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 1Endocrine-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 CCGCCGCCGC, a potential binding site for the EGR-1/WT1 family of transcription factors, some of which appear to function in gonadal development.

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 GAGATCCGCCGCGCCTGACAGTTTTTTTTTTTTTT. Subsequent PCR reactions employed the 3' primer GAGATCCGCCGCGCCTGACAGT and the 5' SRY primers GAGTAAGCCGCCCATGA and CGTCGGAAGGCGAGATG 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 (AAATAAA, 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 CATCTTCCGCTTCGAC. Subsequent PCR reactions employed reagents and 5' primers from the GHBCO-BRL 5' RACE system, using the 5' SRY-specific primers TGCGGATTCTCTCAGTGC and TATCCCCAGCTTCTGCTG 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-
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. L08065). Above the nucleotide sequence is the 228-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/WTF1 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 HindIII restriction sites are shown.
human testis poly(A)+ RNA were further characterized using RT (reverse transcription) PCR. RT-PCR reactions were carried out using a single 3' primer (CATCTTCGCTTCCGAC, within the SRY-FMG box) in combination with various primers located at increasing distances 5'. Positive RT-PCR signals were obtained with ORF primers GAGTGAACGGACCATGA, GAATATTCGCGCTCTCC, and AATAAGTTTCCAGCTTGCGA (the latter at the 5' extreme end of the ORF), but no signal was obtained with ATTGTCAGGTTACTAGGGGG, 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 repitition 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 significant 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 cosm id expressed SRY transcripts. These transcripts had poly(A) tracts beginning at the same site as in our 5' 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 Scholar Program/Chicago Community Trust.

REFERENCES