Vol. 173, No. 1, 1990 BIOCHEMICAL AND BIOPHYSICAL RESEARCH COMMUNICATIONS
November 30, 1990 Pages 141-148

MOLECULAR CLONING OF MOUSE ACID ß-GALACTOSIDASE cDNA: SEQUENCE, EXPRESSION OF CATALYTIC ACTIVITY AND COMPARISON WITH THE HUMAN ENZYME*

Eiji Nanba and Kunihiko Suzuki

Brain and Development Research Center,
Departments of Neurology and Psychiatry
University of North Carolina School of Medicine
Chapel Hill, NC 27599

Received October 12, 1990

ABSTRACT: A full-length cDNA coding for mouse lysosomal acid ß-galactosidase has been isolated on the basis of homology with the human gene. Catalytic activity toward 4-methylumbelliferyl ß-D-galactoside in the COS-1 cell expression system provided positive proof for its authenticity. The sequence analysis showed that the degree of similarity between the human and mouse enzymes was approximately 70% in the nucleotide sequence and nearly 80% in the amino acid sequence. The deduced primary amino acid sequences of the enzymes from the two species indicated that, of the seven possible N-glycosylation sites in the human enzyme, five are conserved in the mouse enzyme. Three additional possible N-glycosylation sites, not present in the human enzyme, are found in the primary amino acid sequence of the mouse enzyme. All seven cysteine residues in the mouse enzyme are conserved in the human enzyme. Although the nucleotide sequence could be aligned to 60% identity with the E. coli ß-galactosidase, similarity in the amino acid sequence was minimal.

Mammalian acid ß-galactosidase is a lysosomal enzyme which hydrolyzes the terminal ß-D-galactoside residues from glycolipids and glycoproteins in the normal catabolic process of tissue constituents. Two groups of disorders are caused by genetic deficiencies of its activity, a severe, rapidly progressive and fatal neurological disorder, GM1-gangliosidosis, and Morquio B disease, primarily a skeletal disorder (1). The phenotypic differences occur presumably because mutations underlying the two diseases differentially affect its specificity toward different natural substrates.

cDNAs coding for human acid ß-galactosidase have been cloned by Oshima et al. (2), Morreau et al. (3) and more recently by Yamamoto et al. (4). The published nucleotide sequences generally agree well with only minor differences (3). We have cloned full-length cDNAs coding for mouse acid ß-galactosidase, taking advantage of the expected homology between the human and murine enzymes. We report here the sequence of the mouse acid ß-

*Reprint requests and correspondence should be addressed to Kunihiko Suzuki, M.D., Brain and Development Research Center, CB#7250, University of North Carolina School of Medicine, Chapel Hill, NC 27599, USA. The nucleotide sequence of the mouse ß-galactosidase cDNA is available through GenBank (accession #M33769). We will provide hard copies upon request or an ASCII file on IBM-PC-compatible format, if a blank diskette is sent with a request.

0006-291X/90 $1.50

galactosidase cDNA, catalytic expression in the COS-1 cell system, and comparison with the human enzyme. As expected, the human and murine acid ß-galactosidases are homologous with a high degree of similarities.

MATERIALS AND METHODS

Commercial Materials: Bethesda Research Laboratories (Gaithersburg, MD), Boerhinger Mannheim (Indianapolis, IN), Pharmacia (Piscataway, NJ), International Biotechnologies Inc. (New Haven, CT) and New England Biolab (Beverley, MA) were the main sources for enzymes, reagents and other molecular biological supplies. Radioisotopes were obtained from ICN Radiochemicals (Irvine, CA). 4-Methylumbelliferyl ß-galactoside and 4-methylumbelliferyl ß-N-acetylglucosaminide were purchased from Sigma Chemical Co. (St. Louis, MO). Transformed African green monkey kidney cells (COS-1) were purchased from the American Type Culture Collection (ATCC)(Rockville, MD). Sources for other non-standard materials will be indicated below as appropriate.

Human Acid ß-Galactosidase cDNA: A full-length human cDNA coding for acid ß-galactosidase was isolated in two segments and subsequently ligated. The 5'-segment from the 5' terminus to the internal Eco RI site was isolated by polymerase chain reaction (PCR) (5) of a human fibroblast cDNA library with two primers based on the sequence published by Oshima et al. (3). The 5'-primer (5'-TATCTAGAGACTGCAGAGCCGGGAGGCTG) was the coding strand sequence at the 5'-terminus with an extra Xba I sequence, and the 3'-primer (5'-CCAGCCAGTATAGAATTCAGA) was the non-coding strand sequence around the internal Eco RI site. Amplified fragments of the expected size, 845 bp, was obtained from a double-stranded cDNA mixture from normal human fibroblasts. A similarly designed amplification for the 3'-segment was unsuccessful and the downstream segment was isolated by screening a cDNA library in λgt11 (6) directly with the above upstream segment obtained by PCR as the probe. The two segments were ligated at the internal Eco RI site to generate a full-length cDNA and subsequently ligated into pGEM 7Zf(+) (Promega Corp., Madison, WI). The sequence of our cDNA clone was compared with that of Oshima et al (3). The human ß-galactosidase cDNA was digested with Hind III to remove the poly-A tract and purified by agarose gel electrophoresis and electroelution for subsequent use as the probe for isolation of mouse ß-galactosidase cDNA.

Mouse cDNA Libraries: cDNA libraries were prepared from cultured fibroblasts and brain tissues of adult C57BL/6J mice. The total RNA was extracted with guanidine isothiocyanate (7) and the poly A+ fraction isolated by oligo(dT)-cellulose chromatography (Collaborative Research, Bedford, MA). Double-stranded cDNAs were generated essentially according to Gubler and Hoffman (8). After attachment of the Eco RI linker, they were electrophoresed in 1% agarose gel, and cDNAs larger than approximately 600 bp were isolated by electroelution. They were ligated into the λ gt10 phage (9).

Isolation of Mouse ß-Galactosidase cDNA: The mouse brain cDNA library was screened under a low-stringency conditions with the human ß-galactosidase cDNA labelled with [α-^{32}P]dCTP by nick translation. Phage plaques were blotted to Nylon membrane (Colony-Plaque Screen, New England Nuclear Corp., Boston, MA). The solution for prehybridization and hybridization was essentially those of Wahl et al. (10). The hybridization solution contained 50% formamide, 5 x SCC, 10% dextran sulfate, 2 x Denhardt's solution, 20 mM sodium phosphate buffer, pH 6.5, 0.5% SDS and 250 μg/ml of denatured salmon sperm DNA. Hybridization was for 40 hrs at 28°. The final washing condition was at 48° in 2 x SCC and 1% SDS. Initial clones isolated on the basis of low-stringency hybridization were then divided into separate cross-hybridizing groups under high-stringency conditions; in the same solutions as described above except for the hybridization temperature at 42°, and the final washes at 60° in 0.2 x SCC and 1% SDS.

COS-1 Cell Expression System: Expression of ß-galactosidase activity in the COS-1 cell expression system was used for testing the authenticity of isolated mouse clones. The putative mouse ß-galactosidase cDNA clone in pGEM 7Zf(+) was further subcloned into the pSVL expression vector with Xba I and Sac I digestion. Since the orientation of the cDNA with respect to the coding strand was not known at this stage, clones of both orientations were generated and purified. The COS-1 cells were transfected by the lipofection procedure (11) according to the instruction manual accompanying the lipofection kit (Lipofectin Reagent, Bethesda Research Laboratories). ß-Galactosidase activities were assayed 48 or 72 hrs after transfection with 4-methylumbelliferyl ß-D-galactoside (12). ß-Hexosaminidase activity was

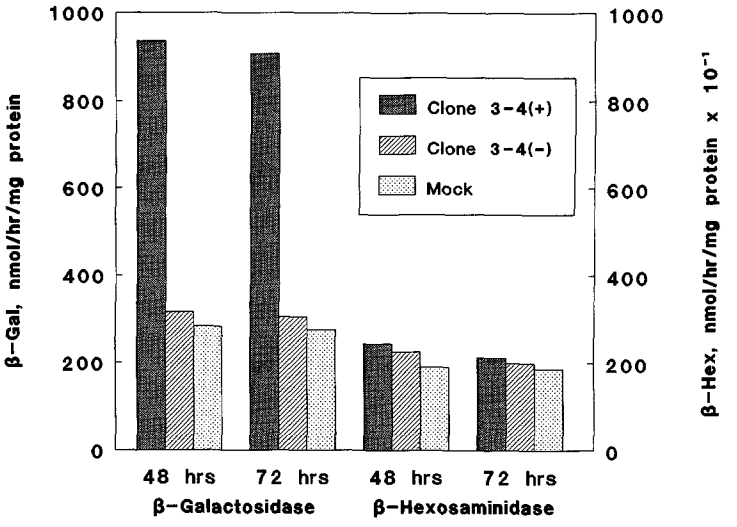

Fig. 1. Expression of ß-galactosidase activity in the COS-1 cell system. The methodological details are described in the text. Activities of ß-galactosidase and ß-hexosaminidase (control) were determined at 48 hrs and 72 hrs after transfection. The clone designation is arbitrary. The "plus (+)" and the "minus (-)" signs indicate different orientations of the cDNA in the vector. Only the clone 3-4 (+) expressed acid ß-galactosidase activity significantly higher than the endogenous baseline activity.

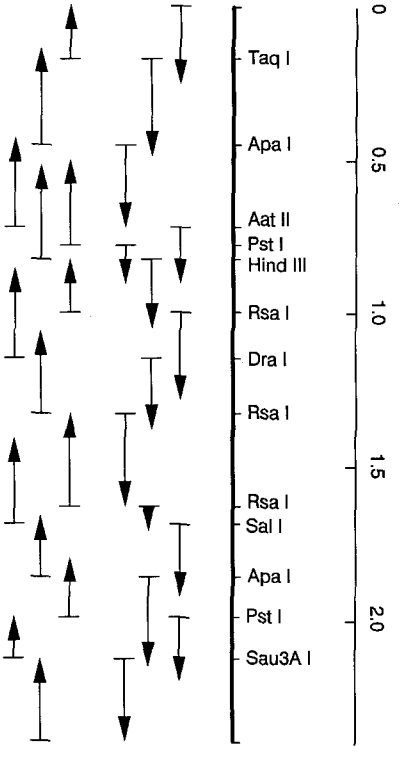

Fig. 2. The restriction map and the sequencing strategy for the murine acid ß-galactosidase cDNA. Both strands were completely sequenced and all restriction sites used for subcloning were bridged to exclude two identical restriction sites present in very close proximity. The scale at the top indicates approximate positions in base pairs from the 5'-terminus.

assayed as the control enzyme with 4-methylumbelliferyl ß-D-N-acetylglucosaminide. The amount of cellular protein was determined by the dye-binding assay (13). Experiments were repeated at least once more and often a few times, with duplicate plates in each experiment to insure reproducibility of the results.

Northern Blotting Analysis: The poly-A+ RNA fraction isolated from cultured mouse fibroblasts, 1 μg, was electrophoresed in 1% agarose/formaldehyde denaturing gel (14), transferred to Gene Screen Plus Nylon membrane (New England Nuclear Corp., Boston, MA), and probed with the isolated mouse ß-galactosidase cDNA labelled with [α-^{32}P]dCTP by nick translation. Hybridization and washing were done according to the high stringency conditions described above for Southern blotting.

DNA Sequence Analysis: The sequences of the inserts were determined in the double-stranded plasmid, pGEM 7Zf(+) or pGEM 3Zf(-) by the dideoxy chain termination method of Sanger et al. (15) with either the T7 promoter primer or SP6 promoter primer and ^{35}S-labelled dATP (16). The DNA polymerase used for sequencing was T7 polymerase (Sequenase, version 2, US Biochem. Corp., Cleveland, OH). Occasionally, the M13 universal primer or 7-deaza-dGTP (US Biochem. Corp, Cleveland, OH) were used.

RESULTS

Isolation of Putative Mouse Acid ß-Galactosidase cDNA: A total of 33 clones positive under the low-stringency conditions were initially obtained by screening 10^6 plaques from an adult mouse brain cDNA library. These clones could be divided into five separate cross-hybridizing groups under the high-stringency conditions. The size distribution of eight clones belonging to one of the five groups, which was later shown to be acid ß-galactosidase, was 0.7-1.3 kb, clearly smaller than expected for full-length ß-galactosidase cDNA. Therefore, another mouse cDNA library constructed from cultured mouse fibroblasts was screened with one of these clones under the high-stringency condition. Out of 2.6 x 10^5 plaques screened, 31 positive clones were obtained and 24 were purified. Of the 24 clones, five had an identical size (2.4 kb) and restriction map. Although the library had been previously amplified, poly-A tracts of different sizes were found among these clones, indicating that they derived from more than one clone in the original library. The 2.4-kb size was consistent with the expected size of the full-length mRNA estimated by Northern blotting (data not shown). One of these 2.4-kb cDNAs was subcloned into the Eco RI site of pGEM 7Zf(+), and clones with the insert in both orientations were purified for sequencing.

Expression of Catalytic Activity: The purified cDNA was subcloned into the pSVL expression vector in both orientations and used to transfect COS-1 cells. At 48 and 72 hrs after transfection, a three-fold increase in acid ß-galactosidase activity was observed in cells transfected with the cDNA in one orientation, while no significant increase was seen in cells transfected with the cDNA in the reverse orientation (Fig. 1). The activity of the control enzyme, ß-hexosaminidase, always remained at the basal level.

Sequence Analysis and Comparison with the Human Enzyme (Fig. 2 and 3): The strategy for restriction enzyme digestion, subcloning and sequencing of the mouse ß-galactosidase cDNA is shown in Fig. 2. The cDNA coding for the mouse acid ß-galactosidase was 2353-bp long, with 73-bp 5'-untranslated and 339-bp 3'-untranslated regions terminating with a poly-A tract. The open frame is 1941-bp long, coding for 647 amino acids. The deduced amino acid sequence includes the first 24 amino acids that satisfy the criteria for a signal sequence and eight possible N-glycosylation sites (Asn-X-Thr or -Ser).

The nucleotide sequence of our human ß-galactosidase cDNA was identical with that of Morreau et al. (3) [GenBank accession #J05124], including the four areas where discrepancies were noted between theirs and that of Oshima et al. (2). Morreau et al. (3) pointed out discrepancies in three areas but appear to have overlooked another discrepancy at nucleotide #34 counted from the A of the initiation codon, ATG. As expected, the human and murine

```
HUMAN   - MPGFLVRILLLLLVLL-LLGPTRGLRNATQRMFEIDYSRDSFLKDGQPFR -49
MOUSE   - MLRVPLCTPLPLLALLQLLGAAHGIYNVTQRTFKLDYSRDRFLKDGQPFR -50

HUMAN   - YISGSIHYSRVPRFYWKDRLLKMKMAGLNAIQTYVPWNFHEPWPGQYQFS -99
MOUSE   - YISGSIHYFRIPRFYWEDRLLKMKMAGLNAIQMYVPWNFHEPQPGQYEFS -100

HUMAN   - EDHDVEYFLRLAHELGLLVILRPGPYICAEWEMGGLPAWLLEKESILLRS -149
MOUSE   - GDRDVEHFIQLAHELGLLVILRPGPYICAEWDMGGLPAWLLEKQSIVLRS -150

HUMAN   - SDPDYLAAVDKWLGVLLPKMKPLLYQNGGPVITVQVENEYGSYFACDFDY -199
MOUSE   - SDPDYLVAVDKWLAVLLPKMKPLLYQNGGPIITVQVENEYGSYFACDYDY -200

HUMAN   - LRFLQKRFRHHLGDDVVLFTTDGAHKTFLKCGALQGLYTTVDFGTGSNIT -249
MOUSE   - LRFLVHRFRYHLGNDVILFTTDGASEKMLKCGTLQDLYATVDFGTGNNIT -250

HUMAN   - DAFLSQRKCEPKGPLINSEFYTGWLDHWGQPHSTIKTEAVASSLYDILAR -299
MOUSE   - QAFLVQRKFEPKGPLINSEFYTGWLDHWGKPHSTVKTKTLATSLYNLLAR -300

HUMAN   - GASVNLYMFIGGTNFAYWNGANSPYAAQPTSYDYDAPLSEAGDLTEKYFA -349
MOUSE   - GANVNLYMFIGGTNFAYWNGANTPYEPQPTSYDYDAPLSEAGDLTKKYFA -350

HUMAN   - LRNIIQKFEKVPEGPIPPSTPKFAYGKVTLEKLKTVGAALDILCPSGPIK -399
MOUSE   - LREVIQMFKEVPEGPIPPSTPKFAYGKVALRKFKTVAEALGILCPNGPVK -400

HUMAN   - SLYPLTFIQVKQHYGFVLYRTTLPQDCSNPAPL-SSPLNGVHDRAYVAVD -448
MOUSE   - SLYPLTFTQVKQYFGYVLYRTTLPQDCSNPKPIFSSPFNGVRDRAYVSVD -450

HUMAN   - GIPQGVLERNNVITLNITGKAGATLDLLVENMGRVNYGAYINDFKGLVSN -498
MOUSE   - GVPQGILDRNLMTALNIRGKAGATLDILVENMGRVNYGRFINDFKGLISN -500

HUMAN   - LTLSSNILTDWTIFPLDTEDAVRSHLGGWGHRDSGHHDEAWAHNSSNYTL -548
MOUSE   - MTINSTVLTNWTVFPLNTEAMVRNHLWGREASDEGHLDGRSTSNSSDLIL -550

HUMAN   - PAFYMGNFSIPSGIPDLPQDTFIQFPGWTKGQVWINGFNLGRYWPARGPQ -598
MOUSE   - PTFYVGNFSIPSGIPDLPQDTFIQFPGWSKGQVWINGFNLGRYWPTMGPQ -600

HUMAN   - LTLFVPQHILMTSAPNTITVLELEWAPCSSDDPELCAVTFVDRPVIGSSV -648
MOUSE   - KTLFVPRNILTTSAPNNITVLELEFAPCSEGTPELCTVEFVDTPVIS    -647

HUMAN   - TYDHPSKPVEKRLMPPPPQKNKDSWLDHV -677
```

Fig. 3. Alignment of the deduced amino acid sequences of the human and murine acid ß-galactosidases. The character to show that two aligned residues are identical is " | ", and the character to show that two aligned residues are similar is ":". Groups of amino acids considered "similar" are: A,S,T; D,E; N,Q; R,K; I,L,M,V; and F,Y,W. With two short gaps in the human enzyme, the two sequences can be aligned to 500 identical and 45 "similar" amino acids. The signal sequences are in bold-face. The conserved putative glycosylations sites are underlined, while those present only in one of the two sequences are double-underlined. Similarly, the seven conserved cysteine residues are in bold-face and underlined, while the sole cysteine residue in the human sequence not found in the mouse sequence is double-underlined.

acid ß-galactosidases [the murine nucleotide sequence available from GenBank, accession #M33769] are homologous with over 75% identity in the base sequences. The sequence homology was well conserved throughout all regions of the cDNA, including the 3'-untranslated region (58%). The similarity between the deduced primary amino acid sequences was even greater at nearly 80% identity (Fig. 3). However, no more than half of amino acids constituting the signal sequences are identical between the two enzymes, and the human enzyme has 30 extra amino acids at the carboxyl terminus. Of the eight possible N-glycosylation sites in the mouse enzyme, five are conserved in the human enzyme. The other three sites are not present in the human enzyme, while the mouse enzyme lacks two glycosylation sites found in the human enzyme. All of the seven cysteine residues in the mouse enzyme sequence are conserved in the human enzyme. The human enzyme has an additional cysteine residue which has no counterpart in the mouse enzyme.

DISCUSSION

The expression of ß-galactosidase activity in the COS-1 cells provided definitive evidence for authenticity of the mouse ß-galactosidase cDNA isolated on the basis of sequence similarity with the human cDNA. It included the entire protein coding sequence. We observed identical 5'-termini in more than one cDNA having poly-A tracts of different lengths. This suggests, although does not definitively prove, that the 5'-terminus of our mouse cDNA may indeed be the transcription initiation site. In humans, Morreau et al. (3) detected a shorter cDNA arising from the ß-galactosidase transcripts by an alternate splicing. We did not find a counterpart in the mouse, although we did not specifically sought for it. We did not see a shorter mRNA on the Northern blotting either; however, again Morreau et al. could detect such mRNA only with immuno-selected polysomal RNA fraction. Thus, our results do not exclude presence of a similar alternately-spliced smaller mRNA species in the mouse.

As expected, the human and murine acid ß-galactosidases are evolutionarily-related homologous proteins. Sequence similarities, particularly that of the amino acid sequence, are very high. The human acid ß-galactosidase is heavily glycosylated. Five glycosylation sites are conserved between the two species. The only stretches of amino acid sequences of substantial lengths where similarity falls at or below 50% are the signal sequence at the N-terminus and the areas of amino acid residues 500-550, in addition to the last 30 amino acids in the human enzyme which have no counterpart in the mouse enzyme.

In contrast, comparison of either of the mammalian enzymes with E. coli ß-galactosidase was of interest in that, although it was possible to align the nucleotide sequences to approximately 60% identity, the amino acid sequences could not be aligned to more than 20% identity (data not shown). These findings might be taken to mean that the mammalian genes evolved from the bacterial gene but that functionality of the proteins as enzymes have diverged. It is known, for example, that the E. coli and some other bacterial ß-galactosidases do not catalyze hydrolysis of the terminal ß-galactose residues from a series of sphingoglycolipids, while mammalian ß-galactosidases do (our own unpublished observations).

Galjart et al. (17) recently reported on the mouse "protective protein" comparing it with the homologous human protein (18). The "protective protein" is a lysosomal glycoprotein. Its function is to protect the ß-galactosidase/neuraminidase complex from degradation in the acidic intra-lysosomal environment. The degree of similarities between the "protective proteins" from the two species was even slightly higher than those between the acid ß-galactosidases. Of potential interest is the relatively high conservation of the 3'-untranslated sequences between the two species in both proteins.

The human ß-galactosidase gene has been mapped to chromosome 3 and the mouse gene to chromosome 9 (19, 20). An earlier observation by de Wit et al. (21) that the human gene maps to chromosome 22 was recently explained on the basis that the gene coding for the protective protein is located on chromosome 22 (22). Utilizing the cDNA as the probe, we have isolated a series of genomic DNA fragments containing the mouse ß-galactosidase sequence. The earlier localization studies of the mouse gene were done with the somatic cell hybridization technique. Availability of the cDNA and genomic clones now makes it feasible to confirm these observations directly by in situ hybridization.

ACKNOWLEDGMENTS

Dr. Yoshiyuki Suzuki, Tokyo Metropolitan Institute for Medical Science, provided us with his human acid ß-galactosidase cDNA as a reference standard. We thank Mr. Jose Langaman for his excellent assistance in the fibroblast cell culture. This investigation was supported in part by research grant, RO1 NS-24289, and Mental Retardation Research Center Core Grant, P30 HD-03110, from the United States Public Health Service.

REFERENCES

1. O'Brien, J. S. (1989) In The Metabolic Basis of Inherited Disease (C. R. Scriver, A. L. Beaudet, W. S. Sly and D. Valle, Eds.), pp. 1797-1806, McGraw-Hill, New York.
2. Oshima, A., Tsuji, A., Nagao, Y., Sakuraba, H. and Suzuki, Y. (1988) Biochem. Biophys. Res. Commun. **157**:238-244.
3. Morreau, H., Galjart, N. J., Gillemans, N., Willemsen, R., van der Horst, G. T. J. and d'Azzo, A. (1989) J. Biol. Chem. **264**:20655-20663.
4. Yamamoto, Y., Hake, C. A., Martin, B. M., Kretz, K. A., Ahernrindell, A. J., Naylor, S. L., Mudd, M. and O'Brien, J. S. (1990) DNA and Cell Biol. **9**:119-127.
5. Saiki, R. K., Gelfand, D. H., Stoffel, S., Scharf, S. J., Higuchi, R., Horn, G. T., Mullis, K. B. and Erlich, H. A. (1988) Science **239**:487-491.
6. Nakano, T. and Suzuki, K. (1989) J. Biol. Chem. **264**:5155-5158.
7. Okayama, H., Kawaichi, M., Brownstein, M., Lee, F., Yokota, T. and Arai, K. (1987) Methods Enzymol. **154**:3-28.
8. Gubler, U. and Hoffman, B. J. (1983) Gene **25**: 263-269.
9. Jendrisak, J., Young, R. A. and Engel, J. D. (1987) Methods Enzymol. **152**:359-371.
10. Wahl, G. M., Stern, M. and Stark, G. R. (1979) Proc. Natl. Acad. Sci., U.S.A. **76**:3683-3687.
11. Felgner, P. L., Gadek, T. R., Holm, M., Roman, R., Chan, H. W., Wenz, M., Northrop, J. P., Ringold, G. M. and Danielsen, M. (1987) Proc. Natl. Acad. Sci., U.S.A. **84**:7413-7417.
12. Suzuki, K. (1987) Methods Enzymol **138**:727-762.
13. Bradford, M. (1976) Anal. Biochem. **74**:5463-5467.
14. Davis, L. G., Dibner, M. D. and Battey, J. F. (1986) In Basic Methods in Molecular Biology, pp. 143-146. Elsevier, Amsterdam.
15. Sanger, F., Nicklen, S. and Coulson, A. R. (1977) Proc. Natl. Acad. Sci., U.S.A. **74**:5463-5467.

16. Biggin, M. D., Gibson, J. J. and Hong, G. F. (1983) Proc. Natl. Acad. Sci., U.S.A. **80**:3963-3965.
17. Galjart, N. J., Gillemans, N., Meijer, D. and d'Azzo, A. (1990) J. Biol. Chem. **265**:4678-4684.
18. Galjart, N. J., Gillemans, N., Harris, A., van der Horst, G. T. J., Verheijen, F. W., Galjaard, H. and d'Azzo, A. (1988) Cell **54**:755-764.
19. Shows, T. B., Scrafford-Wolff, L. R., Brown, J. A. and Meisler, M. H. (1979) Somat. Cell Genet. **5**:147-158.
20. Naylor, S. L., Elliott, R. W., Brown, J. A. and Shows, T. B. (1982) Am. J. Human Genet. **34**:235-244.
21. de Wit, J., Hoeksema, H. L., Halley, D., Hagemeijer, A., Bootsma, D. and Westerveld, A. (1977) Somat. Cell Genet. **3**:351-363.
22. Sips, H. J., de Wit-Verbeek, H. A., de Wit, J., Westerveld, A. and Galjaard, H. (1975) Human Genet. **69**:340-344.