Fine structure of the human galactokinase GALK1 gene.

D J Bergsma, Y Ai, W R Skach, et al.

Genome Res. 1996 6: 980-985
Access the most recent version at doi:10.1101/gr.6.10.980

LETTER

Fine Structure of the Human Galactokinase *GALK1* Gene

Derk J. Bergsma,[1] Yunjun Ai,[3] William R. Skach,[4] Kristin Nesburn,[2] Elizabeth Anoia,[2] Stephanie Van Horn,[1] and Dwight Stambolian[2,3,5]

[1]Department of Molecular Genetics, SmithKline Beecham Pharmaceutical, King of Prussia, Pennsylvania 19046; Departments of [2]Ophthalmology, [3]Genetics, and [4]Molecular and Cellular Engineering, School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104

Defects in the human *GALK1* gene result in galactokinase deficiency and cataract formation. We have isolated this gene and established its structural organization. The gene contains 8 exons and spans ~7.3 kb of genomic DNA. The *GALK1* promoter was localized and found to have many features in common with other housekeeping genes, including high GC content, several copies of the binding site for the Sp1 transcription factor, and the absence of TATA-box and CCAAT-box motifs typically present in eukaryotic Pol II promoters. Analysis by 5′-RACE PCR indicates that the *GALK1* mRNA is heterogeneous at the 5′ terminus, with transcription sites occurring at many locations between 21 and 61 bp upstream of the ATG start site of the coding region. In vitro translation experiments of the *GALK1* cDNA indicate that the protein is cytosolic and not associated with the endoplasmic reticulum membrane.

Genetic defects of galactose metabolism constitute a class of genetic disorders termed galactosemia. One form of galactosemia in humans is caused by an enzyme deficiency of galactokinase (Segal 1989). Individuals with homozygous galactokinase deficiency are symptomatic in the early infantile period with galactosemia, galactosuria, increased galactitol levels, cataracts, and in a few cases, mental retardation (Segal et al. 1979). Heterozygotes for galactokinase deficiency are prone to presenile cataracts at ~20–50 years of age (Stambolian et al. 1986).

The isolation and characterization of the galactokinase gene has been completed in *Escherichia coli, Saccharomyces,* and *Streptomyces lividans* (Citron and Donelson 1984; Debouck et al. 1985; Adams et al. 1988). Recently the cDNA encoding human galactokinase, termed GALK1, was cloned, characterized functionally, and mapped to chromosome 17q24 (Stambolian et al. 1995). Analysis of the sequence predicted a 1.35-kb mRNA that results in a polypeptide chain of 392 amino acids. This was confirmed by detection of a single 1.35-kb mRNA on a Northern blot. The 3′ end of the mRNA was well defined, but the 5′ end was not defined completely. Analysis of the predicted protein sequence showed a conserved galactokinase signature sequence as well as two different ATP-binding motifs that are conserved among all the galactokinases. Additionally, two distinct mutations were identified within *GALK1* gene-coding sequences of two unrelated families with galactokinase deficiency and cataracts.

Definition of the *GALK1* gene structure will improve our understanding of the protein function by allowing experimental manipulation of the gene in vitro and in vivo through gene knockouts. It will also facilitate further genetic studies of patients with galactokinase deficiency. In this study we have determined the complete exon–intron structure of the human *GALK1* gene and examined expression and processing of the encoded protein.

[5]Corresponding author.
E-MAIL stamboli@mail.med.upenn.edu; FAX (215) 573-8590.

RESULTS AND DISCUSSION

Isolation of the Human *GALK1* Gene

Six phage clones were isolated from a human placental genomic library by screening with the full-length *GALK1* cDNA and characterized by restriction mapping followed by Southern blot analysis (Sambrook et al. 1989). Clone GK17 contained a 14-kb insert that, following primer sequence walking, was found to contain the entire structure of the *GALK1* gene. Comparison of the ge-

6:980–985 ©1996 by Cold Spring Harbor Laboratory Press ISSN 1054-9803/96 $5.00

nomic coding sequence with the published cDNA sequence (Stambolian et al. 1995) revealed no sequence differences.

Structure and Organization of the *GALK1* Gene

The human *GALK1* gene is ~7.3 kb in length and consists of 8 exons interrupted by 7 introns (Fig. 1). Exon sizes are relatively small and range in size from 120 to 190 bp, whereas intron sizes range in size from 82 to 4107 bp (Table 1). All intron types are about equally represented in the *GALK1* gene. Type 0 (uninterrupted codon) is found twice, type 1 (splice site after the first nucleotide) is present three times, and type 2 (splice site after the second nucleotide) comprises the remaining two sites (Table 1). All of the exon–intron boundaries conform to the AG/GT rule (Breathnach et al. 1978; Shapiro and Senapathy 1987).

It has been suggested that genes are assembled, through intron-mediated recombination, from exons that code for functional or structural units of proteins (Doolittle 1978; Gilbert 1978). In the case of the *GALK1* gene, the exons seem to correspond to the predicted functional/structural domains of the protein product (Table 1). Exon 1 contains the galactokinase signature sequence, which is conserved among all known galactokinases and may be required for functional activity (Stambolian et al. 1995). Exons 3 and 7 contain separate ATP-binding motifs (Debouck et al. 1985; Tsay and Robinson 1991) that are likewise conserved among galactokinases (Ai et al. 1995).

The last exon of the *GALK1* gene, exon 8, was found to contain 189 bp, calculated from the last intron–exon boundary to the published polyadenylation site (Stambolian et al. 1995). Beginning with the most upstream transcription initiation site as position +1 (see below), this adds up to a maximum size of 1358 nucleotides for the mRNA [not including the poly(A) tail], a value that is compatible with the results of Northern analysis, which showed a single mRNA of ~1.35 kb (Stambolian et al. 1995). Exon 8 is composed of a 72-bp region encoding the carboxyl terminus of the galactokinase protein, followed by a 117-bp 3′-

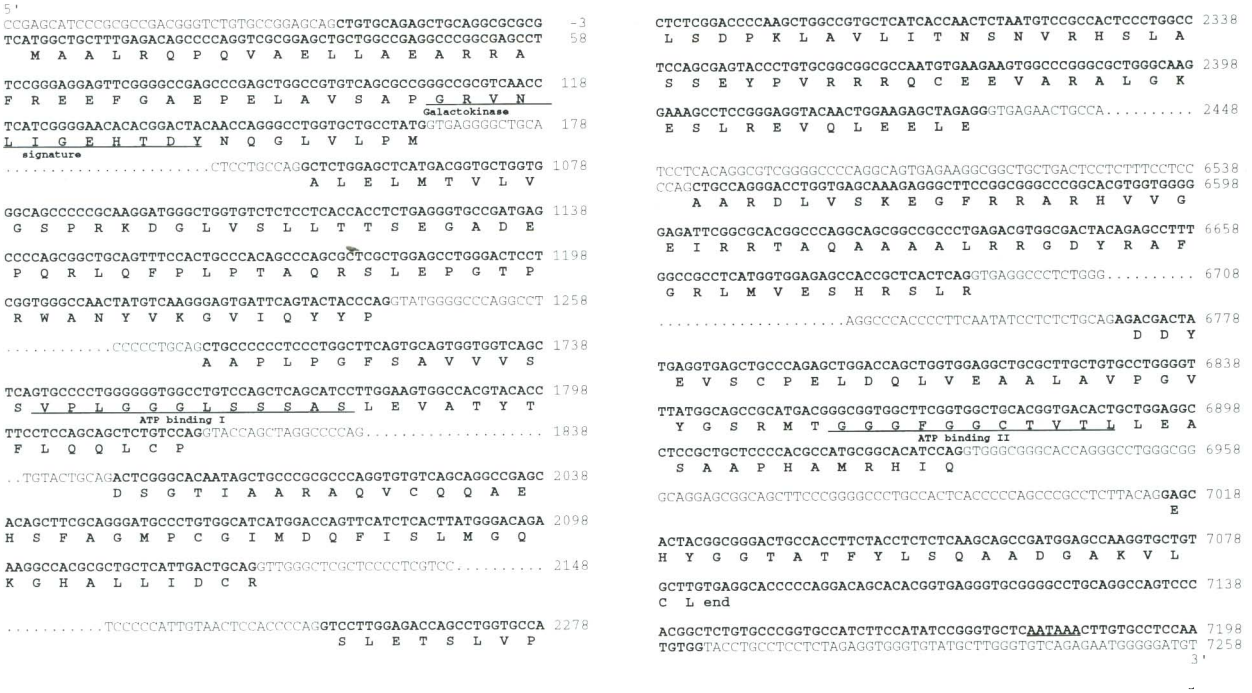

Figure 1 Nucleotide and deduced amino acid sequence of the human *GALK1* gene. Amino acids (represented by single-letter code) are indicated below their respective codons. Nucleotide positions are numbered at *right*. Underscored nucleotide sequence indicates the canonical eukaryotic polyadenylation signal. Nucleotide sequence depicted in boldface type is colinear with *GALK1* cDNA reported previously. The galactokinase signature sequence and two putative ATP-binding domains are underscored and labeled accordingly (GenBank accession nos. L76927).

Table 1. Exon—intron Organization of the Human Galactokinase Gene

Exon	Size (bp)	3′ Splice site (acceptor)		5′ Splice site (donor)	Intron (bp)	Intron type	Protein motif
1	179–227			CCTATG**gt**gagggg	886	0	Galactokinase signature
2	190	tctcctgcc**ag**GCT	–	ACCCAG**gt**atgggg	459	1	
3	120	tccccctgc**ag**CTG	–	GTCCAG**gt**accagc	170	1	ATP binding I
4	136	ctgtactgc**ag**ACT	–	CTGCAG**gt**tgggct	127	2	
5	182	ctccaccccc**ag**GTC	–	TAGAGG**gt**gagaac	4107	1	
6	151	ttcctcccc**ag**CTG	–	ACTCAG**gt**gaggcc	76	2	
7	163	cctctctgc**ag**AGA	–	ATCCAG**gt**gggcgg	82	0	ATP binding II
8	189	gcctcttac**ag**GAG					

Exon and intron sequences are shown in uppercase and lowercase, respectively. The ag/gt exon intron boundaries are in boldface type.

untranslated region containing a polyadenylation signal, AATAAA (Proudfoot and Brownlee 1976), positioned 18 nucleotides upstream of the cleavage/polyadenylation site (Fig. 1).

The *GALK1* Gene Promoter: Identification of Multiple Transcription Initiation Sites

The transcription start site for the *GALK1* cDNA was investigated by 5′-rapid amplification of cDNA ends (RACE) PCR (see Methods). Fragments derived from reverse transcriptase (RT)–PCR reactions were cloned into a plasmid vector and sequenced to help identify mRNA initiation start sites. The DNA sequences of 20 clones examined were all colinear with the genomic sequences present upstream of the coding region of the *GALK1* gene, suggesting that the promoter was adjacent to the coding region and not separated by an intron. Of 20 clones, 12 had different 5′ ends, suggestive of multiple starting sites, which is commonly found among housekeeping genes (Singer-Sam et al. 1984; Ingolia et al. 1986; Mitchell et al. 1986; Patel et al. 1986).

Because *cis* elements of a promoter are arrayed within a few hundred base pairs of the mRNA initiation site (Mitchell and Tjian 1989), we determined ~500 bp of upstream sequence, which is depicted in Figure 2. The computer program PROMOTER SCAN (Prestridge 1995) recognized a 250-bp region within this sequence as a eukaryotic PolII promoter. Notably, this region is positioned upstream of the multiple initiation sites. Typical of housekeeping gene promoters, the *GALK1* promoter lacks CCAAT-box and TATA-box motifs (Hajra and Collins 1995; Dynan 1986). The TATA box is often present ~25–40 bp upstream from the initiation site of many Pol II promoters (Benoist et al. 1980; Corden et al. 1980) and is apparently necessary for accurate mRNA initiation (Grosschedl and Birnstiel 1980).

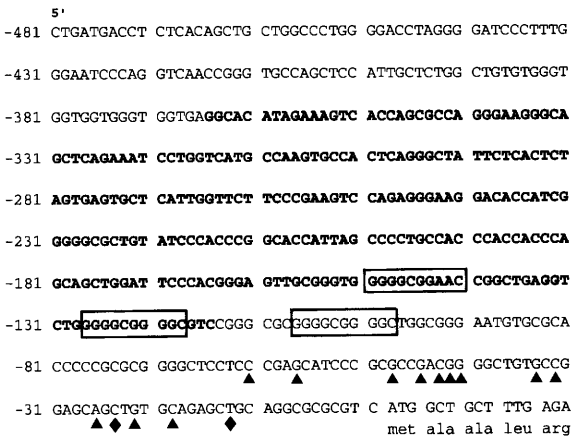

Figure 2 Human *GALK1* promoter. The negative numbering at *left* is based on the number of sequences preceding the A of the ATG used to encode the initiator methionine residue of GALK1 protein, which is shown below the sequence. The multiple transcription initiation sites predicted by RACE–PCR mapping (▲) or cDNA start sites (♦) are indicated below the sequence. Nucleotide sequence in boldface type was predicted by the computer program PROMOTER SCAN (Prestridge 1995) to be a eukaryotic Pol II promoter. The predicted Sp1 transcription-binding sites are boxed.

Therefore, it is possible that the multiple transcription start sites of the *GALK1* gene may be attributable to the absence of the TATA-box motif.

Another hallmark of housekeeping gene promoters is that they have a very high GC content, which reflects in part potential binding sites (consensus sequence 5′-G/TGGGCGGG/AG/AC/T-3′) for the transcription factor Sp1 (Dynan 1986; Kadonaga et al. 1986; Hajra and Collins 1995). Similarly, the *GALK1* promoter is very GC rich and contains three Sp1-binding motifs positioned immediately upstream of the multiple mRNA initiation sites. These motifs may be responsible for activating *GALK1* gene transcription.

Expression and Processing of *GALK1*

Translation of *GALK1* cDNA in the rabbit reticulocyte lysate system generated a protein of 42 kD, which was consistent with the expected size of the GALK1 translation product (Fig. 3). In the presence of microsomal membranes, newly synthesized chains remained fully cytosolic as demonstrated by sensitivity of chains to exogenously added proteinase K (Fig. 3, lanes 3–5). To confirm that newly synthesized GALK1 protein was not associated with membranes, translation products were incubated either at pH 11.5 (0.1 M sodium carbonate) or pH 7.5 before pelleting membranes. Under both conditions, chains remained in the supernatant fraction.

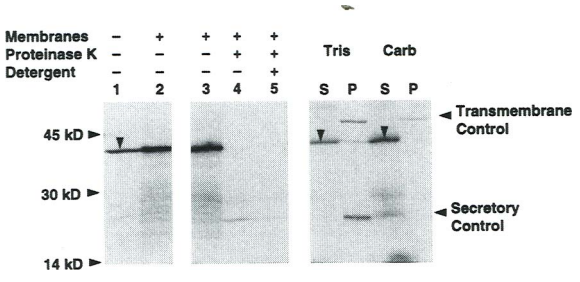

Figure 3 Translation products of *GALK1* cDNA. *GALK1* cDNA was expressed in rabbit reticulocyte lysate supplemented with canine pancreas microsomal membranes, as indicated. Full-length protein (lane *1*, ▼) migrated at the expected size of 42 kD. GALK1 protein synthesized in the presence of translocation-competent microsomal membranes was degraded by exogenous proteinase K (lanes *3–5*) and was recovered in the supernatant (S) but not the membrane pellet (P) following incubation at pH 7.5 (Tris) and pH 11.5 (Carb).

In this study we show that the human *GALK1* gene consists of 8 exons with a signature sequence and two ATP-binding sites dispersed among 3 separate exons. The availability of the human *GALK1* gene structure and organization will be useful in the characterization of *GALK1* gene defects in hereditary galactokinase deficiency and in the determination and analysis of the function of the galactokinase protein.

METHODS

Isolation of Phage Clones Spanning the Human *GALK1* Gene

The full-length human galactokinase cDNA (Stambolian et al. 1995) was used to screen a genomic phage library constructed with DNA prepared from human placenta and inserted into the *Not*I restriction site of the λ FIX II vector (Stratagene, La Jolla, CA). A total of 9×10^5 phage were screened by plaque hybridization (Sambrook et al. 1989) carried out overnight at 65°C in a solution containing 1 M NaCl, 1% SDS, 1× Denhardt's solution, and 100 μg/ml of denatured sheared salmon sperm DNA. Hybridization filters were washed at 60°C with 0.1 × SSC/1% SDS and exposed to autoradiography film for 2 days at −70°C. Six independent phage clones were isolated after secondary and tertiary screening. The clones were mapped with the restriction enzyme *Not*I by complete digestion and probing with various fragments of the human *GALK1* cDNA or by partial digestion and indirect labeling with T7 and T3 oligonucleotides.

Characterization of *GALK1* genomic clone

Three clones were found to contain the entire coding sequence for *GALK1*. The *Not*I digestion fragments from one clone, GK17, containing exonic sequences were isolated by agarose gel electrophoresis and Qiaex purification (Qiagen, Chatsworth, CA). The fragments were subcloned into pBluescript and sequenced on an automated ABI 373A sequencer. Oligonucleotide primers specific to the coding region of *GALK1* were initially used and additional primers for sequencing were synthesized based on the obtained sequence (GenBank accession no. L76927). The exon–intron junctions were identified by comparing the genomic sequence with the cDNA sequence. All splice donor and acceptor sites were compared with the consensus sequences established by Mount (1982).

Determination of the 5′ Ends of Galactokinase mRNA

The 5′ RACE technique was done using the Marathon-Ready cDNA kit (Clontech Laboratories Inc.). Briefly, 0.5 ng of double-stranded cDNA from human placenta was used as a template for PCR amplification with a primer complementary to the anchor sequence (5′-CCATCCTAATACGACTCACTATAGGGC-3′) and a gene-specific primer GK412 (5′-CTCCGCGACCTGGGGCTGTCT-

CAAAG-3′). PCR was done for 30 cycles at 94°C for 30 sec and 68°C for 4 min. Amplification products were cloned into the T/A cloning vector (Invitrogen). Twenty clones containing cDNA fragments were picked and sequenced with the SP6 primer derived from the linker region of the vector.

Expression and Processing of Galactokinase

The *GALK1* cDNA was amplified by 10 cycles of PCR using a sense oligonucleotide (5′-GCGCGCCATGGCTGCTTT-GAGAC-3′) and the antisense oligonucleotide (5′-CGCAAGATCTCCGTGTGCTGTCC-3′). The resulting fragment was digested with *Nco*I and *Bgl*II and ligated into plasmid pBP1 (Skach and Lingappa 1993) (derived from SP64T) (Melton and Krieg 1984) to generate plasmid pSP-GALK1. pSPGALK1 was translated in the rabbit reticulocyte lysate translation system supplemented with microsomal membranes as indicated, using standard translation conditions described previously (Skach and Lingappa 1993). Translation products [^{35}S] methionine labeled) were analyzed by SDS-PAGE and autoradiography. Reticulocyte lysate containing newly synthesized protein was digested with proteinase K (0.2 mg/ml) in the presence and absence of 1% Triton X-100 for 1 hr at 0°C before addition of 10 mM phenylmethylsulfonyl fluoride and transfered into 10 volumes of 0.1 M Tris (pH 8.0) and 1% SDS preheated to 100°C. Translation products were incubated in 200 volumes of 0.1 M sodium carbonate (pH 11.5) or 0.25 M sucrose, 0.1 M Tris (pH 7.5) for 30 min followed by centrifugation at 70,000 rpm for 30 min (Beckman TLA 100.2 rotor) to pellet membranes. *Trans*-membrane and secretory proteins, prolactin, and S.L.S.T.gG.P (Skach and Lingappa 1993) served as secretory and transmembrane controls, respectively.

Other Procedures

Genomic Southern blots were performed using standard procedures (Sambrook et al. 1989), with a final wash of 0.1 × SSC/1% SDS at 60°C–70°C. DNA probes were labeled by random priming (Sambrook et al. 1989). Plasmid DNA was prepared with Qiagen miniprep columns (Qiagen, Chatsworth, CA).

ACKNOWLEDGMENTS

We thank Drs. Christine Debouck, Ganesh Sathe, Mark Hurle, and James Fickett for support or help with computer programs during the course of this work. This work was supported by National Institutes of Health grant RO1 EY09404 (D.S.), CA 01614 (W.R.S.), NIMH 5-T32-MH-18902 (K.N.), the Benjamin and Mary Siddons Measey Foundation (K.N.), and the Pennsylvania Lions Sight Conservation and Eye Research Foundation. D.S. is a Research to Prevent Blindness James S. Adams Scholar. The sequence data described in this paper have been submitted to the GenBank data library under accession no. L76927.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Adams, C.W., J.A. Fornwald, F.J. Schmidt, M. Rosenberg, and M.E. Brawner. 1988. Gene organization and structure of the Streptomyces lividans gal operon. *J. Bacteriol.* **170:** 203–212.

Ai, Y., N.A. Jenkins, N.G. Copeland, D.J. Gilbert, D.J. Bergsma, and D. Stambolian. 1995. Mouse galactokinase: Isolation, characterization, and location on chromosome 11. *Genome Res.* **5:** 53–59.

Benoist, C., K. O'Hare, R. Breathnach, and P. Chambon. 1980. The ovalbumin gene-sequence of putative control regions. *Nucleic Acids Res.* **8**:127–142.

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.* **75**:4853–4857.

Citron, B.A. and J.E. Donelson. 1984. Sequence of the Saccharomyces GAL region and its transcription in vivo. *J. Bacteriol.* **158:** 269–278.

Cordon, J., B. Wasylyk, A. Buchwalder, P. Sassone-Corsi, C. Kedinger, and P. Chambon. 1980. Promoter sequences of eukaryotic protein coding genes. *Science* **209:** 1406–1414.

Debouck, C., A. Riccio, D. Schumperli, K. McKenney, J. Jeffers, C. Hughes, and M. Rosenberg. 1985. Structure of the galactokinase gene of E. coli, the last (?) gene of the gal operon. *Nucleic Acids Res.* **13:** 1841–1853.

Doolittle, W.F. 1978. Genes in pieces: Were they ever together? *Nature* **272:** 581–582

Dynan, W.S. 1986. Promoters for housekeeping genes. *Trends Genet.* **2:** 196–197.

Gilbert, W. 1978. Why genes in pieces? *Nature* **271:** 501.

Grosschedl, R. and M.L. Birnstiel. 1980. Identification of regulatory sequences in the prelude sequences of an H2A histone gene by the study of specific deletion mutants in vivo. *Proc. Natl. Acad. Sci.* **77:** 1432–1436.

Hajra, A. and F.S. Collins. 1995. Structure of the leukemia-associated human CBFB gene. *Genomics* **26:** 571–579.

Ingolia, D.E., M.R. Al-Ubaidi, C.Y. Yeung, H.A. Bigo, D. Wright, and R.E. Kellems. 1986. Molecular cloning of the murine adenosine deaminase gene from a genetically enriched source: Identifiction and charaterization of the promoter region. *Mol. Cell. Biol.* **6:** 4458–4466.

Kadonage, J.T. and R. Tjian. 1986. Affinity purification of sequence-specific DNA binding proteins. *PNAS* **83:** 5889–5893.

Melton, D.A. and P.A. Krieg. 1984. Efficient in vitro

synthesis of biologically active RNA and RNA hybridization probes from plasmids containing a bacterial SP6 promoter. *Nucleic Acids Res.* **12:** 7035–7056.

Mitchell, P.J., A.M. Carothers, J.H. Han, J.D. Harding, E. Kas, L. Venolia, and L.A. Chasin. 1986. Multiple transcription start sites, Dnase I-hypersensitive sites, and an opposite-strand exon in the 5′ region of the CHO dhfr gene. *Mol. Cell. Biol.* **6:** 425–440.

Mitchell, P.J. and R. Tjian. 1989. Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins. *Science* **245:** 371–378.

Mount, S.M. 1982. A catalogue of splice junction sequences. *Nucleic Acids Res.* **10:** 459–472.

Patel, P.I., P.E. Framson, C.T. Caskey, and C. Chinault. 1986. Fine structure of the human hypoxanthine phosphoribosyltransferase gene. *Mol. Cell. Biol.* **6:** 393–403.

Prestridge, D.S. 1995. Predicting Pol II promoter sequences using transcription factor binding sites. *J. Mol. Biol.* **249:** 923–932.

Proudfoot, N.J. and G.G. Brownlee. 1976. 3′ Non-coding region sequences in eukaryotic mRNA. *Nature* **263:** 211.

Sambrook, J., E.F. Fritsch, and T. Maniatis. 1989. *Molecular cloning: A laboratory manual*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Segal, S. 1989. Disorders of galactose metabolism. In: *The metabolic basis of inherited disease,* Vol. I, (ed. C.R. Scriver et al.), pp. 453–480. McGraw-Hill, New York, NY.

Segal, S., J.Y. Rutman, and G.W. Frimpter. 1979. Galactokinase deficiency and mental retardation. *J. Peds.* **95:** 750–752.

Shapiro, M.B. and P. Senapathy. 1987. RNA splice junctions of different classes of eukaryotes: Sequence statistics and functional implications in gene expression. *Nucleic Acid Res.* **15:** 7155–7174.

Singer-Sam, J., D.H. Keith, K. Tani, R.L. Simmer, L. Shively, S. Lindsay, A. Yoshida, and A.D. Riggs. 1984. Sequence of the promoter region of the gene for human X–linked 3-phosphoglycerate kinase. *Gene* **32:** 409–417.

Skach, W. and V. Lingappa. 1993. Amino terminus assembly of human P-glycoprotein at the endoplasmic reticulum is directed by cooperative actions of two internal sequences. *J. Biol. Chem.* **268:** 23552–23561.

Stambolian, D., Y. Ai, D. Sidjanin, K. Nesburn, G. Sathe, M. Rosenberg, and D.J. Bergsma. 1995. Cloning of the galactokinase cDNA and identification of mutations in two families with cataracts. *Nature Genet.* **10:** 307–317.

Stambolian, D., V. Scarpino-Myers, R.C. Eagle, B. Hodes, and H. Harris. 1986. Cataracts in patients heterozygous for galactokinase deficiency. *Invest. Opthal. Vis. Sci.* **27:** 429–433.

Tsay, Y.H. and G.W. Robinson. 1991. Cloning and characterization of ERG8, an essential gene of Saccharomyces cerevisiae that encodes phosphomevalonate kinase. *Mol. Cell. Biol.* **11:** 620–631.

Received April 23, 1996; accepted in revised form June 21, 1996.