GENOMICS 19, 310–319 (1994)

The Structure of the *Zfx* Gene on the Mouse X Chromosome

SHIUH-WEN LUOH AND DAVID C. PAGE[1]

Howard Hughes Research Laboratories at Whitehead Institute and Department of Biology, Massachusetts Institute of Technology, 9 Cambridge Center, Cambridge, Massachusetts 02142

Received June 21, 1993; revised October 31, 1993

Genes homologous to mouse *Zfx* have been identified on the X and Y chromosomes of all placental mammals examined. The genes of this *ZFX/ZFY* family appear to encode proteins comprising an amino-terminal acidic domain, a putative nuclear localizing signal, and a carboxy-terminal domain of 13 zinc fingers. These proteins likely function as transcription activators. Although roles for these proteins in sex determination, Turner syndrome, and spermatogenesis have been proposed, the biological processes in which these proteins function are not known. No comprehensive studies of gene structure have been reported for any member of the *ZFX/ZFY* family. Here, we report that mouse *Zfx* spans 50 kb and contains at least 11 exons. Exons 1 through 4 contain 5′ untranslated sequences, exons 5 through 10 encode the acidic domain, exon 10 also encodes the putative nuclear localizing signal, and exon 11 encodes 13 zinc fingers and contains the 3′ untranslated sequences. The 5′ untranslated exons exhibit complex patterns of differential splicing. At the 5′ end of this widely expressed gene, a 1.5-kb CpG island encompasses multiple transcription initiation sites as well as the first and second exons. The 5′ portion of the CpG island displays promoter activity. This knowledge of the *Zfx* gene structure allowed us to reconstruct the splicing and retroposition events by which the *Zfa* gene on mouse chromosome 10 arose from a *Zfx* transcript.

INTRODUCTION

Several genes are common to the X and Y chromosomes in humans or in mice. Some X–Y gene pairs are located in the pseudoautosomal region, where X–Y recombination is a normal and frequent event that maintains sequence identity between the two chromosomes (Goodfellow *et al.*, 1986; Keitges *et al.*, 1987; Gough *et al.*, 1990; Ellison *et al.*, 1992; Schiebel *et al.*, 1993; Slim *et al.*, 1993). Other X–Y genes are located in strictly sex-linked regions, where the X and Y chromosomes do not recombine (Page *et al.*, 1987; Fisher *et al.*, 1990; Kay *et al.*, 1991; Mitchell *et al.*, 1991; Salido *et al.*, 1992).

[1] To whom correspondence should be addressed.

The first X–Y genes to be identified in the strictly sex-linked portions of the mammalian X and Y chromosomes were *ZFX* and *ZFY* (Page *et al.*, 1987). Unlike pseudoautosomal sequences, where X–Y nucleotide identity is uninterrupted, high similarity between mammalian *ZFX* and *ZFY* sequences is intermittent and largely restricted to exons. *ZFX/ZFY* introns and flanking sequences generally exhibit much less nucleotide similarity (Schneider-Gädicke *et al.*, 1989a; Shimmin *et al.* 1993). Similarities in coding DNA sequences of *ZFX* and *ZFY* suggested that they derived from a single, common ancestral gene (Schneider-Gädicke *et al.*, 1989a). Homologs of the human *ZFX* and *ZFY* genes were found on the X and Y chromosomes of all placental mammals tested, suggesting that divergence of *ZFX* and *ZFY* from a common ancestral gene began prior to the radiation of placental mammals at least 60 to 80 million years ago (Page *et al.*, 1987).

Mammalian *ZFX* and *ZFY* encode similar proteins composed of two large domains (Ashworth *et al.*, 1989; Lau and Chan, 1989; Mardon and Page, 1989; Schneider-Gädicke *et al.*, 1989b; Mardon *et al.*, 1990; Palmer *et al.*, 1990). The amino-terminal half of these proteins is rich in acidic residues and has been shown to activate transcription when fused to the DNA-binding domain of yeast GAL4 (Mardon *et al.*, 1990). The carboxy-terminal half of these proteins consists of 13 zinc fingers and, by analogy to other zinc-finger proteins (Pabo and Sauer, 1992), probably binds to DNA (and/or RNA) in a sequence-specific manner. Thus, the ZFX and ZFY proteins may function as transcription activators.

The genomes of humans and most other placental mammals appear to contain two such genes: a single *ZFX* gene on the X chromosome and a single *ZFY* gene on the Y chromosome (Page *et al.*, 1987). By contrast, in the mouse, there are four related genes: *Zfx* on the X chromosome, *Zfy-1* and *Zfy-2* on the Y chromosome, and *Zfa* on chromosome 10 (Page *et al.*, 1987; Mardon *et al.* 1989; Mitchell *et al.*, 1989; Nagamine *et al.*, 1989; Ashworth *et al.*, 1990; Mardon *et al.*, 1990; Page *et al.*, 1990a). The coexistence of *Zfy-1* and *Zfy-2* on the mouse Y chromosome is the result of an intrachromosomal duplication that occurred during rodent evolution (Mardon *et*

0888-7543/94 $6.00

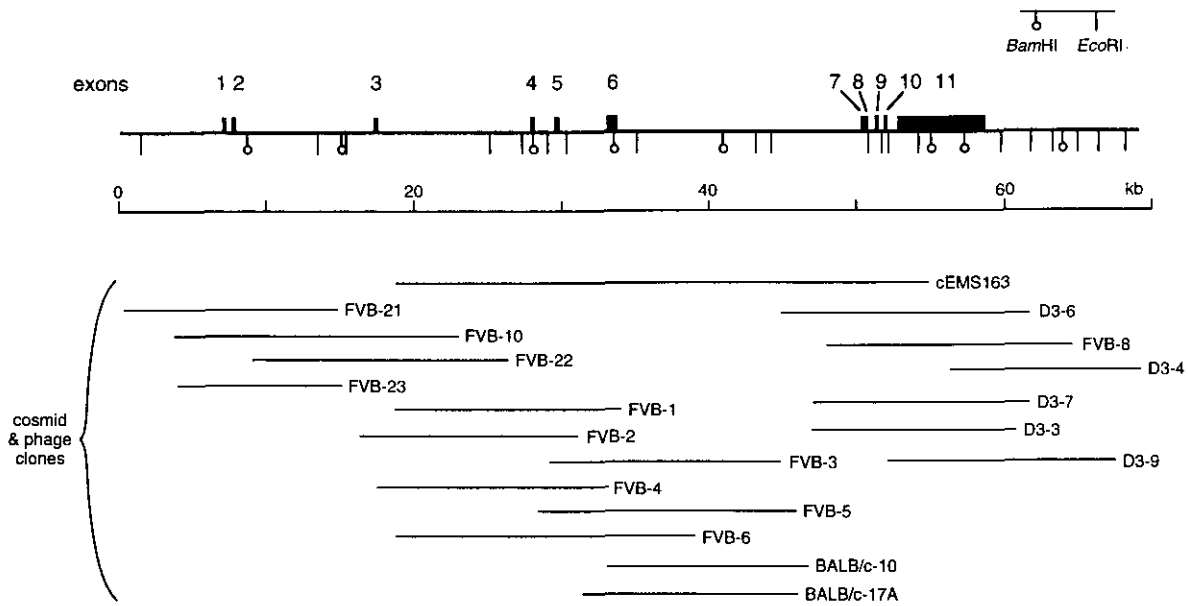

FIG. 1. Mouse *Zfx* genomic locus. A schematic representation of a 70-kb portion of the mouse X chromosome containing the entire *Zfx* transcription unit is shown at the top. The positions of 11 exons (black rectangles) are indicated, as are all *Eco*RI and *Bam*HI restriction sites. Shown below are 19 overlapping genomic DNA clones that span the region. These phage and cosmid clones are named according to the source of genomic DNA from which libraries were prepared. Hybridization and sequencing data allowed all exons to be precisely placed with respect to the *Eco*RI/*Bam*HI map with the following exceptions: (1) Exon 3 maps within a 2-kb segment present in phage FVB-4 but absent in FVB-6 and (2) Exons 7–9 are in close proximity, as indicated by the fact that PCR using cloned DNA as template, a "forward" (sense) primer from the 5′ end of exon 7, and a "reverse" (antisense) primer from the 3′ end of exon 9 yields a 1-kb amplification product (not shown).

al., 1989). The *Zfa* gene on mouse chromosome 10 is the result of retroposition of a processed *Zfx* transcript during rodent evolution (Ashworth *et al.*, 1990; Mardon *et al.*, 1990).

Many hypotheses have been forwarded concerning the biological functions of the mammalian *ZFX* and *ZFY* genes. Possible roles in gonadal sex determination, Turner syndrome, and spermatogenesis have been vigorously debated (Page *et al.*, 1987, 1990b; Burgoyne, 1989; Koopman *et al.*, 1989; Mardon *et al.*, 1989; Palmer *et al.*, 1989; Koopman *et al.*, 1991; Simpson and Page, 1991). In reality, the biological processes in which these proteins function are not known.

Although many facets of *ZFX/ZFY* molecular biology have been explored in mammals, no comprehensive study of the structure of any gene in this family has been reported. Here, we describe the structure of the mouse *Zfx* gene.

MATERIALS AND METHODS

Isolation of genomic DNA clones. *Zfx* clones were obtained from the following four libraries, all prepared using mouse genomic DNAs partially digested with *Mbo*I and inserted into the *Bam*HI site of lambda phage or cosmid cloning vectors: (1) a library of FVB/N female liver DNA in phage EMBL3 (this study), (2) a library of 129 male (D3 embryonic stem cell line) DNA in λ DASH II (Li *et al.*, 1992; a gift of En Li and Rudolph Jaenisch), (3) a library of BALB/c DNA in λ EMBL3 (a gift of Douglas Gray and Rudolph Jaenisch), and (4) a library of FVB/N male liver DNA in cosmid vector c2RB (Simpson and Page, 1991; kindly provided by Elizabeth M. Simpson). *Zfx* phage and cosmid clones in these libraries were identified by filter hybridization using radiolabeled portions of *Zfx* cDNAs pDP1115 and pDP1119 (Mardon *et al.*, 1990) as probes. To ensure specificity, all hybridizations were carried out at high stringency at 47°C in 50% formamide, 5× SSC (1× SSC = 0.15 *M* NaCl, 15 m*M* Na citrate, pH 7.4), 1× Denhardt's (0.02% Ficoll 400, 0.02% polyvinyl pyrrolidone, 0.02% bovine serum albumin), 20 m*M* $NaPO_4$, pH 6.6, 50 μg/ml denatured salmon sperm DNA, 1% SDS (sodium dodecyl sulfate), and 10% dextran sulfate. Following hybridization, filters were washed three times for 15 min each at 65°C in 0.1× SSC, 0.1% SDS prior to exposure with X-ray film. Clones found to hybridize with both the 5′ and 3′ portions of the *Zfx* cDNAs were considered likely to contain not *Zfx*, but *Zfa*, a small autosomal gene derived from a processed *Zfx* transcript by retroposition (Ashworth *et al.*, 1990; Mardon *et al.*, 1990). Such clones were not purified or analyzed further. However, 19 clones hybridized with only a portion of the *Zfx* cDNAs and were purified. These included 11 phages from the FVB/N library, 5 from the 129 library, 2 from the BALB/c library, and 1 FVB/N cosmid. Analysis of these 19 clones by restriction mapping and hybridization demonstrated that they formed a single contiguous array (Fig. 1). To generate templates for intron/exon boundary sequencing, exon-containing genomic fragments were subcloned from phage and cosmid into plasmid vector pBluescript KS^- (Stratagene) in both orientations.

Isolation of cDNA clones. *Zfx* clones were obtained from the following two mouse cDNA libraries, both prepared using poly(A)$^+$ RNAs: (1) a library prepared from FVB/N adult testis with oligo(dT) priming of first-strand synthesis (Mardon and Page, 1989; this library kindly provided by Graeme Mardon) and (2) a library prepared from skin fibroblasts with random priming of first-strand synthesis (a gift of Reinhard Faessler and Rudolph Jaenisch). *Zfx* clones in these libraries were initially identified by filter hybridization at high stringency using a combination of two 700-bp *Eco*RI–*Bam*HI fragments—one derived from the 5′-most portion of *Zfx* cDNA pDP1115 and the other from the 5′-most portion of *Zfx* cDNA pDP1119 (Mardon *et al.*, 1990). Nine cDNA clones initially identified by these 700-bp probes were subsequently tested by hybridization with two oligonucleotides, 5′-GGGCGCTCAGAACGGCGGTCG-3′ and 5′-CCGGCCTGCAGCACCCGCCA-3′, representing the 5′ termini of cDNAs pDP1119 and pDP1115, respectively. Prior to hybridization, oligonucleotides were end-labeled using T4 polynucleotide kinase and [γ-^{32}P]ATP. Membranes were prehybridized and hybridized at 42°C in 6× SSC, 5× Den-

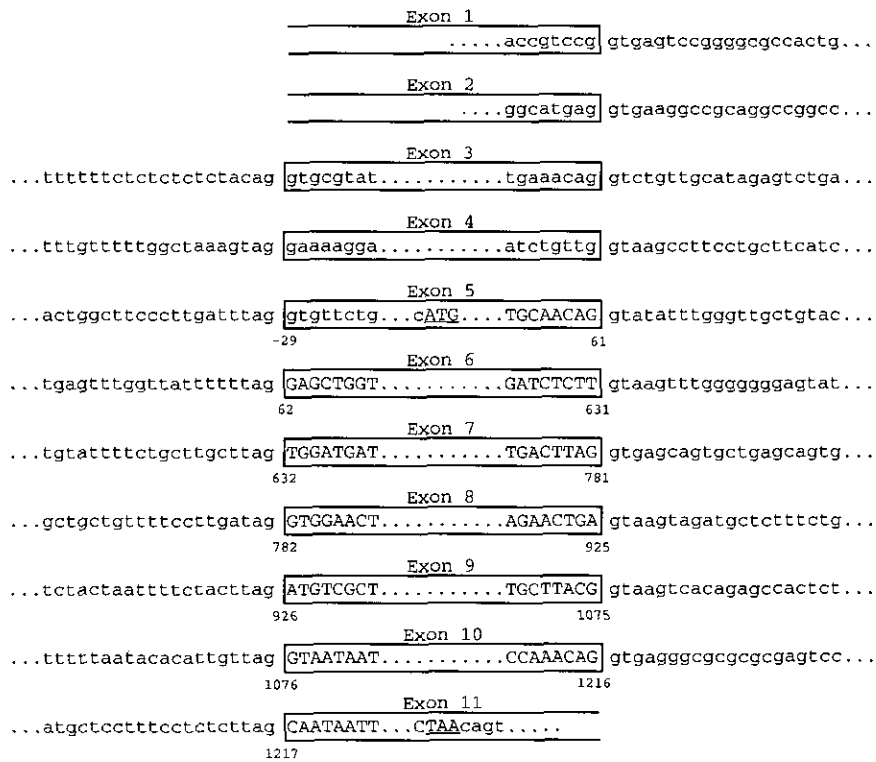

FIG. 2. Intron/exon boundaries in the mouse *Zfx* gene. Nucleotide sequences at splice sites for all 11 identified exons are shown. Noncoding and intervening sequences are in lowercase letters. Coding sequences are in uppercase letters. Translation initiation (ATG) and termination (TAA) codons are underlined. The first and last nucleotides of each coding exon are numbered according to their positions in cDNAs, relative to the first nucleotide of the initiator codon.

hardt's, 0.5% SDS, 0.5% $Na_4P_2O_7$; prehybridization also included 100 mg/ml sheared salmon sperm DNA. Membranes were then washed three times for 20 min at 56°C in 6× SSC, 0.1% SDS prior to exposure to X-ray film. Three of the nine cDNA clones identified in the initial screening were found to hybridize with the 5′ oligonucleotides, suggesting that they extended at least as far 5′ as cDNAs pDP1115 and pDP1119. The cDNA inserts of the three clones were transferred into the *Eco*RI site of pBluescript KS$^-$, generating plasmids pDP1296 (from the testis library) and pDP1536 and pDP1537 (both from the fibroblast library). The other six cDNA clones identified in the initial screening did not hybridize with the oligonucleotides and were not analyzed further.

DNA sequencing. Nucleotide sequencing of *Zfx* cDNA and genomic DNA clones was carried out by dideoxynucleotide chain termination (Sanger *et al.*, 1977) using synthetic oligonucleotide primers and modified T7 polymerase (Sequenase 2.0, U.S. Biochemical). Sequencing templates were either (1) single-stranded DNAs rescued from pBluescript KS$^-$ constructs using helper phage VCS-M13 or (2) supercoiled, double-stranded DNAs that had been twice purified by CsCl gradient centrifugation. Cloned cDNAs were sequenced by primer walking. To map intron/exon boundaries, genomic subclones were partially sequenced using oligonucleotide primers based on cDNA sequence. A 2.9-kb segment of genomic DNA (including a CpG island) from the 5′ end of the *Zfx* gene was sequenced using as templates a variety of plasmid subclones, some generated by restriction digestion and others generated using *Exo*III and S1 nucleases (Henikoff, 1984). Both strands of this 2.9-kb segment were completely sequenced. The high G+C content of this region made analysis of the sequence difficult. To resolve ambiguities due to compressions and other artifacts, we employed the substituted nucleotides 7-deaza-dGTP and dITP (U.S. Biochemical) (Mizusawa *et al.*, 1986) and resolved extension products on 6% polyacrylamide/8 *M* urea/20% formamide gels.

Plasmid constructs for transcriptional analysis. To map transcription initiation sites and promoter activity, 5′ portions of the mouse *Zfx* gene were placed upstream of human growth hormone (*hGH*) or bacterial chloramphenicol acetyltransferase (CAT) reporter genes. Four different *Zfx*–CAT constructs were produced as described in Fig. 8.

A single *Zfx–hGH* construct was produced as follows: A 2-kb *Bss*HII–*Eag*I fragment (approximately nucleotides −2000 to +73; nucleotides numbered as in Fig. 5A) from the mouse *Zfx* genomic locus was purified and the 5′ overhanging ends were "filled in" using the DNA polymerase Klenow fragment. pSGO (a promoterless, genomic *hGH* construct) (Selden *et al.*, 1986; a gift of Richard Selden) was digested with *Sal*I and *Bam*HI, and the resulting 5′ overhanging ends were repaired; the 2-kb *Zfx* fragment was ligated to pSGO so that *Zfx* was upstream of *hGH* in the sense orientation. During these steps, the *Bss*HII, *Eag*I, and *Bam*HI sites were regenerated. The *Bss*HII site of the *Zfx–hGH* construct was then cleaved, and the resulting 5′ overhanging ends were repaired and ligated to *Eco*RI linkers. This allowed the *Zfx–hGH* minigene to be released from the vector by *Eco*RI digestion and inserted into the single *Eco*RI site of plasmid pSVori, which contains the SV40 replication origin. (Plasmid pSVori was derived from pgTat-CMV (Chang and Sharp, 1989; a gift of Phillip Sharp) by digestion with *Spe*I and *Eco*RI, repair of 5′ overhanging ends, and intramolecular ligation to recircularize the construct and regenerate the single *Eco*RI site.) The final result of these manipulations is a plasmid containing an SV40 replication origin and a promoterless *hGH* gene driven by *Zfx* 5′ genomic sequences. All plasmids were purified twice by CsCl gradient centrifugation prior to transfection.

S1 nuclease protection and CAT reporter gene assays. COS-7, Hela, 3T3, and HepG2 cell lines obtained from the American Type Culture Collection were cultured in DMEM supplemented with 10% fetal calf serum.

The *Zfx–hGH* construct was electroporated into COS-7 cells (ECM600 apparatus; BTX, San Diego, CA). Briefly, COS-7 cells grown to 40–60% confluency were harvested, washed twice, and resuspended in ice-cold PBS (with no calcium or magnesium) at a density of 5 to 10 million cells per ml. Resuspended cells were electroporated

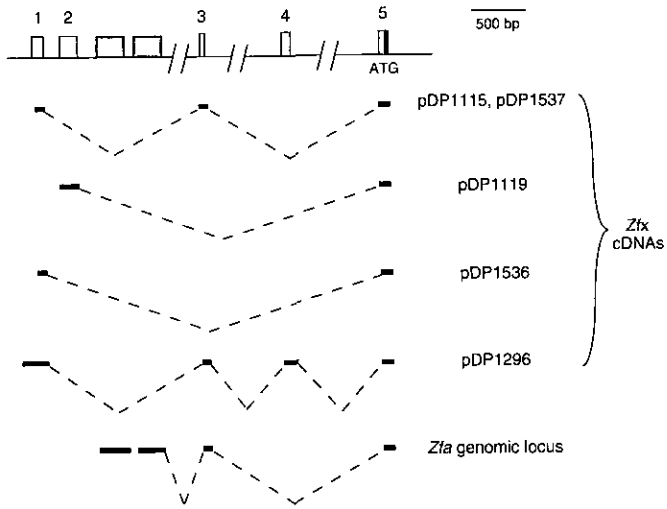

FIG. 3. Alternative splicing of *Zfx* 5′ UTR exons as revealed by *Zfx* cDNAs and by *Zfa* retroposon. A schematic representation of the first five exons of the mouse *Zfx* gene is shown at the top. Exons 1–4 are not translated (open boxes). Exon 5 contains the initiator codon and some coding sequence (black box). Located between exons 2 and 3 are DNA sequences (stippled boxes) highly similar to the 5′ portion of the *Zfa* gene. Nucleotide sequencing of five mouse *Zfx* cDNAs revealed the patterns of splicing shown. The five cDNAs showed no alternative splicing farther downstream. At the bottom, the splicing pattern of the transcript whose retroposition produced the *Zfa* gene is shown.

with 3 to 7 μg of plasmid DNA at 1750 to 2000 V/cM and 50 μF with a pulse length of 0.35 to 0.4 ms. Cells were then chilled on ice for 10 min before plating; they were harvested 3 to 4 days after electroporation. Cytoplasmic RNAs were isolated from transfected and nontransfected cells (Chang and Sharp, 1989) and used in S1 nuclease protection assays (Sambrook *et al.*, 1989). The probe used was a double-stranded *Sac*II/*Eag*I fragment (nucleotides −454 to +73; numbered as in Fig. 5A) carrying a single ^{32}P end-label at the *Eag*I site. Thirty thousand counts per minute of probe were hybridized to 5 μg of COS-7 cytoplasmic RNA or 20 μg of yeast tRNA at 65°C overnight. Digestions were carried out at room temperature for 45 to 60 min with 100 to 1600 units of S1 nuclease. Digestion products were precipitated and resolved on a denaturing gel containing 6% acrylamide and 8 *M* urea.

Four different *Zfx*–CAT constructs (1.3S, 1.3AS, 0.9S, and 0.9AS) were transfected into Hela, 3T3, and HepG2 cells by calcium phosphate precipitation (Sambrook *et al.*, 1989). Cells were harvested 48 h after transfection, and lysates were assayed for CAT activity by thin-layer chromatography (Sambrook *et al.*, 1989).

RESULTS

Intron/Exon Structure of the Mouse Zfx Gene

We had previously characterized two mouse *Zfx* cDNA clones, pDP1115 and pDP1119 (Mardon *et al.*, 1990), both having poly(A) tracts at their 3′ ends. The two cDNAs have identical coding regions but differ in their 5′ and 3′ untranslated regions (Mardon *et al.*, 1990). To clone the mouse *Zfx* gene, genomic DNA libraries were screened by hybridization with pDP1115 and pDP1119. A total of 19 genomic clones (18 lambda phage and one cosmid) were isolated, and they formed a single overlapping array 70 kb in length (Fig. 1). As described below, this contig appears to encompass the entire mouse *Zfx* gene as well as 8 kb of 5′ flanking and 12 kb of 3′ flanking DNA sequences.

We then determined the intron/exon structure of the mouse *Zfx* gene. To roughly localize exons, portions of cDNAs, or synthetic oligonucleotides based on cDNA sequence, were hybridized to restriction digests of genomic DNA clones. The exonic portions of the corresponding genomic subclones were then sequenced. We sequenced genomic DNA segments corresponding to not only the inserts of cDNAs pDP1115 and pDP1119 but also those of three new *Zfx* cDNA clones whose identification is reported below. The genomic DNA sequences were completely consistent with previously and newly reported *Zfx* cDNA sequences (Mardon *et al.*, 1990; see below). A total of 11 exons were defined. These exons, distributed across 50 kb of genomic DNA (Fig. 1), contain a total of 7216 nucleotides. Exons 1 through 4 contain 5′ untranslated sequences, exon 5 contains the initiator codon, and exon 11 contains the stop codon and 3′ untranslated sequences. All exons but exon 11 were completely sequenced. In the case of exon 11, the largest and most 3′ exon, comparative restriction mapping demonstrated collinearity of cDNA and genomic DNA clones, suggesting that the region contains no introns. The nucleotide sequences at the intron/exon boundaries (Fig. 2) are typical of those observed near splice sites, with all introns conforming to the GT–AG rule (Padgett *et al.*, 1986). Earlier studies had identified splice sites bounding a few of the more 3′ exons of mouse *Zfx* (Ashworth *et al.*, 1990; Mardon *et al.*, 1990), and these previous results agree with those shown in Figs. 1 and 2.

Alternative Splicing of Zfx 5′ UTR Exons

Mouse *Zfx* cDNA clones pDP1115 and pDP1119 had previously been shown to differ in their 5′ untranslated regions, perhaps as the result of differential splicing or alternative promoter usage (Mardon *et al.*, 1990). To further explore the 5′ portion of the transcription unit, we cloned three additional mouse *Zfx* cDNAs that appeared to extend at least as far 5′ as pDP1115 and pDP1119. Complete sequencing of these three cDNAs (pDP1296, pDP1536, and pDP1537) revealed that they lack poly(A) tracts and have 3′ ends within exon 7 but are otherwise collinear with the coding regions of pDP1115 and pDP1119 (i.e., exons 5 and 6 and the 5′ portion of exon 7). We suspect that the three new cDNAs are artifactually truncated at their 3′ ends, perhaps because of internal priming during reverse transcription. The 5′ terminal nucleotide of one of the new cDNAs, pDP1536, is identical to that of the previously characterized cDNA pDP1115. The other two new cDNAs, pDP1537 and pDP1296, extend, respectively, 7 and 260 nucleotides farther 5′.

Within the 5′ untranslated region, four different pat-

```
GAAAAGGAGA ATTAAGGTCA TTGGGATGTC TTAAAAGAAG TGGGCCTTAG
GAAGTTCTGT GTTCAGGATC CAGAAATCTG TTG
```

FIG. 4. Nucleotide sequence of *Zfx* exon 4.

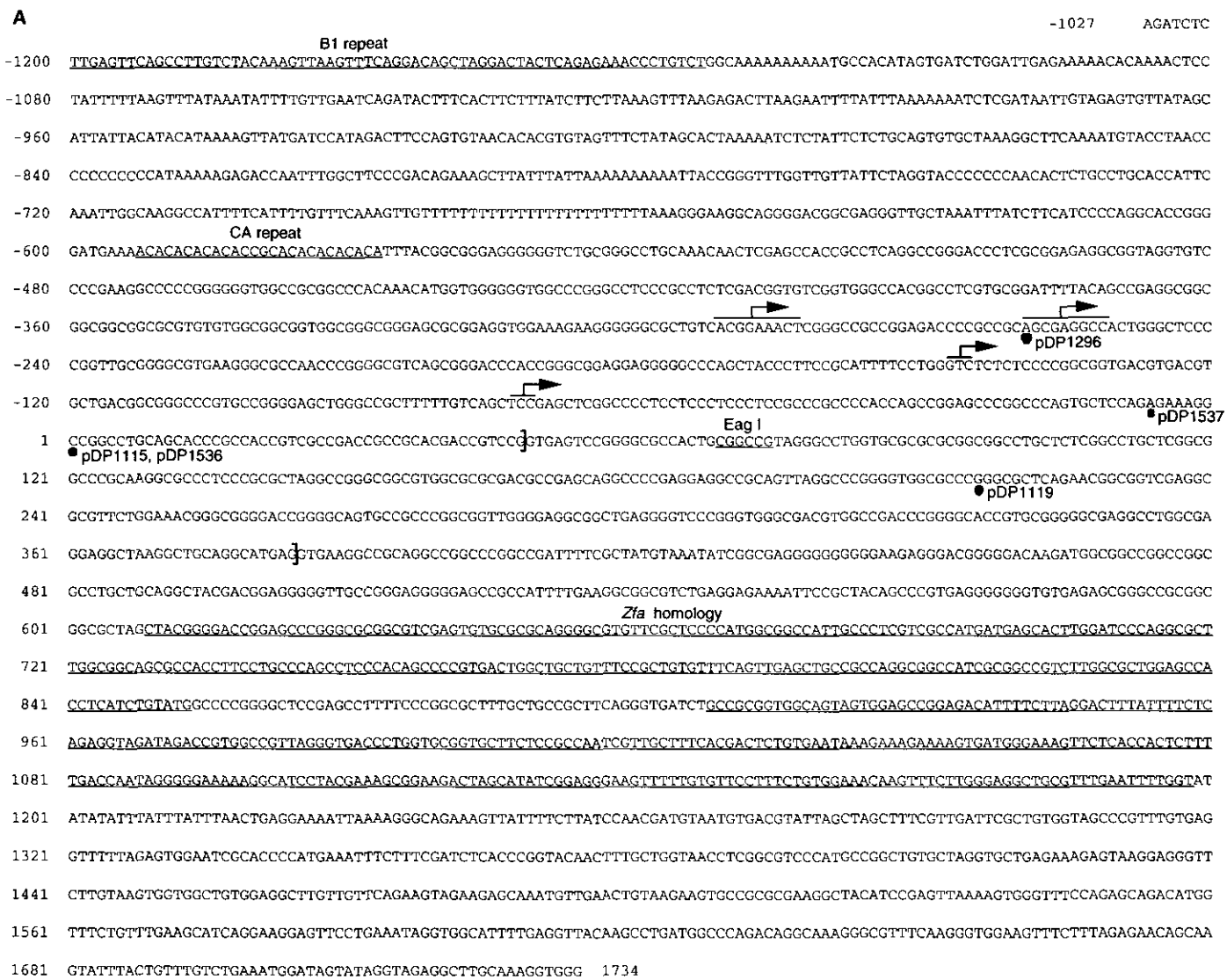

FIG. 5. 5′ portion of the *Zfx* transcription unit. (**A**) Nucleotide sequence of a 2.9-kb segment of the mouse X chromosome (GenBank Accession No. L19715) that includes (1) a B1 repeat (underlined), (2) a CA simple sequence repeat (underlined), (3) four transcription initiation sites (arrows above sequence) identified by S1 nuclease protection, (4) 5′ ends of five cDNAs (dots beneath sequence; GenBank Accession Nos. pDP1296, L19714; pDP1115, M32308; pDP1119, M32309), (5) 3′ boundaries of exons 1 and 2 (right brackets), (6) an *Eag*I site end-labeled in S1 nuclease protection studies and used in constructing a *Zfx–hGH* minigene, and (7) a 590-bp region of 95% identity to *Zfa* (underlined; note 54-bp discontinuity in center of this region). Nucleotides are numbered according to the 5′ nucleotide in cDNAs pDP1115 and pDP1536. (**B**) Schematic representation of the same 2.9-kb region. A restriction map is shown at the top (*Bgl*II, Bg; *Pst*I, P; *Sac*II, SII; *Eag*I, E; *Bam*HI, B); the position of a 1.5-kb CpG island is indicated by the thick black line. The positions of (1) a 0.9-kb *Pst*I fragment exhibiting promoter activity, (2) *Zfa*-homologous sequences, (3) the 5′ portions (exons 1 and 2) of five mouse *Zfx* cDNAs (and of one human *ZFX* cDNA (pDP1132; cDNA 1 in Schneider-Gädicke *et al.*, 1989b) whose 5′ sequence is nearly identical to that of mouse *Zfx* exon 1), and (4) the hybridization probe used and products observed in S1 nuclease protection studies are shown below.

terns of splicing are found among the five cDNA clones (Fig. 3). Four of the five cDNA clones contain at least part of exon 1. Exon 2 is present only in cDNA pDP1119. Exon 4, found only in cDNA pDP1296, has not been reported previously; its sequence is presented in Fig. 4. As discussed below, analysis of *Zfa*, a *Zfx*-derived retroposon, suggests a fifth pattern of 5′ UTR splicing.

A CpG Island at the 5′ End of the Zfx Transcription Unit

CpG islands are unusual segments of the genome that have a high G + C content and in which the dinucleotide CpG is abundant and unmethylated. They are typically about a kilobase in length. In many mammalian genes, a CpG island overlaps the transcription start site (Bird, 1987; Larsen *et al.*, 1992). CpG islands are known to be located near the 5′ ends of both the human *ZFY* and *ZFX* genes. Hybridization studies suggested that the nucleotide sequences of the human *ZFX* and *ZFY* CpG islands are similar and have been conserved during the evolutionary radiation of placental mammals (Page *et al.*, 1987; Schneider-Gädicke *et al.*, 1989a).

All five mouse *Zfx* cDNAs have a high G + C content near their 5′ termini (corresponding to exons 1 and 2).

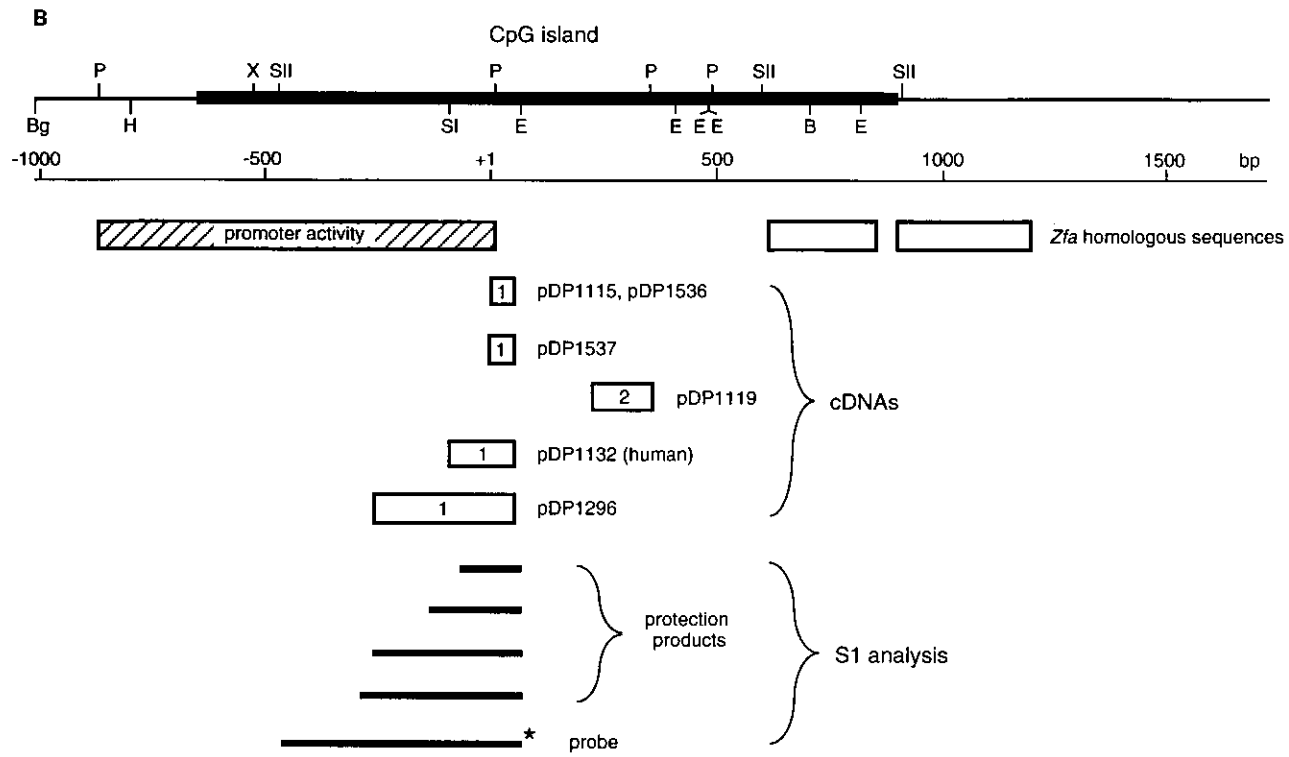

FIG. 5—*Continued*

To further characterize the 5′ portion of the gene, we sequenced a 2.9-kb segment of genomic DNA containing *Zfx* exons 1 and 2 (Fig. 5A). Within this 2.9-kb segment, we identified a 1.5-kb region in which 80% of the nucleotides are either G or C and in which CpG and GpC dinucleotides are equally abundant. This CpG island appears to contain the entirety of exons 1 and 2, both of which are untranslated (Fig. 5B).

The 5′ portion of a previously described human *ZFX* cDNA, pDP1132, has a high G + C content (Schneider-Gädicke *et al.*, 1989b) and its sequence is very similar to exon 1 of mouse *Zfx*. In fact, the human cDNA clone is spliced at a position corresponding precisely to the 3′ boundary of mouse *Zfx* exon 1. Upstream of this splice, the human cDNA is collinear with mouse *Zfx* cDNA pDP1296 (Fig. 5B). A detailed comparison of the CpG islands found in the mouse *Zfx* and human *ZFX* and *ZFY* genes will be presented elsewhere (S.-W.L., D.C.P., and colleagues, in preparation).

A simple sequence repeat is found just 5′ of the CpG island in the mouse *Zfx* gene (Fig. 5A). The length of this CA dinucleotide repeat varies among inbred mouse strains, making it useful as a marker in genetic linkage studies (Dietrich *et al.*, 1992).

The Zfx CpG Island and the Genesis of an Autosomal Retroposon, Zfa

Mouse chromosome 10 carries the *Zfa* gene, a transcribed retroposon derived from a processed *Zfx* transcript (Mitchell *et al.*, 1989; Nagamine *et al.*, 1989; Ashworth *et al.*, 1990; Mardon *et al.*, 1990; Page *et al.*, 1990a). Thus, the *Zfa* gene is akin to a *Zfx* cDNA. Comparison of the sequence of the *Zfa* genomic locus (Ashworth *et al.*, 1990) with that of the *Zfx* CpG island and exons allowed us to more fully reconstruct the splicing pattern of the *Zfx* transcript that gave rise to *Zfa*. This comparative sequence analysis revealed that *Zfa* derived from a processed transcript comprising one or two novel exons (not present in any available *Zfx* cDNA clone) followed by *Zfx* exons 3, 5, 6, and 8–11.

The existence of these novel exons was inferred from nucleotide sequence comparisons of *Zfa* (Ashworth *et al.*, 1990) and the *Zfx* CpG island, which are 95% identical over a span of more than 500 bp (Figs. 5A and 6). The 3′ boundary of the *Zfx*/*Zfa* similarity (i.e., the 3′ boundary of the novel exon(s)) is defined by a GT dinucleotide that could serve as a splice donor (Fig. 6). Although no available *Zfx* cDNA employed this splice site, the *Zfa* retroposon appears to have been cleanly spliced to *Zfx* exon 3 at this point (Fig. 6). The 500-bp region of similarity between *Zfa* and the *Zfx* CpG island is interrupted by a 54-bp deletion in *Zfa* (Figs. 5A and 6). The 54-bp deletion in *Zfa* could represent either a *Zfx* intron that was spliced out prior to retroposition or a deletion that occurred during or following retroposition. The presence of this 54-bp sequence in the CpG island of human *ZFX* (unpublished results) argues against an insertion into mouse *Zfx* following the retroposition of *Zfa*. Thus, the 500-bp region of similarity between *Zfa* and the *Zfx* CpG island might contain either one or two novel exons.

The 5′ border of the 500-bp region of *Zfa*/*Zfx* similarity is defined by a 10-bp sequence, CATATTTCCT (boxed in Fig. 6). This sequence is present in *Zfa* but

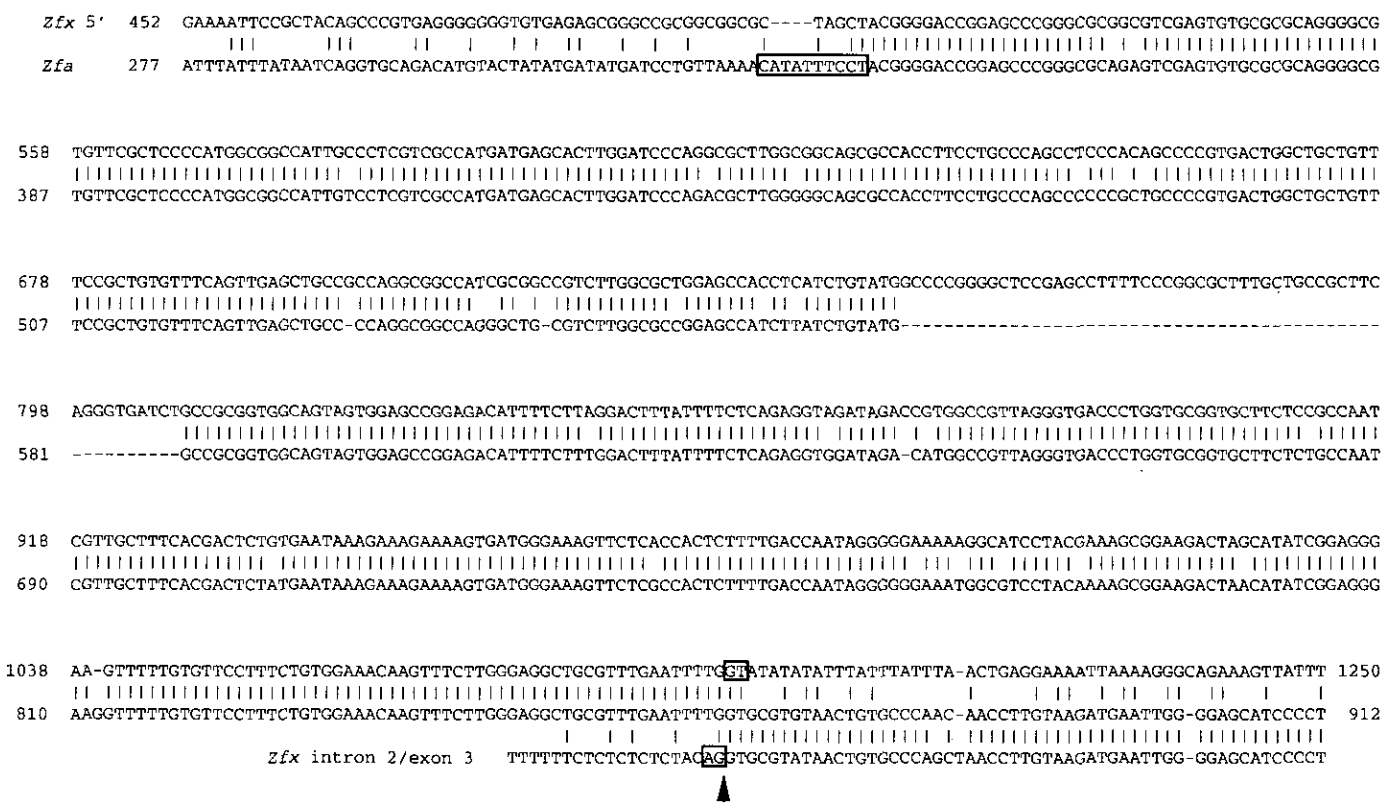

FIG. 6. Alignment of DNA sequences from 5′ portions of mouse *Zfx* and *Zfa* genes. A 636-bp portion of the *Zfa* gene (nucleotides numbered according to Ashworth *et al.*, 1990) is aligned with two portions of the *Zfx* gene (nucleotides numbered as in Fig. 5A). Dashes indicate gaps in one sequence as compared with another. A 10-nucleotide sequence demarcating the 5′ boundary of *Zfx/Zfa* similarity (and also present at the 3′ end of *Zfa*; Ashworth *et al.*, 1990) is boxed. Also boxed are (1) the GT dinucleotide (a putative splice donor) demarcating the 3′ boundary of *Zfx/Zfa* similarity within *Zfx* intron 2 and (2) the AG splice acceptor immediately 5′ of *Zfx* exon 3. To the 3′ side of this AG, *Zfa* is 96% identical to *Zfx* exon 3. The 5′ boundary of *Zfx* exon 3 is marked by an arrowhead.

absent in *Zfx*. As pointed out by Ashworth *et al.*, (1990), an identical 10-bp sequence defines the 3′ boundary of similarity between *Zfx* and *Zfa* genomic sequences, i.e., between the 3′-most exon of *Zfx* and sequences 3′ of the *Zfa* open reading frame on chromosome 10. Again, the 10-bp sequence is present in *Zfa* but absent in *Zfx*, and in this case, the sequence is immediately preceded by a short poly(A) tract (Ashworth *et al.*, 1990). Taken together, these results demonstrate unequivocally that *Zfa* arose by retroposition of a *Zfx* transcript and that the retroposed unit is delimited by the 10-bp direct repeats.

Transcription Initiation within the CpG Island

Transcription initiation sites have not been reported and promoters have not been defined for the *ZFX* (or *ZFY*) genes in any species. To determine the initiation site(s) for mouse *Zfx*, we performed S1 nuclease protection assays. Initial experiments employed poly(A)$^+$ RNAs from female mouse tissues but yielded no conclusive results (not shown), perhaps because *Zfx* transcripts are of low abundance in these tissues. To overcome this problem, we exploited SV40 origin-dependent amplification in COS-7 monkey kidney cells. In brief, 2 kb of *Zfx* genomic sequence (containing and extending 5′ of exon 1) was placed upstream of a promoterless segment of human growth hormone genomic DNA (a reporter) in a plasmid containing the SV40 replication origin. Following electroporation into COS-7 cells, the *Zfx–hGH* fusion transcript was produced at high levels (Northern analysis not shown). S1 nuclease protection analysis of these transcripts demonstrated major products of 146 and 220 nucleotides and minor products of 329 and 362 nucleotides (Fig. 7). These results suggest the existence of major initiation sites at −73 and −147 and minor sites at approximately −256 and −289 (see Fig. 5A for positions of these four start sites). All four sites fall within the CpG island, and there are no upstream TATA or CCAAT elements. The S1 protection products are shown schematically in Fig. 5B, where they can be compared with the 5′ portions of cloned cDNAs.

Given these results, we suspected that the promoter of *Zfx* would be located within or just 5′ of the CpG island. This prediction was tested by assaying the ability of segments of *Zfx* genomic sequence to drive expression of a promoterless CAT reporter gene. The *Zfx* sequences tested were a 1.3-kb fragment (nucleotides −871 to +490, containing exons 1 and 2 and upstream sequences) and a 0.9-kb fragment (nucleotides −871 to +10, containing part of exon 1 and upstream sequences). These sequences were inserted in both sense and antisense orientations upstream of the CAT reporter gene.

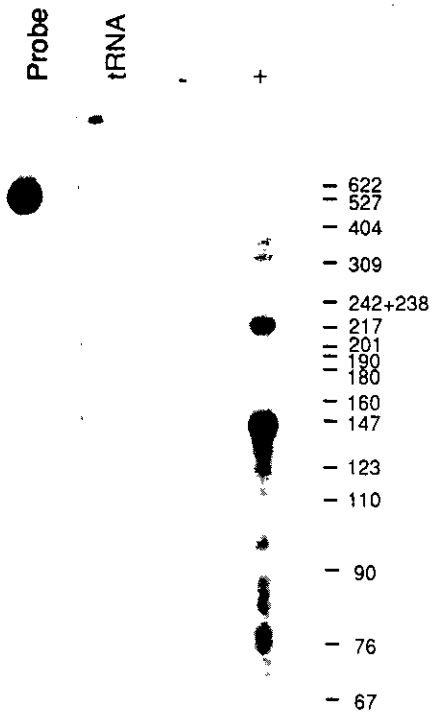

FIG. 7. Mapping of *Zfx* transcription initiation sites by S1 nuclease protection. The probe was a double-stranded 530-bp *Sac*II/*Eag*I fragment carrying a single ^{32}P end-label at the *Eag*I site. This probe was hybridized overnight at 65°C with 20 μg of yeast tRNA or with 5 μg of cytoplasmic RNA from nontransfected (−) or transfected (+) COS-7 cells. Results shown were produced using 400 units of S1 nuclease per assay. Essentially identical results were obtained using 100 or 1600 units of S1 nuclease. Major products of 146 and 220 nucleotides and longer, less abundant products of 329 and 362 nucleotides are observed. We suspect that the shorter, low-abundance products result from degradation of major products.

After transient transfection into 3T3, Hela, or HepG2 cell lines, both the 1.3- and the 0.9-kb fragments were able to drive CAT expression—but only when inserted in the sense orientation (Fig. 8). These results confirm that orientation-dependent promoter activity is found in the 5′ portion of the CpG island, in the immediate vicinity of transcription initiation sites (Fig. 5B).

DISCUSSION

Mouse Zfx Gene Structure and Differential Splicing

The *Zfx* gene has 11 exons distributed across 50 kb of the mouse X chromosome (Figs. 1 and 2). Exons 1–4, which contain 5′ untranslated sequences, display at least four different patterns of splicing (Fig. 3). Analysis of the *Zfa* retroposon suggests a fifth pattern of 5′ UTR splicing and the existence of one or two more 5′ UTR exons (Fig. 3). It is likely that additional modes of 5′ UTR splicing remain to be discovered. *Zfx* may contain novel 5′ UTR exons not represented in currently available cDNAs. Differential splicing of 5′ UTR exons also occurs in human *ZFX* transcripts (Schneider-Gädicke *et al.*, 1989b). We do not know whether the pattern of 5′ UTR splicing differs among tissues or developmental stages.

The Zfx protein, encoded by exons 5 through 11, consists of two large domains. The amino-terminal half, which is highly acidic and activates transcription (Mardon *et al.*, 1990), is encoded by exons 5–10. Exon 10 also encodes a putative nuclear localizing signal. The carboxy-terminal half, which consists of 13 zinc fingers, is encoded by exon 11, which also contains the entire 3′ UTR. The ATG initiator codon is located in exon 5. Intriguingly, each of the other coding exons, exons 6 through 11, begins with the second nucleotide of a codon triplet. Thus, any coding exon or combination of coding exons could be spliced out without disrupting the reading frame downstream. Differential splicing could generate a wide variety of Zfx protein isoforms, all containing the same zinc-finger domain but with differing acidic, activating domains and, potentially, different functional properties. Indeed, sequencing of mouse *Zfx* RT-PCR products (Ashworth *et al.*, 1990) suggested that exon 7 or exon 10 is sometimes spliced out. In humans, cDNA and RT-PCR analyses revealed alternative splicing that would generate two ZFX protein isoforms differing in the length of their acidic domains (Schneider-Gädicke *et al.*, 1989b). The description of the *Zfx* gene structure provided here should facilitate systematic study of differential splicing affecting coding and noncoding sequences. The information and reagents described here should also facilitate efforts to create mutations in *Zfx* by homologous recombination.

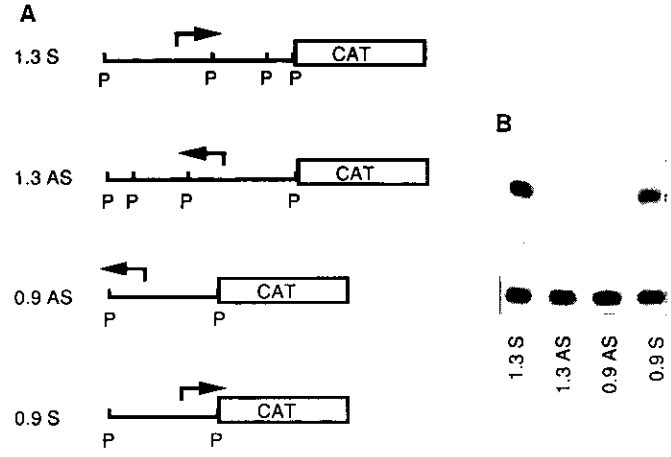

FIG. 8. Detection of *Zfx* promoter activity using CAT reporter gene constructs. (**A**) Schematic representation of four *Zfx*–CAT constructs produced by transferring 1.3- or 0.9-kb genomic *Pst*I fragments (nucleotides −871 to +490 or −871 to +10, respectively; nucleotides numbered as in Fig. 5A) to the *Pst*I site of pCATBasic (Promega) in both orientations, generating plasmids "1.3S" and "1.3AS" (sense and antisense orientations, respectively) and "0.9S" and "0.9AS." Arrows indicate transcription initiation sites defined by S1 nuclease protection (Fig. 6). (**B**) CAT activities in lysates from Hela cells transfected with the four *Zfx*–CAT constructs. Essentially identical results were obtained with 3T3 and HepG2 cells transfected with the same constructs.

A CpG Island at the 5′ End of the Zfx Gene

Exons 1 and 2 of *Zfx* fall within a 1.5-kb CpG island (Fig. 5). *Zfx* is typical of genes with CpG islands near their 5′ ends in three respects (Bird, 1987; Larsen *et al.*, 1992). First, *Zfx* transcription initiates within the CpG island (Fig. 7). Second, the *Zfx* CpG island has promoter activity (Fig. 8). Third, the *Zfx* gene is transcribed in many tissues (Mardon *et al.*, 1990). Like some housekeeping genes with CpG island promoters (Larsen *et al.*, 1992), *Zfx* appears to have no TATA or CCAAT element 5′ to its transcription initiation sites.

Evolution of the Mammalian ZFX/ZFY Gene Family from a Single Ancestral Gene

It has been hypothesized that the *ZFX* and *ZFY* genes diverged from a single common ancestral gene prior to the radiation of placental mammals (Page *et al.*, 1987). Consistent with this model, the mammalian *ZFX* and *ZFY* genes have been found to encode similar proteins (Ashworth *et al.*, 1989; Mardon and Page, 1989; Schneider-Gädicke *et al.*, 1989b; Mardon *et al.*, 1990; Palmer *et al.*, 1990). Additionally, the model predicted that mammalian *ZFX* and *ZFY* genes would have highly similar intron/exon structures. Detailed structural information is presently available only for the mouse *Zfx* gene. However, we can assess the limited structural information available for the mouse *Zfy*, human *ZFX*, and human *ZFY* genes in light of the mouse *Zfx* gene structure presented here. Like mouse *Zfx*, the mouse *Zfy* and human *ZFX* genes have at least seven coding exons, six of which encode an amino-terminal acidic domain (Schneider-Gädicke *et al.*, 1989b; Simpson and Page, 1991). Moreover, three splice junctions have been found to be conserved among the mouse *Zfx*, human *ZFY*, and human *ZFX* genes. These three conserved junctions are the exon 1 splice donor, the exon 10 donor, and the exon 11 acceptor (Fig. 2; Schneider-Gädicke *et al.*, 1989a,b; Shimmin *et al.*, 1993; our unpublished results). It remains to be determined whether all splice sites are conserved among the members of this gene family. Nonetheless, all presently available evidence indicates that the structures of the mammalian *ZFX* and *ZFY* genes are highly similar. Thus, it is likely that the extant mammalian *ZFX* and *ZFY* genes are very similar in structure to the single common ancestral gene from which they diverged at least 60 to 80 million years ago.

Elucidation of *ZFX/ZFY* gene structure will have ramifications for molecular studies of mammalian evolution. Already, comparative sequencing of small portions of *ZFX* and *ZFY* genes has allowed molecular phylogeneticists to decipher evolutionary relationships among extant mammalian species and to gain quantitative insights into the evolution of the sex chromosomes (e.g., Shimmin *et al.*, 1993). The complete description of the mouse *Zfx* gene structure provided here should accelerate such molecular phylogenetic studies.

ACKNOWLEDGMENTS

We thank Elizabeth M. Simpson, Graeme Mardon, Douglas Gray, En Li, Reinhard Faessler, and Rudolf Jaenisch for providing genomic DNA and cDNA libraries; Douglas Gray for assistance in library construction; Richard Selden and Phillip Sharp for providing plasmids; and Mark Behlke, Karin Jegalian, Rebecca Mosher, Roberto Polakiewicz, Renee Reijo, Elizabeth M. Simpson, and Andrew Zinn for comments on the manuscript. This work was supported by National Institutes of Health and Searle Scholars Program/Chicago Community Trust.

REFERENCES

Ashworth, A., Swift, S., and Affara, N. (1989). Sequence of cDNA for murine *Zfy-1*, a candidate for *Tdy*. *Nucleic Acids Res.* **17:** 2864.

Ashworth, A., Skene, B., Swift, S., and Lovell-Badge, R. (1990) *Zfa* is an expressed retroposon derived from an alternative transcript of the *Zfx* gene. *EMBO J.* **9:** 1529–1534.

Bird, A. P. (1987). CpG islands as gene markers in the vertebrate nucleus. *Trends Genet.* **3:** 342–347.

Burgoyne, P. S. (1989). Mammalian sex determination: Thumbs down for zinc finger? *Nature* **342:** 860–862.

Chang, D. D., and Sharp, P. A. (1989). Regulation by HIV Rev depends upon recognition of splice sites. *Cell* **59:** 789–795.

Dietrich, W., Katz, H., Lincoln, S. E., Shin, H. S., Friedman, J., Dracopoli, N. C., and Lander, E. S. (1992). A genetic map of the mouse suitable for typing intraspecific crosses. *Genetics* **131:** 423–447.

Ellison, J. W., Ramos, C., Yen, P. H., and Shapiro, L. J. (1992). Structure and expression of the human pseudoautosomal gene XE7. *Hum. Mol. Genet.* **1:** 691–696.

Fisher, E. M., Beer-Romero, P., Brown, L. G., Ridley, A., McNeil, J. A., Lawrence, J. B., Willard, H. F., Bieber, F. R., and Page, D. C. (1990). Homologous ribosomal protein genes on the human X and Y chromosomes: Escape from X inactivation and possible implications for Turner syndrome. *Cell* **63:** 1205–1218.

Goodfellow, P. J., Darling, S. M., Thomas, N. S., and Goodfellow, P. N. (1986). A pseudoautosomal gene in man. *Science* **234:** 740–743.

Gough, N. M., Gearing, D. P., Nicola, N. A., Baker, E., Pritchard, M., Callen, D. F., and Sutherland, G. R. (1990). Localization of the human GM-CSF receptor gene to the X–Y pseudoautosomal region. *Nature* **345:** 734–736.

Henikoff, S. (1984). Unidirectional digestion with exonuclease III creates targeted breakpoints for DNA sequencing. *Gene* **28:** 351–359.

Kay, G. F., Ashworth, A., Penny, G. D., Dunlop, M., Swift, S., Brockdorff, N., and Rastan, S. (1991). A candidate spermatogenesis gene on the mouse Y chromosome is homologous to ubiquitin-activating enzyme E1. *Nature* **354:** 486–489.

Keitges, E. A., Schorderet, D. F., and Gartler, S. M. (1987). Linkage of the steroid sulfatase gene to the *Sex-Reversed* mutation in the mouse. *Genetics* **116:** 465–468.

Koopman, P., Gubbay, J., Collignon, J., and Lovell-Badge, R. (1989). *Zfy* gene expression patterns are not compatible with a primary role in mouse sex determination. *Nature* **342:** 940–942.

Koopman, P., Ashworth, A., and Lovell-Badge, R. (1991). The *ZFY* gene family in humans and mice. *Trends Genet.* **7:** 132–136.

Larsen, F., Gundersen, G., Lopez, R., and Prydz, H. (1992). CpG islands as gene markers in the human genome. *Genomics* **13:** 1095–1107.

Lau, Y. F., and Chan, K. M. (1989). The putative testis-determining factor and related genes are expressed as discrete-sized transcripts in adult gonadal and somatic tissues. *Am. J. Hum. Genet.* **45:** 942–952.

Li, E., Bestor, T. H., and Jaenisch, R. (1992). Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. *Cell* **69:** 915–926.

Mardon, G., Mosher, R., Disteche, C. M., Nishioka, Y., McLaren, A., and Page, D. C. (1989). Duplication, deletion, and polymorphism in

the sex-determining region of the mouse Y chromosome. *Science* **243:** 78–80.

Mardon, G., and Page, D. C. (1989). The sex-determining region of the mouse Y chromosome encodes a protein with a highly acidic domain and 13 zinc fingers. *Cell* **56:** 765–770.

Mardon, G., Luoh, S.-W., Simpson, E. M., Gill, G., Brown, L. G., and Page, D. C. (1990). Mouse Zfx protein is similar to Zfy-2: Each contains an acidic activating domain and 13 zinc fingers. *Mol. Cell. Biol.* **10:** 681–688.

Mitchell, M., Simon, D., Affara, N., Ferguson-Smith, M., Avner, P., and Bishop, C. (1989). Localization of murine X and autosomal sequences homologous to the human Y located testis-determining region. *Genetics* **121:** 803–809.

Mitchell, M. J., Woods, D. R., Tucker, P. K., Opp, J. S., and Bishop, C. E. (1991). Homology of a candidate spermatogenic gene from the mouse Y chromosome to the ubiquitin-activating enzyme E1. *Nature* **354:** 483–486.

Mizusawa, S., Nishimura, S., and Seela, F. (1986). Improvement of the dideoxy chain termination method of DNA sequencing by use of deoxy-7-deazaguanosine triphosphate in place of dGTP. *Nucleic Acids Res.* **14:** 1319–1324.

Nagamine, C. M., Chan, K. M., Kozak, C. A., and Lau, Y. F. (1989). Chromosome mapping and expression of a putative testis-determining gene in mouse. *Science* **243:** 80–83.

Pabo, C. O., and Sauer, R. T. (1992). Transcription factors: Structural families and principles of DNA recognition. *Annu. Rev. Biochem.* **61:** 1053–1095.

Padgett, R. A., Grabowski, P. J., Konarska, M. M., Seiler, S., and Sharp, P. A. (1986). Splicing of messenger RNA precursors. *Annu. Rev. Biochem.* **55:** 1119–1150.

Page, D. C., Mosher, R., Simpson, E. M., Fisher, E. M., Mardon, G., Pollack, J., McGillivray, B., de la Chapelle, A., and Brown, L. G. (1987). The sex-determining region of the human Y chromosome encodes a finger protein. *Cell* **51:** 1091–1104.

Page, D. C., Disteche, C. M., Simpson, E. M., de la Chapelle, A., Andersson, M., Alitalo, T., Brown, L. G., Green, P., and Akots, G. (1990a). Chromosomal localization of ZFX—a human gene that escapes X inactivation—and its murine homologs. *Genomics* **7:** 37–46.

Page, D. C., Fisher, E. M., McGillivray, B., and Brown, L. G. (1990b). Additional deletion in sex-determining region of human Y chromosome resolves paradox of X,t(Y;22)female. *Nature* **346:** 279–281.

Palmer, M. S., Sinclair, A. H., Berta, P., Ellis, N. A., Goodfellow, P. N., Abbas, N. E., and Fellous, M. (1989). Genetic evidence that *ZFY* is not the testis-determining factor. *Nature* **342:** 937–939.

Palmer, M. S., Berta, P., Sinclair, A. H., Pym, B., and Goodfellow, P. N. (1990). Comparison of human *ZFY* and *ZFX* transcripts. *Proc. Natl. Acad. Sci. USA* **87:** 1681–1685.

Salido, E. C., Yen, P. H., Koprivnikar, K., Yu, L. C., and Shapiro, L. J. (1992). The human enamel protein gene amelogenin is expressed from both the X and the Y chromosomes. *Am. J. Hum. Genet.* **50:** 303–316.

Sambrook, J., Fritsch, E. F., and Maniatis, T. (1989). "Molecular Cloning: A Laboratory Manual," Cold Spring Harbor Laboratory Press, New York.

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

Schiebel, K., Weiss, B., Wohrle, D., and Rappold, G. (1993). A human pseudoautosomal gene, ADP/ATP translocase, escapes X-inactivation whereas a homologue on Xq is subject to X-inactivation. *Nature Gene.* **3:** 82–87.

Schneider-Gädicke, A., Beer-Romero, P., Brown, L. G., Nussbaum, R., and Page, D. C. (1989a). *ZFX* has a gene structure similar to *ZFY*, the putative human sex determinant, and escapes X inactivation. *Cell* **57:** 1247–1258.

Schneider-Gädicke, A., Beer-Romero, P., Brown, L. G., Mardon, G., Luoh, S.-W., and Page, D. C. (1989b). Putative transcription activator with alternative isoforms encoded by human *ZFX* gene. *Nature* **342:** 708–711.

Selden, R. F., Howie, K. B., Rowe, M. E., Goodman, H. M., and Moore, D. D. (1986). Human growth hormone as a reporter gene in regulation studies employing transient gene expression. *Mol. Cell. Biol.* **6:** 3173–3179.

Shimmin, L. C., Chang, B. H.-J., and Li, W.-H. (1993). Male-driven evolution of DNA sequences. *Nature* **362:** 745–747.

Simpson, E. M., and Page, D. C. (1991). An interstitial deletion in mouse Y chromosomal DNA created a transcribed *Zfy* fusion gene. *Genomics* **11:** 601–608.

Slim, R., Levilliers, J., Lüdecke, H.-J., Claussen, U., Nguyen, V. C., Gough, N. M., Horsthemke, B., and Petit, C. (1993). A human pseudoautosomal gene encodes the ANT3 ADP/ATP translocase and escapes X inactivation. *Genomics* **16:** 26–33.