholars Program/Chicago Community Trust.

REFERENCES

1. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* **215:** 403–410.
2. Bird, A. P. (1986). CpG-rich islands and the function of DNA methylation. *Nature* **321:** 209–213.
3. Cao, X. M., Koski, R. A., Gashler, A., McKiernan, M., Morris, C. F., Gaffney, R., Hay, R. V., and Sukhatme, V. P. (1990). Identification and characterization of the Egr-1 gene product, a DNA-binding zinc finger protein induced by differentiation and growth signals. *Mol. Cell. Biol.* **10:** 1931–1939.
4. Frohman, M. A., Dush, M. K., and Martin, G. R. (1988). Rapid production of full-length cDNAs from rare transcripts: Amplification using a single gene-specific oligonucleotide primer. *Proc. Natl. Acad. Sci. USA* **85:** 8998–9002.
5. Gardiner-Garden, M., and Frommer, M. (1987). CpG islands in vertebrate genomes. *J. Mol. Biol.* **196:** 261–282.
6. Gubbay, J., Collignon, J., Koopman, P., Capel, B., Economou, A., Munsterberg, 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.
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). DNA binding activity of recombinant SRY from normal males and XY females. *Science* **255:** 453–456.
8. 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.
9. Lemaire, P., Vesque, C., Schmitt, J., Stunnenberg, H., Frank, R., and Charnay, P. (1990). The serum-inducible mouse gene Krox-24 encodes a sequence-specific transcriptional activator. *Mol. Cell. Biol.* **10:** 3456–3467.
10. Madden, S. L., Cook, D.M., Morris, J. F., Gashler, A., Sukhatme, V. P., and Rauscher, F. (1991). Transcriptional repression mediated by the WT1 Wilms tumor gene product. *Science* **253:**1550–1553.
11. McElreavey, K., Vilain, E., Abbas, N. Costa, J. M., Souleyreau, N., Kucheria, K. Boucekkine, C., Thibaud, E., Brauner, R., Flamant, F., and Fellous, M. (1992). XY sex reversal associated with a deletion 5′ to the SRY HMG box in the testis-determining region. *Proc. Natl. Acad. Sci. USA* **89:**11016–11020.
12. Page, D. C., Fisher, E. M., McGillivray, B., and Brown, L. G. (1990). Additional deletion in sex-determining region of human Y chromosome resolves paradox of X,t(Y;22) female. *Nature* **346:** 279–281.
13. Pelletier, J., Bruening, W., Kashtan, C. E., Mauer, S. M., Manivel, J. C., Striegel, J. E., Houghton, D. C, Junien, C., Habib, R., Fouser, L., Fine, R. N. Silverman, B. L., Haber, D. A., and Housman, D. (1991). Germline mutations in the Wilms' tumor suppressor gene are associated with abnormal urogenital development in Denys–Drash syndropme. *Cell* **67:** 437–447.
14. Pelletier, J., Bruening, W., Li, F. P., Haber, D. A., Glaser, T., and Housman, D. E. (1991). WT1 mutations contribute to abnormal genital system development and hereditary Wilms' tumour. *Nature* **353:** 431–434.
15. Pelletier, J., Schalling, M., Buckler, A. J., Rogers, A., Haber, D. A., and Housman, D. (1991). Expression of the Wilms' tumor gene WT1 in the murine urogenital system. *Genes Dev.* **5:** 1345–1356.
16. Rauscher, F. J., Morris, J. F., Tournay, O. E., Cook, D. M., and Curran, T. (1990). Binding of the Wilms' tumor locus zinc finger protein to the EGR-1 consensus sequence. *Science* **250:** 1259–1262.
17. 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.
18. Su, H., and Lau, Y-C. (1993). Identification of the transcription unit, structural organization, and promoter sequence of the human sex-determining region Y (*SRY*) gene, using a reverse genetic approach. *Am. J. Hum. Genet.* **52:** 24–38.
19. Uberbacher, E. C., and Mural, R. J. (1991). Locating protein-coding regions in human DNA sequences by a multiple sensor-neural network approach. *Proc. Natl. Acad. Sci. USA* **88:** 11261–11265.