Genomic Organization and Sequence of the *Gus-s^a* Allele of the Murine β -Glucuronidase Gene

BRURIA FUNKENSTEIN,[†] SUSAN L. LEARY,[‡] JILL C. STEIN,[§] AND JAMES F. CATTERALL*

The Population Council and The Rockefeller University, New York, New York 10021

Received 24 September 1987/Accepted 30 November 1987

The *Gus-s^a* allele of the mouse β -glucuronidase gene exhibits a high degree of inducibility by androgens due to its linkage with the *Gus-r^a* regulatory locus. We isolated *Gus-s^a* on a 28-kilobase pair fragment of mouse chromosome 5 and found that it contains 12 exons and 11 intervening sequences spanning 14 kilobase pairs of this genomic segment. The mRNA cap site was identified by ribonuclease protection and primer extension analyses which revealed an unusually short 5' noncoding sequence of 12 nucleotides. Proximal regulatory sequences in the 5'-flanking DNA and the complete sequence of the *Gus-s^a* mRNA transcript were also determined. Comparison of the amino acid sequence determined from the *Gus-s^a* nucleotide sequence with that of human β -glucuronidase indicated that the two human mRNA species differ due to alternate splicing of an exon homologous to exon 6 of the mouse gene.

β -Glucuronidase is a lysosomal acid hydrolase that comprises four identical subunits of molecular mass 75 kilodaltons (38). The enzyme is also found in the microsomal membranes of liver and kidney in association with a second protein, egasyn, which can bind to each subunit and anchor the complex to the membrane (37). The richest known sources of the enzyme are the rat preputial gland (39) and the mouse kidney, where the enzyme is synthesized and excreted under the control of androgens (2, 35).

The increase in β -glucuronidase activity and synthesis in response to androgen treatment are genetically regulated (36). The β -glucuronidase genetic locus, [*Gus*], is composed of the structural gene, *Gus-s*; a tightly linked *cis*-acting androgen regulator, *Gus-r*; a *cis*-acting systemic regulator, *Gus-u*; and a *trans*-acting temporal regulator, *Gus-t* (29). The two most common allelic forms of *Gus-s* and *Gus-r* are the *Gus-s^b* and *Gus-r^b* and the *Gus-s^a* and *Gus-r^a* alleles. Animals carrying the *a* alleles exhibit approximately 10-fold-greater androgen inducibility than animals carrying the *b* alleles. The *Gus-s* and *Gus-r* alleles are always inherited together and determine the [*Gus*] haplotype of the animal so that the [*Gus*]^A haplotype comprises the *Gus-s^a* and *Gus-r^a* alleles.

The identification of regulatory sequences that are responsible for the androgenic control of gene expression is a necessary step in understanding the mechanism of action of androgens. To identify the *Gus-r* androgen regulatory locus, the *Gus-s* structural gene must first be characterized. We report here the structure of the *Gus-s^a* allele of the [*Gus*] locus. The genomic organization, site of initiation of transcription, and the sequence of the mRNA derived from the *Gus-s^a* allele are described. Finally, comparison of the mouse β -glucuronidase sequence to that of humans suggests that the two species of human β -glucuronidase mRNA (28) arose by differential mRNA splicing.

MATERIALS AND METHODS

β -Glucuronidase cDNA clones. The cDNA clones pGUS7, pGUS5, pGUS48, and pGUS22 (Fig. 1) were described previously (7, 41). The cDNA from which pGUS142 was derived was synthesized from mouse (strain A/J) kidney poly(A)⁺ RNA by reverse transcription from an oligo(dT) primer by the Gubler and Hoffmann (11) modification of the method of Okayama and Berg (27). The pGUS121 cDNA clone was obtained from a mouse kidney cDNA library prepared in the λ gt11 vector and provided by Seth Alper (Massachusetts Institute of Technology, Cambridge).

Screening of genomic DNA libraries. Two libraries containing genomic DNA from [*Gus*]^A haplotype mice were screened. The first was constructed in the Charon 4A vector from spleen DNA from BALB/c mice by partial digestion with *Eco*RI (provided by R. M. Perlmutter, California Institute of Technology, Pasadena). The second library was derived from the total DNA of embryos of 12-day-old mice of the same strain by partial *Mbo*I digestion followed by insertion into the *Bam*HI site of Charon 28 (provided by J. Ravetch, Memorial Sloan-Kettering Cancer Center, New York, N.Y.).

The BALB/c spleen DNA library was screened with cDNA plasmids that recognized seven *Gus-s^a*-containing clones among the 1.7×10^6 plaques screened in three separate experiments. Clones selected for further analysis were designated GusA21 and GusA12 (Fig. 2). The BALB/c embryo DNA library was screened with cDNA and genomic subclones, resulting in the identification of four additional *Gus-s^a* clones (of a total 1.1×10^6 plaques), including GusA17 (Fig. 2). *Gus-s^a* genomic DNA fragments were routinely subcloned into pBR322 or pT3/T7 (Bethesda Research Laboratories, Gaithersburg, Md.) for further analysis.

DNA-DNA hybridization. Hybridization of ³²P-labeled DNA probes to nitrocellulose replicas of recombinant phage plaques was done at 42°C for 40 h in 50% formamide containing 4× SSC (0.6 M NaCl plus 0.06 M sodium citrate, pH 7.0), 5× Denhardt solution (0.1% [wt/vol] each of Ficoll, polyvinylpyrrolidone, and bovine serum albumin), 20 mM sodium phosphate (pH 7.0), 0.1% sodium dodecyl sulfate (SDS), and 10 μ g of poly(A), 10 μ g of poly(C), and 50 μ g of heat-denatured salmon sperm DNA per ml.

* Corresponding author.

[†] Present address: Israel Oceanographic and Limnological Research, Tel Shikmona, Haifa 31080, Israel.

[‡] Present address: Lifecodes Corp., Valhalla, NY 12595.

[§] Present address: Roche Institute for Molecular Biology, Nutley, NJ 07110.

FIG. 1. Mouse β -glucuronidase cDNA map and sequencing strategy. The physical map includes restriction endonuclease cleavage sites used in nucleotide sequence analysis of β -glucuronidase cDNA clones. The cloned cDNA inserts are shown schematically above the map. The positions of the intervening sequences (see Fig. 2 and 3) are indicated (X) on the line designated pGUS121. Below the map, the arrows represent the fragments from which the sequence was determined. Vertical lines denote the radioactively labeled ends, and the arrowheads indicate the direction and length of the deduced sequence. Solid and broken lines designate that the sequence was determined from the mRNA or cDNA equivalent strands, respectively. Restriction enzyme sites included are as follows: A, *Ava*I; B, *Bam*HI; BE, *Bst*EII; C, *Cl*aI; H, *Hind*III; HC, *Hinc*II; HF, *Hinf*I; K, *Kpn*I; P, *Pst*I; PV, *Pvu*II; RV, *Eco*RV; and S, *Sst*I.

Restriction endonuclease digests of cloned genomic DNA inserts were analyzed by Southern blot hybridization at 42°C for 18 h in a solution containing 6× SSC, 5× Denhardt solution, 0.1% SDS, 10 mM EDTA, and 25 μ g of poly(A), 50 μ g of tRNA, and 50 μ g of heat-denatured salmon sperm DNA per ml in 50% formamide. Nitrocellulose filters were washed to a final stringency of 0.2× SSC–0.1% SDS at 50°C.

DNA sequence analysis. DNA sequencing was performed by the method of Maxam and Gilbert (22) with modifications as previously reported (9).

RNAse protection assay. A 350-base-pair (bp) genomic fragment containing the 5' terminus of β -glucuronidase mRNA was subcloned into pT3/T7-18 at the *Pst*I site in the polylinker. ³²P-labeled RNA was transcribed from this template, after cleavage with *Acc*I, by using T3 RNA polymerase. The labeled transcript was purified by electrophoresis through an 8% polyacrylamide–7 M urea gel in TBE running buffer (90 mM Tris hydrochloride, 90 mM boric acid, 1 mM EDTA).

Mouse kidney poly(A)⁺ RNA (30 μ g) was mixed with 2.5 × 10⁵ cpm labeled RNA precipitated with ethanol, dried and resuspended in 30 μ l of a solution containing 0.04 M PIPES [piperazine-*N,N'*-bis(2-ethanesulfonic acid); pH 6.7], 0.4 M NaCl, and 1 mM EDTA in 80% formamide. After incubation at 85°C for 5 min, the solution was rapidly transferred to

50°C, and incubation was continued for 12 to 16 h. RNase A (Sigma type X-A; 0.3 ml of a 10- μ g/ml solution in 10 mM Tris hydrochloride [pH 7.5]–5 mM EDTA) was added, followed by incubation at 37°C for 30 min. The reaction was adjusted to 0.5% SDS, and 50 μ g proteinase K was added for an additional 15 min at 37°C. The reaction was extracted with an equal volume of phenol-chloroform (1:1) and precipitated with ethanol in the presence of 10 μ g of carrier tRNA. The protected labeled RNA was subjected to electrophoresis on a 10% polyacrylamide–7 M urea gel at 40 W constant power for 3 h.

Primer extension. Synthetic oligonucleotide primers complementary to β -glucuronidase mRNA were made using the Pharmacia Gene Assembler (Pharmacia, Inc., Piscataway, N.J.) and labeled with [γ -³²P]ATP and T4 polynucleotide kinase. After purification by gel electrophoresis, the primer (ca. 3 × 10⁷ cpm) was hybridized with 10 μ g of kidney poly(A)⁺ RNA in 0.1 M Tris (pH 8.2)–0.2 M ammonium acetate–0.38 M β -mercaptoethanol–0.1 M MgCl₂–RNasin (40 U; Promega Biotec, Madison, Wis.) at 70°C for 10 min. The hybridization mixture was allowed to cool slowly to 46°C, and the deoxynucleoside triphosphates (1 mM each) and reverse transcriptase (Life Sciences type XL; 100 U) were added. The reaction mixture was incubated for 75 min at 46°C and precipitated with ethanol. The dried pellet was

FIG. 2. Genomic organization of the *Gus-s^a* allele. The 12 exons of the *Gus-s^a* allele of the [*Gus*] complex on mouse chromosome 5 were localized by an analysis of 28 kbp of genomic DNA. The genomic segment was contained in three clones obtained from genomic libraries of strain BALB/c, which are represented in the lower part of the figure. The *Eco*RI (R) and *Hind*III (H) cleavage sites in this 28-kbp section of chromosome 5 are indicated.

suspended in 20 μ l of H₂O and diluted to 200 μ l with 50% formamide containing 20 mM PIPES (pH 6.4), 0.2% SDS, 0.4 M NaCl, and 80 μ g of tRNA per ml. Several small pieces of nitrocellulose to which β -glucuronidase cDNA was bound were added to this solution. Hybridization was carried out for 3 h at 50°C, and the hybridization solution was then replaced with 2 ml of washing buffer (0.01 M Tris hydrochloride [pH 7.6], 0.15 M NaCl, 0.001 M EDTA, 0.5% SDS) at 60°C. This wash was repeated 10 times followed by two washes with 2 ml of the same buffer without SDS at 60°C. The nitrocellulose was transferred to a 1.5-ml tube and eluted with 0.3 ml of sterile H₂O containing 3 μ g of tRNA by heating for 1 min in a boiling-water bath. The eluted sample was snap-frozen in dry ice-acetone. The sizes of the labeled hybridization-selected primer extension products were determined by electrophoresis through a 10% polyacrylamide-7 M urea gel and comparison to a known DNA sequencing ladder used as a standard.

RESULTS

Isolation and characterization of cDNA clones. We have previously reported the isolation and characterization of partial cDNA clones for mouse renal β -glucuronidase mRNA (7, 41). In addition to these clones, which were all prepared from kidney poly(A)⁺ RNA from A/J mice, we prepared pGUS142, the insert of which is a 1,451-bp *Hind*III fragment of β -glucuronidase cDNA from strain A/J, and isolated pGUS121 from a λ gt11 cDNA library prepared from mouse renal poly(A)⁺ RNA of strain CD-1. None of these cDNA clones was full-length, even though it could be demonstrated that full-length double-stranded cDNA was synthesized (6). Figure 1 shows a composite cDNA map derived from the overlapping cDNA inserts and Southern mapping of full-length cDNA. The strategy for determining the nucleotide sequence is also shown. The 5' and 3' ends of the cloned cDNA sequence were determined from the longest cDNA insert, pGUS121, which was obtained from the cDNA library of strain CD-1. This strain of mice is outbred, so we could not be certain of its [*Gus*] haplotype. Since most of our data were derived from [*Gus*]^A strains, genomic clones were isolated from a strain (BALB/c) with this haplotype. Analysis of *Gus-s^a* genomic clones allowed precise mapping of *Gus-s^a*, permitted the unambiguous determination of the nucleotide sequence of the allele-specific transcript, and aided in the identification of the mRNA cap site by nuclease mapping.

Genomic organization of the *Gus-s^a* structural allele. Initially, high-molecular-weight liver DNA from [*Gus*]^A mice was digested with *Eco*RI and analyzed by Southern blot hybridization to cloned cDNA probes. Four fragments of 6.0, 2.15, 2.85, and 9.4 kbp were detected (data not shown). To completely characterize the *Gus-s^a* allele, a BALB/c genomic library prepared by partial *Eco*RI cleavage was screened by using the various cDNAs described in Fig. 1. Two types of clones were identified in the library, one of which contained *Eco*RI fragments of 2.15, 2.85, and 9.4 kbp (represented by GusA21 in Fig. 2), while the other contained 6.0- and 8.3-kbp *Eco*RI fragments as in GusA12 in Fig. 2. The order of the fragments was determined to be 8.3, 6.0, 2.15, 2.85, and 9.4 kbp (5' to 3') by restriction mapping and Southern blot hybridization with cDNA fragments specific for the 5' and 3' ends of the gene. The 8.3-kbp *Eco*RI fragment of GusA12 did not hybridize to any of the cloned cDNA fragments [or renal poly(A)⁺ RNA], suggesting that the 5' terminus of the gene was in the 6.0-kbp fragment. To

prove that the 6.0- and 2.15-kbp *Eco*RI fragments, which were derived from separate cloned DNA fragments, were indeed contiguous in genomic DNA, a second library, prepared by partial *Mbo*I cleavage, was screened and an overlapping clone (GusA17) was isolated. Characterization of this genomic segment eliminated the possibility of an additional *Eco*RI fragment between the 6.0- and 2.15-kbp fragments of GusA12 and GusA21, respectively (Fig. 2).

The physical map of 28 kbp of genomic DNA represented in the GusA clones (Fig. 2) shows the organization of the 12 exons and 11 introns of *Gus-s^a* on mouse chromosome 5. The exon-intron structure of the gene is shown in greater detail in Fig. 3, which also shows the strategy for sequencing the *Gus-s^a* exons. Exon sequences were identified by comparison with known cDNA sequences using the SEQFIT program of Staden (34). Adjacent intron sequences were also determined in order to place each exon precisely and to determine the primary structure of the sites for splicing of the primary transcript. The splicing junctions (Table 1) in every case conform to the consensus sequences determined for the splicing of eucaryotic mRNA precursors (1, 5, 8, 24).

Characterization of the *Gus-s^a* transcription unit. The nucleotide sequence of the *Gus-s^a* allele transcript (Fig. 4) was determined primarily from the sequences of the cDNA clones and was confirmed by the exon sequences derived from the genomic clones. No sequence differences were found between the A/J cDNA sequences and the exon sequences of *Gus-s^a*. However, in two regions of the sequence (nucleotides 1 to 581 and 2306 to 2454) the *Gus-s^a* sequence was derived mainly from the genomic clones. Except for nucleotides 1 to 43, which were not present in any cDNA clone, these two regions of the cDNA are represented only in pGUS121, which appears to have been derived from another allele of the *Gus-s* gene (unpublished results). Of 1,701 nucleotides of the pGUS121 sequence that were determined, seven differences from the *Gus-s^a* sequence were documented. These sequence variations may represent the divergence of *Gus-s^a* from the wild-type allele *Gus-s^b*, but this cannot be substantiated by the present data.

The amino acid sequence deduced from the large open reading frame in the composite cDNA sequences revealed a high percentage of sequence identity with the previously published rat β -glucuronidase amino acid sequence (26). Examination of genomic sequences adjacent to the cDNA sequence led to the identification of the N-terminal codons of the open reading frame. The N-terminal amino acid sequence of the enzyme secreted from the rat preputial gland (10) matched amino acids 23 to 31 of the mouse protein sequence (8 identical amino acids of the 9 determined). This indicated that the first 22 amino acids of the mouse sequence comprised the signal peptide (Fig. 4).

To determine whether the transcription initiation site was contained within the sequences adjacent to the protein-coding region, various restriction fragments from the 6.0-kb *Eco*RI fragment of the GusA12 genomic clone were labeled and hybridized to β -glucuronidase mRNA by Northern (RNA) blot hybridization. Only restriction fragments containing sequences proximal to the open reading frame hybridized to β -glucuronidase mRNA. Synthetic oligonucleotides representing the 5'-most sequences of the cDNA and sequences progressively 5' from this sequence (Fig. 5) were labeled and hybridized to β -glucuronidase mRNA. Oligomers 1 and 7 hybridized to β -glucuronidase mRNA, whereas oligomers 2 to 6 did not. These experiments indicated that the 350-bp *Pst*I fragment of GusA12 (Fig. 5) contained the 5' noncoding sequence of the mRNA. This fragment was

FIG. 3. Fine mapping of the exons of *Gus-s^a*. The four *EcoRI* fragments that contain the *Gus-s^a* allele are presented schematically and designated by their lengths (in kilobase pairs) at the left. Solid boxes represent the exons. Uninterrupted lines represent actual lengths of introns and interrupted lines indicate introns or flanking DNA, the lengths of which have been proportionally reduced to conserve space but maintain the relative distances between exons. The restriction endonuclease sites shown are only those used in determining the nucleotide sequences of the exons and exon-intron junctions. Arrows start at the labeled restriction site and show the length and direction of the sequences determined. Restriction endonuclease sites are abbreviated as described in the legends to Fig. 1 and 2 except that AV is *AvaII* and HP is *HpaII*.

subcloned into the pT3/T7 plasmid, and labeled antisense RNA was prepared with the T3 RNA polymerase. When this RNA was hybridized to β -glucuronidase mRNA and the hybrids were treated with RNase, a major protected fragment of approximately 185 nucleotides was produced (Fig. 6A). From this data, the 5' terminus of β -glucuronidase mRNA was estimated to be 13 nucleotides 5' of the initiator methionine codon.

Since this result could not distinguish between the mRNA cap site and the presence of an additional intervening se-

quence, primer extension analysis was performed using two separate oligonucleotide primers. The primers used contained sequences complementary to nucleotides 20 to 39 (*Gus-1*) and 90 to 109 (*Gus-7*) (Fig. 5). Each was used to prime synthesis of cDNA from mouse renal poly(A)⁺ RNA. Primer extension products of 39 and 108 nucleotides were obtained from the *Gus-1* and *Gus-7* primers, respectively (Fig. 6B). The difference in the length of these cDNAs was identical to the distance between the primers, indicating that each synthesis terminated at the same position, 12 nucleo-

TABLE 1. Nucleotide sequences at the exon-intron junctions

Exon no.	Exon size (bp)	Sequence at exon-intron junction		Intron length (kbp)
		5' Splice site	3' Splice site	
1	222	C G G G A G g t g c g c	c c c t a g T C G G G C	1.48
2	186	G T T G T G g t c a g t	c t g c a g T G G G T G	0.21
3	182	C T C C A T g t g a g c	c c c t a g G T A T C C	0.28
4	143	A C A T C G g t g a g c	c c a c a g G G C T G G	0.65
5	188	T T G G A G g t a a c g	t c c t a g G T G A A G	0.76
6	144	T C A G A T g t g a g t	t c c t a g A T C C G A	0.20
7	179	G C T A C C g t g a g t	c c a c a g T C A G A G	0.10
8	147	C T T T A A g t g a g t	g g a c a g G A C G C T	2.15
9	85	C T G G G G g t g t a g	t c t t a g G C C C C G	1.85
10	177	C A C G A G g t a a g c	a t c c a g G A C C C G	3.30
11	136	A C C A G T g t g a g t	t t t t a g C A C C A C	0.28
12	665

presence of a TATA box-like sequence (TAAAAA) sequence 28 nucleotides 5' of the site (Fig. 5). Such TATA consensus sequences are typically located approximately 30 nucleotides upstream of the transcription initiation sites of eucaryotic genes (25).

The 3' terminus of the *Gus-s^a* gene contains an A₁₄ sequence which is preceded by an AATAAA polyadenyly-

```

-170      -160      -150      -140      -130      -120
CTGCAGAGTC CAGCCTTGGG GCCTCTAGAT AGCCTTGAGC CAGCAGCTGT ACAAGCGCTT

-110      -100      -90      -80      -70      -60
GCCCTCAACC AAGTTGGCGC AGGCGGCTTC TCTCGAGAAC ATGTGACTCT TCCAGGTCCC
  2          1          3          4          5          6

-50      -40      -30      -20      -10      1
GCCTCTGCCC TCCCAGAGAG GAGGCTAAAA AGCCAGCCG CTGGTCGGTG GGCAGCCTGG
  5          6          7          8          9          10

10      20      30      40      50      60
CCAGTATGTC CCTAAAAATGG AGTGCCTGTT GGTTCGCGCT GGGCCAGCTG CTGTGCAGCT
          1          2          3          4          5          6

70      80      90      100      110      120
GCGCGCTGGC TCTGAAGGGC GGGATGCTGT TCCCGAAGGA GAGCCCGTCG CGGGAGCTCA
          7          8          9          10          11          12

130      140      150      160      170      180
AGGCGCTGGA CGGACTGTGG CACTTCCGCG CCGACCTCTC GAACAACCGG CTGCAG

```

FIG. 5. Sequence of the 5'-flanking DNA of the *Gus-s^a* allele. The sequence of the 350-bp *Pst*I fragment of GusA12 that contains the mRNA cap site is shown. Underlined sequences 1 to 7 represent the synthetic oligonucleotides used for β -glucuronidase mRNA hybridization mapping. Oligomers 1 and 7 were also used as primers in the determination of the 5' end of the mRNA (Fig. 6). The overlined sequence indicates the AT-rich TATA box consensus sequence at -28 from the mRNA cap site. The numbering of the sequence was assigned on the basis of primer extension determination of the mRNA cap site (Fig. 6).

lation signal 15 nucleotides upstream. The actual site of poly(A) addition to the mRNA occurs at or within this A_{14} sequence, since oligonucleotides complementary to the mRNA strand sequence of the genomic DNA immediately 3' of the A_{14} sequence do not hybridize to β -glucuronidase mRNA. The complete *Gus-s^a* transcription unit is therefore 14 kbp long and contains 12 exons and 11 intervening sequences that are processed into an mRNA that is 2,454 nucleotides long. These data agree well with our previous estimates of the length of β -glucuronidase mRNA (7), if one assumes a poly(A) tail of 100 to 150 nucleotides.

The mouse β -glucuronidase sequence. Mouse β -glucuronidase mRNA encodes a 648-amino-acid precursor from which the 626-amino-acid lysosomal enzyme is derived. Mouse and rat β -glucuronidase are identical in length and share 88.2% sequence identity. Like the rat enzyme (26), mouse β -glucuronidase shows amino acid sequence homology to *Escherichia coli* β -galactosidase. However, no significant homology was detected with any other protein when murine β -glucuronidase was compared with protein sequences from two protein sequence data bases (Protein Identification Resource, Protein Sequence Database National Biomedical Research Foundation, Washington, D.C., release 12; and GenBank, Bolt, Beranek and Newman, Cambridge, Mass., release 50, May 1987) using the FASTP program of Lipman and Pearson (16). Mouse and human β -glucuronidase (28) are 75.3% identical, but the human enzyme precursor (651 amino acids) differs in length from the rodent precursors (648 residues) (Fig. 7). This variation in length is due to sequence divergence in three regions of the sequences. First, the human sequence is lengthened by one amino acid relative to the rodent sequences by an apparent substitution of four amino acids (P, S, R, and L) for the three amino acids at positions 163 to 165 (T, T, and C; Fig. 7). Second, the C terminus of the human protein is truncated by a single amino acid relative to the rodent sequences. Third, there is a three-amino-acid insertion in the human sequence relative to the rodent sequences between positions 306 and 307 in the mouse sequence. The latter variation occurs in a 51-amino-acid sequence homologous to that derived from mouse exon 6 (Fig. 7). These 51 amino

FIG. 6. Identification of the 5' terminus of β -glucuronidase mRNA by ribonuclease protection and primer extension. (A) A 32 P-labeled transcript of the 350-bp *Pst*I genomic fragment of GusA12 (Fig. 5) was hybridized to kidney poly(A)⁺ RNA and digested with RNase A. The protected products were analyzed by electrophoresis on a 10% polyacrylamide gel containing 7 M urea. The size of the major protected fragment was determined from a DNA sequencing ladder of a known sequence and run on the same gel. (B) Oligomers Gus-1 (lane 1) and Gus-7 (lane 2) were used to prime cDNA synthesis from kidney poly(A)⁺ RNA. Primer extension products were separated, and their lengths were determined by gel electrophoresis as described above.

FIG. 7. Comparison of the amino acid sequences of mouse (upper) and human (lower) β -glucuronidase. Output from the FASTP2 program of Lipman and Pearson (16) comparing the deduced mouse β -glucuronidase amino acid sequence with the human sequence (28) is shown. An asterisk (*) designates potential sites of N-linked glycosylation in the two sequences. ∇ , Boundaries of exon 6 in the mouse gene which match those of a fragment missing in one species of human mRNA (28); X, limits of the longest region of uninterrupted sequence similarity between the two species.

acids are encoded in the 153 nucleotides missing in a variant of human β -glucuronidase mRNA (28).

Three of the four consensus sequences for N-linked oligosaccharide addition sites at positions 172, 416, and 627 of the mouse sequence are conserved between the rodent and human sequences. The fourth site, at position 591 in the mouse, was lost in the human due to a conservative amino acid substitution of glutamic acid for asparagine. However, the total number of these sites is conserved since the human sequence contains a fourth site at position 272 (Fig. 7).

DISCUSSION

The mouse renal β -glucuronidase gene complex, [*Gus*], consists of four independent loci as shown by classical genetic analysis (29). Besides the structural gene for the enzyme (*Gus-s*), three regulatory loci have been identified: the *cis*-acting androgen-response regulator, *Gus-r* (17, 36); a *cis*-acting systemic regulator, *Gus-u* (18); and a *trans*-acting temporal regulator, *Gus-t* (13, 18, 23). The best characterized of the regulatory loci is *Gus-r*, which regulates the steady-state accumulation of β -glucuronidase mRNA in re-

sponse to androgen treatment (30, 31, 40). *Gus-r* is tightly linked to *Gus-s* (36), and its regulatory activity is specific for β -glucuronidase enzyme activity (36) and mRNA accumulation (40). An important step in the characterization of the structure and molecular mechanism of action of *Gus-r* is the determination of the structure of the *Gus-s* gene on which it acts. Three common alleles of the *Gus-s* gene have been characterized, two of which code for enzymes that can be separated electrophoretically. *Gus-s^b* codes for the wild-type β -glucuronidase, and *Gus-s^a* produces a faster-migrating electrophoretic variant (17). A third allele encoding a heat-stable variant of the enzyme, *Gus-s^h*, has also been described (19). More recently, two new structural alleles, *Gus-s^{cs}* and *Gus-s^{cl}* were reported (32). The *Gus-s^a* allele is linked to the *Gus-r^a* regulatory allele which confers a high degree of androgen inducibility (3, 29). Our present data reveal the genomic organization of the *Gus-s^a* structural allele, including precise location of the mRNA cap site and the complete nucleotide sequence of its mRNA as determined from the sequences of cloned cDNA and the exons in genomic clones from [*Gus*]^A haplotype strains A/J and BALB/c, respectively.

Gus-s^a consists of 12 exons dispersed among 11 intervening sequences covering approximately 14 kbp of mouse chromosome 5. The positions of exons 4 to 12 were determined by comparing *Gus-s^a* sequences with the sequence of [*Gus*]^A haplotype cDNA (pGUS5, -7, -22, -48, and -142 in Fig. 1), and exons 1 to 3 were placed by using the pGUS121 cDNA sequence. The pGUS121 cDNA was derived from a *Gus-s* allele other than *Gus-s^a*, but its sequence is 99.6% identical (seven differences in 1,701 nucleotides of pGUS121 have been determined to date) to [*Gus*]^A cDNA. The mRNA product of *Gus-s^a* is 2,454 bases long, a finding which agrees with our previous measurements of the size of β -glucuronidase mRNA (7), assuming the presence of a poly(A) tail of approximately 150 bases. The mRNA consists of a 12-nucleotide 5' noncoding region, a 1,947-nucleotide peptide-coding region of which 22 codons specify the signal peptide, and a 495-nucleotide 3' noncoding region.

The sequences of β -glucuronidase cDNA from a rat (26) and a human (28) were reported previously. In each case, the complete amino acid sequence was derived but no attempt was made to characterize the 5' or 3' terminus of β -glucuronidase mRNA. There were six nucleotides in the rat cDNA 5' noncoding sequence, whereas the human 5' noncoding sequence is at least 26 nucleotides long. Our data, derived from RNase protection and primer extension analyses, indicate that this region is short (12 nucleotides) in mouse β -glucuronidase mRNA, as are the 5' noncoding sequences of several pancreatic secretory proteins (4, 12, 20, 21, 33), which are 11 to 21 nucleotides long. The restriction map at the 5' end of mouse β -glucuronidase cDNA (Fig. 1) differs from that in a preliminary report from this laboratory (6); the two 5'-most sites in the original map are not present in cloned cDNA or genomic sequences. The original data were drawn from Southern mapping of double-stranded cDNA, a procedure subject to experimental error in the size determination of the full-length cDNA and the restriction fragments.

The 3' noncoding region of β -glucuronidase mRNA contains a single AATAAA polyadenylation signal. However, none of the cDNA clones analyzed included a poly(A) segment. In the genomic DNA, the AATAAA sequence is followed by an A₁₄ sequence at a 16 position nucleotides downstream. A synthetic oligomer containing the 21 nucleotides adjacent to the A₁₄ sequence in *Gus-s^a* does not

hybridize to β -glucuronidase mRNA, suggesting that polyadenylation occurs at or within the encoded A_{14} sequence. On the other hand, the rat β -glucuronidase cDNA contains an A_{18} sequence, preceded by a polyadenylation signal, that is apparently encoded in the mRNA (26). The 17-nucleotide sequence downstream of the rat A_{18} site may be an artifact due to the ligation of unrelated DNA during cDNA cloning. The sequence in this region of the rat cDNA is not similar to the mouse genomic sequence 3' to the A_{14} , whereas the 3' noncoding regions of the two mRNAs preceding the poly(A) stretches are 72% identical. However, the possibility of a splicing event to form the 3' end of the mRNA cannot be excluded.

Comparison of the mouse β -glucuronidase sequence with the sequence of the human enzyme revealed that the human peptide was three amino acids longer than the rodent sequences. The major reason for the difference is a sequence variation causing a three-amino-acid gap in the optimal alignment between positions 306 and 307 in the mouse. Interestingly, this gap occurs at the beginning of a 153-nucleotide segment that is lacking in one of the two human β -glucuronidase mRNA species (28). The boundaries of this sequence are identical to those of exon 6 in the mouse gene. This suggests that the two human mRNAs are derived by alternative splicing. Furthermore, the sequence variation adjacent to the 5' splice site may adversely affect the splice site and therefore explain the failure to splice the homologous exon in the human gene resulting in a "skipped" exon (15).

Sites involved in the regulation of gene expression by several steroid hormones have been characterized. However, similar sites have not been identified for androgen-regulated genes. Sequence similarities have been described between androgen-regulated genes in the rat prostate and seminal vesicle (14), but a broader search did not reveal consensus sequences for androgen regulatory elements (42). Knowledge of the *Gus-s^a* gene structure will allow the identification and analysis of upstream sequences of other alleles of *Gus-s*. Comparison of these sequences may lead to the identification of androgen regulatory elements or of candidates for the *Gus-r* regulatory locus. When combined with a functional analysis of potential regulatory sites by nuclear protein binding or by expression of gene constructs in transfected cells, these data will provide insights into the molecular mechanism of androgen action.

ACKNOWLEDGMENTS

We thank R. M. Perlmutter and J. Ravetch for providing aliquots of mouse genomic libraries, and we thank Seth Alper for making the CD-1 mouse kidney cDNA library available to us. The assistance of Jean Schweis in the preparation of the manuscript is also gratefully acknowledged.

This work was supported by Public Health Service grant HD-13541 from the National Institutes of Health and grants from the Surdna Foundation, Inc., and the Andrew W. Mellon Foundation.

LITERATURE CITED

1. Aebi, M., H. Hornig, R. A. Padgett, J. Reiser, and C. Weissmann. 1986. Sequence requirements for splicing of higher eukaryotic nuclear pre-mRNA. *Cell* 47:555-565.
2. Bardin, C. W., T. R. Brown, N. C. Mills, C. Gupta, and L. P. Bullock. 1978. The regulation of the β -glucuronidase gene by androgens and progestins. *Biol. Reprod.* 18:74-83.
3. Bardin, C. W., and J. F. Catterall. 1981. Testosterone, a major determinant of extragenital sexual dimorphism. *Science* 211:1285-1294.
4. Bell, G. I., C. Quinto, M. Quiroga, P. Valenzuela, C. S. Craik, and W. J. Rutter. 1984. Isolation and sequence of a rat chymotrypsin B gene. *J. Biol. Chem.* 259:14265-14270.
5. Breathnach, R., C. Benoist, K. O'Hare, F. Gannon, and P. Chambon. 1978. Ovalbumin gene: evidence for a leader sequence in mRNA and DNA sequences at the exon-intron boundaries. *Proc. Natl. Acad. Sci. USA* 75:4853-4857.
6. Catterall, J. F., K. K. Kontula, C. S. Watson, P. J. Seppänen, B. Funkenstein, E. Melanitou, N. J. Hickok, C. W. Bardin, and O. A. Jänne. 1986. Regulation of gene expression by androgens in murine kidney. *Rec. Prog. Horm. Res.* 42:71-109.
7. Catterall, J. F., and S. L. Leary. 1983. Detection of early changes in androgen-induced mouse renal β -glucuronidase mRNA using cloned cDNA. *Biochemistry* 22:6049-6053.
8. Catterall, J. F., B. W. O'Malley, M. A. Robertson, R. Staden, Y. Tanaka, and G. G. Brownlee. 1978. Nucleotide sequence homology at 12 intron-exon junctions in the chick ovalbumin gene. *Nature (London)* 257:510-513.
9. Catterall, J. F., J. P. Stein, P. Kristo, A. R. Means, and B. W. O'Malley. 1980. Primary sequence of chicken ovomucoid mRNA as determined from cloned complementary DNA. *J. Cell Biol.* 87:480-487.
10. Erickson, A. H., G. E. Conner, and G. Blobel. 1984. Biosynthesis of lysosomal enzymes, p. 235-249. In J. A. Barranger and R. O. Brady (ed.), *Molecular basis of lysosomal storage diseases*. Academic Press, Inc., Orlando, Fla.
11. Gubler, U., and B. J. Hoffman. 1983. A simple and very efficient method for generating cDNA libraries. *Gene* 25:263-269.
12. Han, J. H., C. Stratowa, and W. J. Rutter. 1987. Isolation of full-length putative rat lysophospholipase cDNA using improved methods for mRNA isolation and cDNA cloning. *Biochemistry* 26:1617-1625.
13. Herrup, K., and R. J. Mullen. 1977. Biochemical and genetic factors in the heat inactivation of murine β -glucuronidase. *Biochem. Genet.* 15:641-653.
14. Kandala, J. C., M. K. Kistler, and W. S. Kistler. 1985. Androgen-regulated genes from prostate and seminal vesicle share upstream sequence homologies. *Biochem. Biophys. Res. Commun.* 126:948-952.
15. Leff, S. E., M. G. Rosenfeld, and R. M. Evans. 1986. Complex transcriptional units: diversity in gene expression by alternative RNA processing. *Annu. Rev. Biochem.* 55:1091-1117.
16. Lipman, D. J., and W. R. Pearson. 1985. Rapid and sensitive protein similarity searches. *Science* 227:1435-1441.
17. Lusis, A. J., V. M. Chapman, C. Herbstman, and K. Paigen. 1980. Quantitation of *cis versus trans* regulation of mouse β -glucuronidase. *J. Biol. Chem.* 255:8959-8962.
18. Lusis, A. J., V. M. Chapman, R. W. Wagenstein, and K. Paigen. 1983. *trans*-Acting temporal locus within the β -glucuronidase gene complex. *Proc. Natl. Acad. Sci. USA* 80:4398-4402.
19. Lusis, A. J., and K. Paigen. 1978. The large-scale isolation of mouse β -glucuronidase and comparison of allozymes. *J. Biol. Chem.* 253:7336-7345.
20. MacDonald, R. J., S. J. Stary, and G. Swift. 1982. Two similar but nonallelic rat pancreatic trypsinogens. *J. Biol. Chem.* 257:9724-9732.
21. MacDonald, R. J., G. Swift, C. Quinto, W. Swain, R. Pictet, W. Nikovits, and W. J. Rutter. 1982. Primary structure of two distinct rat pancreatic preproelastases determined by sequence analysis of the complete cloned mRNA sequences. *Biochemistry* 21:1453-1463.
22. Maxam, A. M., and W. Gilbert. 1977. A new method for sequencing DNA. *Proc. Natl. Acad. Sci. USA* 74:560-564.
23. Meredith, S. A., and R. E. Ganschow. 1978. Apparent *trans* control of murine β -glucuronidase synthesis by a temporal genetic element. *Genetics* 90:725-734.
24. Mount, S. M. 1982. A catalogue of splice junction sequences. *Nucleic Acids Res.* 10:459-472.
25. Nevins, J. R. 1983. The pathway of eukaryotic mRNA formation. *Annu. Rev. Biochem.* 52:441-466.
26. Nishimura, Y., M. G. Rosenfeld, G. Kreibich, U. Gubler, D. D. Sabatini, M. Adesnik, and R. Andy. 1986. Nucleotide sequence of rat preputial gland β -glucuronidase cDNA and *in vitro*

- insertion of its encoded polypeptide into microsomal membranes. *Proc. Natl. Acad. Sci. USA* **83**:7292-7296.
27. Okayama, H., and P. Berg. 1982. High-efficiency cloning of full-length cDNA. *Mol. Cell. Biol.* **2**:161-170.
 28. Oshima, A., J. W. Kyle, R. D. Miller, J. W. Hoffmann, P. P. Powell, J. H. Grubb, W. S. Sly, M. Tropak, K. S. Guise, and R. A. Gravel. 1987. Cloning, sequencing, and expression of cDNA for human β -glucuronidase. *Proc. Natl. Acad. Sci. USA* **84**:685-689.
 29. Paigen, K. 1979. Acid hydrolases as models of genetic control. *Annu. Rev. Genet.* **13**:417-466.
 30. Paigen, K., C. Labarca, and G. Watson. 1979. A regulatory locus for mouse β -glucuronidase induction, *gur*, controls messenger RNA activity. *Science* **203**:554-556.
 31. Palmer, R., P. M. Gallagher, W. L. Boyko, and R. E. Ganschow. 1983. Genetic control of levels of murine kidney glucuronidase mRNA in response to androgen. *Proc. Natl. Acad. Sci. USA* **80**:7596-7600.
 32. Pfister, K., V. Chapman, G. Watson, and K. Paigen. 1985. Genetic variation for enzyme structure and systemic regulation in two new haplotypes of the β -glucuronidase gene of *Mus musculus castaneus*. *J. Biol. Chem.* **260**:11588-11594.
 33. Quinto, C., M. Quiroga, W. F. Swain, W. C. Nikovits, D. N. Standring, R. L. Pictet, P. Valenzuela, and W. J. Rutter. 1982. Rat preprocarboxypeptidase: cDNA sequence and preliminary characterization of the gene. *Proc. Natl. Acad. Sci. USA* **79**:31-35.
 34. Staden, R. 1977. Sequence data handling by computer. *Nucleic Acids Res.* **4**:4037-4051.
 35. Swank, R. T., K. Paigen, R. Davey, V. Chapman, C. Labarca, G. Watson, R. Ganschow, E. J. Brandt, and E. Novak. 1978. Genetic regulation of mammalian glucuronidase. *Rec. Prog. Horm. Res.* **34**:401-436.
 36. Swank, R. T., K. Paigen, and R. E. Ganschow. 1973. Genetic control of glucuronidase induction in mice. *J. Mol. Biol.* **81**:225-243.
 37. Tomino, S., and K. Paigen. 1975. Egasyn, a protein complexed with microsomal β -glucuronidase. *J. Biol. Chem.* **250**:1146-1148.
 38. Tomino, S., K. Paigen, D. R. P. Tulsiani, and O. Touster. 1975. Purification and chemical properties of mouse liver (L) β -glucuronidase. *J. Biol. Chem.* **250**:8503-8509.
 39. Tulsiani, D. R. P., R. K. Keller, and O. Touster. 1975. The preparation and chemical composition of multiple forms of β -glucuronidase from the female rat preputial gland. *J. Biol. Chem.* **250**:4770-4776.
 40. Watson, C. S., and J. F. Catterall. 1986. Genetic regulation of androgen-induced accumulation of mouse renal β -glucuronidase mRNA. *Endocrinology* **118**:1081-1086.
 41. Watson, C. S., D. Salomon, and J. F. Catterall. 1984. Structure and expression of androgen-regulated genes in mouse kidney. *Ann. N.Y. Acad. Sci.* **438**:101-114.
 42. Williams, L., C. McDonald, and S. Higgins. 1985. Sequence organization of rat seminal vesicle F gene: location of transcriptional start point and sequence comparison with six other androgen-related genes. *Nucleic Acids Res.* **13**:659-672.