Proc. Natl. Acad. Sci. USA
Vol. 93, pp. 6101–6105, June 1996
Medical Sciences

The molecular basis of Sanfilippo syndrome type B

(α-*N*-acetylglucosaminidase/mucopolysaccharidosis III B/lysosomal storage disease)

HONG G. ZHAO*, HONG HUA LI*, GIDEON BACH†, ARTUR SCHMIDTCHEN*, AND ELIZABETH F. NEUFELD*‡

*Department of Biological Chemistry, Brain Research Institute and Molecular Biology Institute, University of California at Los Angeles, Los Angeles, CA 90095-1737; and †Department of Human Genetics, Hadassah Medical Center, Jerusalem, Israel

Contributed by Elizabeth F. Neufeld, February 16, 1996

ABSTRACT **The Sanfilippo syndrome type B is a lysosomal storage disorder caused by deficiency of α-*N*-acetylglucosaminidase; it is characterized by profound mental deterioration in childhood and death in the second decade. For understanding the molecular genetics of the disease and for future development of DNA-based therapy, we have cloned the cDNA and gene encoding α-*N*-acetylglucosaminidase. Cloning started with purification of the bovine enzyme and use of a conserved oligonucleotide sequence to probe a human cDNA library. The cDNA sequence was found to encode a protein of 743 amino acids, with a 20- to 23-aa signal peptide immediately preceding the amino terminus of the tissue enzyme and with six potential N-glycosylation sites. The 8.5-kb gene (*NAGLU*), interrupted by 5 introns, was localized to the 5′-flanking sequence of a known gene, *EDH17B*, on chromosome 17q21. Five mutations were identified in cells of patients with Sanfilippo syndrome type B: 503del10, R297X, R626X, R643H, and R674H. The occurrence of a frameshift and a nonsense mutation in homozygous form confirms the identity of the *NAGLU* gene.**

The Sanfilippo syndrome (mucopolysaccharidosis III) is an autosomal recessive disorder that comprises four subtypes (reviewed in ref. 1). Each is due to deficiency of one of four lysosomal enzymes that participate in the removal of sulfated *N*-acetylglucosamine residues during the degradation of heparan sulfate: heparan *N*-sulfatase (A subtype), α-*N*-acetylglucosaminidase (B subtype), acetylCoA:α-glucosaminide acetyltransferase (C subtype), and *N*-acetylglucosamine 6-sulfatase (D subtype). In the absence of any one of these enzymes, undegraded or partially degraded heparan sulfate accumulates in lysosomes and is excreted in urine. Profound mental retardation with relatively mild somatic manifestations is characteristic of the Sanfilippo syndrome. The disease is particularly difficult for families because during childhood, developmental delay is often accompanied by intractable hyperactivity; this phase is followed by a more quiet period with progressive loss of mental function. Magnetic resonance imaging shows atrophy of the brain (2). There is clinical variability between and within the subtypes, and even intrafamilial variability in the B subtype (3, 4). Patients with the B subtype usually live until the late teens, but longer survival occurs among mildly affected patients.

The Sanfilippo syndrome is a rare disorder that may often remain undiagnosed because of the nonspecific nature of early manifestations. Geographic distribution reported for European populations is uneven, with the A subtype prevalent in the British Isles, the B subtype most common in Southern Europe, and the A, B, and C subtypes distributed in a 3:2:1 ratio in the Netherlands (1). The Sanfilippo syndrome has been studied relatively little, and the biochemical and cellular mechanisms that underlie the behavioral disturbances and neurodegeneration are not understood. Recent interest in the Sanfilippo syndrome has resulted in the cloning of the cDNAs encoding *N*-acetylglucosamine 6-sulfatase (5) and of heparan *N*-sulfatase (6), as well as development of a caprine model of Sanfilippo type D (7). We undertook the cloning of the Sanfilippo B cDNA and gene in order to permit molecular studies of phenotypic variability, brain degeneration, and potential therapy. Preliminary accounts of this work have been published in abstract form (8, 9).

MATERIALS AND METHODS

Purification of α-*N*-acetylglucosaminidase. Enzyme activity was assayed as described (10), using the fluorogenic substrate, 4-methylumbelliferyl-α-*N*-acetylglucosaminide (Calbiochem). Protein concentration was determined by the bicinchoninic acid assay (11).

An animal source was used as starting material for enzyme purification to avoid exposure to viral pathogens in human tissues. A preliminary survey showed bovine testes (Pel-Freez Biologicals) to be satisfactory with respect to enzyme activity and cost. The enzyme was purified from 3-kg batches by sequential chromatography on concanavalin A-Sepharose, DEAE-Sepharose, and phenyl-Sepharose CL-4B (12). After this partial purification (≈2000-fold), the enzyme was subjected to preparative SDS/PAGE without preheating, a procedure that preserved α-*N*-acetylglucosaminidase activity. After electrophoresis overnight at 4°C, a sample strip of the gel was soaked for 30 min at 4°C in 0.1 M sodium acetate buffer (pH 4.3) containing 2.5% Triton X-100 in order to remove the SDS. The gel strip was then laid down for a few minutes on a piece of Whatman no. 1 filter paper impregnated with 0.2 mM substrate in 150 mM sodium acetate (pH 4.3), and the fluorescent product of the enzyme reaction was detected under UV light. Two fluorescent bands, of apparent mass of 170 kDa and 87 kDa, were observed (Fig. 1*A*). The larger one corresponded precisely to a well-defined protein band and was therefore selected for sequence analysis of amino-terminal and internal peptides. Tryptic digestion was carried out *in situ* (13). The 170-kDa band from separate preparations was eluted for generation of internal peptides by cyanogen bromide cleavage and for amino-terminal sequencing; for the latter, it was subjected to an additional electrophoretic purification and blotted onto a polyvinyledene difluoride membrane. Amino acid sequence analysis was performed by the University of California at Los Angeles Protein Microsequencing Facility. The amino-terminal sequence of human α-*N*-acetylglucosa-

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviations: RT-PCR, reverse transcriptase PCR; SSCP, single strand conformation polymorphism; RACE, rapid amplification of cDNA ends.

Data Deposition: The sequences reported in this paper have been deposited in the GenBank data base (accession nos. U43572 and U43573).

‡To whom reprint requests should be addressed at: Department of Biological Chemistry, University of California at Los Angeles School of Medicine, Los Angeles, CA 90095-1737. e-mail: liz@ biochem.medsch.ucla.edu

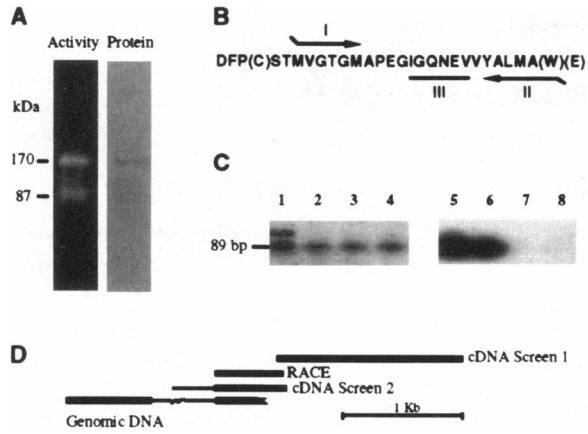

FIG. 1. Strategy for cloning cDNA encoding α-*N*-acetylglucosaminidase. (*A*) Comigration of α-*N*-acetylglucosaminidase activity and 170-kDa protein of enzyme preparation subjected twice to SDS/PAGE. (*B*) Internal tryptic peptide from bovine α-*N*-acetylglucosaminidase, and positions of corresponding fully degenerate oligonucleotides I, II, and III. (*C*) Amplification products from the RT-PCR visualized by ethidium bromide staining (lanes 1–4) or Southern hybridization with oligonucleotide III (lanes 5–8); lanes 1, 2, 5, and 6 are derived from bovine fibroblast RNA, and lanes 3, 4, 7, and 8 are derived from human fibroblast RNA. The odd and even numbered lane of each pair indicate a $MgCl_2$ concentrations of 1.25 and 1.0 mM, respectively, in the PCR. (*D*) Alignment of partial cDNA clones and 5′-end of the genomic clone.

minidase was kindly provided by S. Tomatsu (Gifu University, Gifu, Japan).

Cloning of cDNA. The longest internal peptide sequence was used to synthesize three fully degenerate oligonucleotides, sense primer I and anti-sense primer II for reverse transcriptase-PCR (RT-PCR) and oligonucleotide III for hybridization, as indicated in Fig. 1*B*. Their sequences were as follows: I, 5′-GATCGAATTCATGGT(GATC)GG(GATC)AC(GATC)GG(GATC)AT-3′; II, 5′-CATGCTGCAGCA-(GATC)GCCAT(GATC)A(GA)(GATC)GC(GA)TA-3′; and III, 5′-AT(CTA)GG(GATC)CA(GA)AA(CT)GA(GA)GT-3′. Total RNA (14) from cultured bovine or human fibroblasts (2.5 mg) was reverse transcribed with 16 units of avian myeloblastosis virus reverse transcriptase (Promega) in the presence of 450 pmol of antisense primer II, 1.3 mM each dNTP, 1.3 mg of BSA per ml, 50 mM KCl, 3.3 mM $MgCl_2$, 20 units of RNase inhibitor (RNAguard, Pharmacia), and 20 mM Tris·HCl (pH 8.3); the reaction was incubated at 23°C for 10 min and then at 37°C for 60 min. The reverse transcription reaction was terminated by heating at 95°C for 3 min. After addition of the sense primer I, the reaction mixture was adjusted to 0.4 mM of each dNTP, 0.4 mg of BSA per ml, and varying concentrations of $MgCl_2$. After preheating at 72°C for 10 min, 2.5 units *Taq* DNA polymerase was added. The PCR was carried out sequentially for 1.5 min at 95°C, 1 min at 48°C, and 1.5 min at 72°C for 15 cycles and then for 1.5 min at 95°C, 1 min at 55°C, and 1.5 min at 72°C for 40 cycles. The RT-PCR products were separated by electrophoresis and subjected to Southern hybridization. A 89-bp bovine amplification product hybridized to the degenerate internal oligonucleotide III (Fig. 1*C*). Even though the human amplification product did not hybridize, both products were subcloned into T vector (15) and sequenced. The bovine product corresponded precisely to the tryptic peptide sequence, whereas the human product differed in one amino acid (a difference of two nucleotides in the internal region, explaining its failure to hybridize to oligonucleotide III). The extensive homology between the RT-PCR segments from the two species enabled us to shift at this stage to cloning the human cDNA.

A 41-mer, 5′-ATGGCCCCCGAGGGCATCAGCCAGAA-CGAAGTGGTCTACGC-3′, was synthesized from the sequence of the human RT-PCR product and used to screen a λgt11 human testis cDNA library (Clontech). About 3×10^6 phages were screened. Six positive clones, containing inserts from 1.3 kb to 1.5 kb, were subcloned into pBluescript II KS (+) and sequenced, but all were found to be missing the 5′-end.

5′-End rapid amplification of cDNA ends (RACE; ref. 16) was performed to extend sequence information further upstream. Total RNA (6 μg) was reverse transcribed as described above for RT-PCR, except for the substitution of 50 pmol of perfectly matched antisense primer. The reaction was carried out at 37°C for 10 min, 42°C for 60 min, and 52°C for 30 min. Excess 3′-primer was removed by centrifuge using Centricon 30 (Amicon). The product was incubated at 37°C for 10 min with 0.1 M potassium cacodylate (pH 7.2), 2 mM $CoCl_2$, 0.2 mM DTT, 0.2 mM dATP, and 15 units of terminal deoxynucleotidyl transferase (GIBCO/BRL), then heated at 65°C for 15 min. The reaction mixture was then diluted to 500 μl with TE (10 mM Tris·HCl/1 mM EDTA, pH 8.0) buffer, and 10-μl aliquots were used for amplification. Forty cycles of PCR were carried out at 95°C for 1 min, 57°C for 2 min, and 72°C for 3 min with 10 pmol of 5′-end adapter primer and 50 pmol each of 5′-end and 3′-end amplification primers. The RACE product was directly subcloned into T vector and sequenced.

The RACE segment, which extended the sequence by 0.6 kb, was used as a probe to rescreen the library. An 0.8-kb segment was isolated but still did not contain sequence corresponding to the amino terminus of the bovine or human enzyme; furthermore, its 5′-end was subsequently found to contain intronic sequence. The sequence corresponding to the amino terminus was finally located in a genomic clone, and RT-PCR of fibroblast RNA was performed in order to identify the exonic sequence. Fig. 1*D* shows the alignment of the clones.

Isolation of the *NAGLU* gene. DNA isolated from the normal human fibroblast line IMR 90 (Coriell Institute for Medical Research, Camden, NJ) was digested with *Eco*RI. A sample was used for Southern hybridization with the upstream 480 bp of the cDNA clone from screen 2 and the 1.5-kb cDNA from screen 1 (Fig. 1*D*). DNA of the size corresponding to the Southern blot results was used for constructing two individual λ DASHR II genomic libraries. Two clones hybridizing to the available cDNA segments were isolated from the libraries and subcloned into pBluescript II KS(+). The presence of the 5′-end of the coding sequence was verified by hybridization to degenerate oligonucleotides corresponding to the N-terminus of the enzyme. The longest clone, 12 kb, was used for the studies reported here. The most upstream 500 bases of the genomic sequence are from a cosmid kindly provided by Mary-Claire King and Lori Friedman (University of Washington, Seattle). Sequencing was performed manually by the cycle sequencing method, using kits from Perkin–Elmer and Promega, with the suppliers' instructions. Complete sequencing was performed to the junction with GenBank accession no. M84472 (nucleotide 3095 of our sequence, U43572, corresponds to nucleotide 19 of M84472); beyond that point, we sequenced introns only partially, the rest of intronic sequence coming from M84472.

Analysis of Mutations. Fibroblast lines GM 00156 and 02552 from Sanfilippo B patients and GM 04390 and 03348 from normal individuals were obtained from the Human Mutant Cell Repository, Coriell Institute for Medical Research. Fibroblast lines A and H were from the Hadassah Medical Center (Jerusalem, Israel) cell collection, and line IT 154 was provided by Paola Di Natale (University of Naples, Naples, Italy). Cultures were maintained as described (17).

For localizing mutations by single strand conformation polymorphism (SSCP; ref. 18), the coding sequence and exon–intron borders of *NAGLU* were first amplified in 11 segments, using primers and conditions for obtaining a single

product (A.S. and E.F.N., unpublished data). Genomic DNA from normal or Sanfilippo B cells (0.6 μg) was annealed to 50 pmol each of sense and antisense primers and amplified with 2.5 units of *Taq* DNA polymerase in 100 μl of 50 mM KCl, 10 mM Tris·HCl (pH 9.0), 0.1% Triton X-100, 0.2 mM dNTP, and 1.5 mM $MgCl_2$. The PCR was carried out for 35 cycles of 1 min denaturation at 95°C, 1 min annealing at the temperature indicated below, and 30 sec extension at 72°C. The hot start technique, using Ampliwax (Perkin–Elmer) was used according to the manufacturer's instructions. After amplification, 3 μl of the reaction was mixed with 3 μl of 95% formamide, 10 mM NaOH, 0.25% bromphenol blue, and 0.25% xylene cyanol, heated at 95°C for 4 min and chilled on ice. Electrophoresis was performed on an MDE (mutation detection enhancement) gel essentially as described by the manufacturer (AT Biochem, Malvern, PA). The electrophoresis was carried out at 3 W for 14–16 hr at ambient temperature. The double-stranded DNA had usually migrated out of the gel. The single-stranded DNA was visualized by silver staining, using the protocol recommended by Promega for staining sequencing gels. Except for samples A and H, those amplified segments that showed an abnormal SSCP pattern were separated from salt, primers, and unincorporated nucleotides by use of the QIAquick Spin PCR Purification kit (Qiagen) and subjected to cycle sequencing with *Taq* polymerase (Amplicycle Sequencing kit, Perkin–Elmer) and ^{33}P-dCTP. Both strands were sequenced.

The PCR primers that produced segments with abnormal migration on SSCP, and the corresponding annealing temperature, were as follows: sense primer, 5′-GCTGGCTAGTGACAGCCGCTT-3′, and antisense primer, 5′-CTGGTGCTGTTGGAAAGGGAT-3′, 54°C, for GM 00156 (R626X), GM 02552 (R643H), and A and H (R674H); sense primer, 5′-AAACCAGGAGCTGTAGAGAAGT-3′, and antisense primer, 5′-CTGCCTACCCCTACTGACATCT-3′, 54°C, for GM 02552 (R297X); sense primer, 5′-CCCTGCCCATCTGTTAGACT-3′, and antisense primer, 5′-GCACGTTGAAAGCACTTCTA-3′, 53°C, for IT 154 (503del10).

The procedures followed for cell lines A and H, the first to be analyzed, differed in that the sequence analysis that revealed the R674H mutation was first performed on RT-PCR amplified samples of total fibroblast RNA. The PCR-amplified segment of genomic DNA was not sequenced; instead, it was incubated with the restriction nuclease *Bsr*I in order to cleave the sensitive site created by the mutation. For allele-specific hybridization, PCR-amplified products from 100 genomic DNA samples (Hadassah Medical Center) were dot-blotted onto a nylon membrane and hybridized with γ-^{32}P end-labeled primers corresponding to the normal sequence, 5′-ACACCCCTCGCTGGCGGCT-3′ or the mutant sequence 5′-ACACCCCCTCACTGGCGGCT-3′. Boldface type indicates mutant sequence.

RESULTS AND DISCUSSION

Characterization of cDNA. Fig. 2 shows the nucleotide sequence of cDNA, determined from the overlapping cDNA and genomic clones, as well as the deduced amino acid sequence. The coding region consists of 743 aa. The translation start site was established by the absence of inframe methionine codons further upstream. Nucleotides −3 to +4 (accATGG) fit the Kozak model of a "strong" initiator site (19). A hydrophobic stretch of 23 aa, consistent with a signal peptide, extends to the amino terminus of the purified enzyme. Cleavage by signal peptidase at amino acids 20–23 conforms to von Heijne's predictions (20). Thus in contrast to many other lysosomal enzymes (21), α-*N*-acetylglucosaminidase of bovine testis and human liver is not processed by further proteolytic cleavage.

The deduced amino acid sequence has six potential N-glycosylation sites of the commonly used NXS/T structure at asparagine residues 261, 272, 435, 503, 526, and 532; an additional site, N513, might not be used because of an adjacent proline (22). The sequence also includes the rarely used NXC glycosylation signal (22) at N134. The amino acid sequences of eleven of 12 peptides from the bovine enzyme were over 50% identical to the corresponding amino acid sequences in the human enzyme, with an overall identity of 73% (see legend to Fig. 2); the remaining peptide from the bovine enzyme, which showed only 18% identity to any peptide in the human sequence, may have been derived from a nonconserved region or from an impurity in the enzyme preparation. The amino acid sequence of the human enzyme showed similarity only to very short stretches of sequences in the SwissProt data base (release 32, Dec. 1995).

The molecular size of the mature protein (720 aa or 80 kDa, not including carbohydrate residues) as well as its amino terminal sequence indicate that the two activity bands seen on SDS/PAGE of the bovine enzyme were the monomeric and dimeric forms of α-*N*-acetylglucosaminidase. The dimer, which we had used for amino acid sequence analysis, apparently does not readily dissociate even under denaturing conditions. Others have also shown enzyme of ≈80 kDa as well as larger forms likely to be dimers and tetramers (12, 23, 24).

Chromosomal Locus and Architecture of the *NAGLU* Gene. Most of the cDNA sequence had been found in the 5′-flanking region of *EDH17B*, the gene encoding 17-β-hydroxysteroid dehydrogenase, GenBank accession no. M84472 (25). Because the locus of *EDH17B* is chromosome 17q21 (26), this is also the locus of *NAGLU*. A longer sequence flanking *EDH17B*, encompassing the entire *NAGLU* gene (GenBank accession no. U34879), became available during preparation of this manuscript.

The architecture of the *NAGLU* gene is shown in Fig. 3. The placement of introns was determined by comparison with the cDNA except in the case of intron 1, which was identified by RT-PCR of the corresponding section of fibroblast RNA. The *NAGLU* gene, interrupted by five introns, is 8.2 kb long from translation start to polyadenylylation site. The first exon is indicated provisionally as containing an additional 0.3 kb of untranslated sequence, based on primer extension studies that showed an apparent transcription start site 332 and 321 nucleotides upstream of the initiating methionine (H.G.Z. and E.F.N., unpublished results). But because that region contains neither TATA box nor SP1 sites and is not particularly G+C-rich, the untranslated region may extend even further upstream.

Mutations in Patients with Sanfilippo B Syndrome. Several mutations were identified by preliminary screening of the *NAGLU* gene by PCR-SSCP (18) followed by sequence analysis of the amplified segments that had altered mobility. Cell line IT 154 proved homozygous for a 10-nt deletion starting with nucleotide 503 (Fig. 4*A*); the deletion results in a frameshift and predicts termination 14 codons later. Cell line GM 00156 proved homozygous for a C→T transition at nucleotide 1876, resulting in a termination codon at position 626 instead of the normal arginine (Fig. 4*B*). Cell line GM 02552 proved to be compound heterozygous for a C→T transition at nucleotide 889, resulting in a stop in lieu of arginine at codon 297 (Fig. 4*C*); the other allele had a G→A transition and a substitution of histidine for arginine at codon 643 (data not shown). Finally cell lines designated A and H were first shown by RT-PCR screening to have a G→A transition at position 2021, resulting in a substitution of histidine for arginine at codon 674. Homozygosity was demonstrated by restriction nuclease analysis of the corresponding genomic segment after PCR amplification; Fig. 4*D* shows essentially complete cleavage of the normal 262-bp segment by *Bsr*I at the site created by the mutation.

The base substitutions causing replacement of arginine by a termination codon or by histidine, R297X, R626X, R643H,

```
1    ATGGAGGCGGTGGCGGTGGCCGCGGCGGTGGGGGTCCTTCTCCTGGCCGGGGCCGGGGGCGCGGCAGGCGACGAGGCCCGGGAGGCGGCGGCCGTGCGGGCGCTCGTGGCCCGGCTGCTG
1    M  E  A  V  A  V  A  A  A  V  G  V  L  L  L  A  G  A  G  G  A  A  G  D  E  A  R  E  A  A  A  V  R  A  L  V  A  R  L  L

121  GGGCCAGGCCCCGCGGCCGACTTCTCCGTGTCGGTGGAGCGCGCTCTGGCTGCCAAGCCGGGCTTGGACACCTACAGCCTGGGCGGCGGCGGCGCGGCGCGCGTGCGGGTGCGCGGCTCC
41   G  P  G  P  A  A  D  F  S  V  S  V  E  R  A  L  A  A  K  P  G  L  D  T  Y  S  L  G  G  G  G  A  A  R  V  R  V  R  G  S

241  ACGGGCGTGGCGGCCGCCGCGGGGCTGCACCGCTACCTGCGCGACTTCTGTGGCTGCCACGTGGCCTGGTCCGGCTCTCAGCTGCGCCTGCCGCGGCCACTGCCAGCCGTGCCGGGGGAG
81   T  G  V  A  A  A  A  G  L  H  R  Y  L  R  D  F  C  G  C  H  V  A  W  S  G  S  Q  L  R  L  P  R  P  L  P  A  V  P  G  E

361  CTGACCGAGGCCACGCCCAACAGGTACCGCTATTACCAGAATGTGTGCACGCAAAGCTACTCCTTCGTGTGGTGGGACTGGGCCCGCTGGGAGCGAGAGATAGACTGGATGGCGCTGAAT
121  L  T  E  A  T  P  N  R  Y  R  Y  Y  Q  N  V  C  T  Q  S  Y  S  F  V  W  W  D  W  A  R  W  E  R  E  I  D  W  M  A  L  N

481  GGCATCAACCTGGCACTGGCCTGGAGCGGCCAGGAGGCCATCTGGCAGCGGGTGTACCTGGCCTTGGGCCTGACCCAGGCAGAGATCAATGAGTTCTTTACTGGTCCTGCCTTCCTGGCC
161  G  I  N  L  A  L  A  W  S  G  Q  E  A  I  W  Q  R  V  Y  L  A  L  G  L  T  Q  A  E  I  N  E  F  F  T  G  P  A  F  L  A

601  TGGGGGCGAATGGGCAACCTGCACACCTGGGATGGCCCCCTGCCCCCCTCCTGGCACATCAAGCAGCTTTACCTGCAGCACCGGGTCCTGGACCAGATGCGCTCCTTCGGCATGACCCCA
201  W  G  R  M  G  N  L  H  T  W  D  G  P  L  P  P  S  W  H  I  K  Q  L  Y  L  Q  H  R  V  L  D  Q  M  R  S  F  G  M  T  P

721  GTGCTGCCTGCATTCGCGGGGCATGTTCCCGAGGCTGTCACCAGGGTGTTCCCTCAGGTCAATGTCACGAAGATGGGCAGTTGGGGCCACTTTAACTGTTCCTACTCCTGCTCCTTCCTT
241  V  L  P  A  F  A  G  H  V  P  E  A  V  T  R  V  F  P  Q  V  N  V  T  K  M  G  S  W  G  H  F  N  C  S  Y  S  C  S  F  L

841  CTGGCTCCGGAAGACCCCATATTCCCCATCATCGGGAGCCTCTTCCTGCGAGAGCTGATCAAAGAGTTTGGCACAGACCACATCTATGGGGCCGACACTTTCAATGAGATGCAGCCACCT
281  L  A  P  E  D  P  I  F  P  I  I  G  S  L  F  L  R  E  L  I  K  E  F  G  T  D  H  I  Y  G  A  D  T  F  N  E  M  Q  P  P

961  TCCTCAGAGCCCTCCTACCTTGCCGCAGCCACCACTGCCGTCTATGAGGCCATGACTGCAGTGGATACTGAGGCTGTGTGGCTGCTCCAAGGCTGGCTCTTCCAGCACCAGCCGCAGTTC
321  S  S  E  P  S  Y  L  A  A  A  T  T  A  V  Y  E  A  M  T  A  V  D  T  E  A  V  W  L  L  Q  G  W  L  F  Q  H  Q  P  Q  F

1081 TGGGGGCCCGCCCAGATCAGGGCTGTGCTGGGAGCTGTGCCCCGTGGCCGCCTCCTGGTTCTGGACCTGTTTGCTGAGAGCCAGCCTGTGTATACCCGCACTGCCTCCTTCCAGGGCCAG
361  W  G  P  A  Q  I  R  A  V  L  G  A  V  P  R  G  R  L  L  V  L  D  L  F  A  E  S  Q  P  V  Y  T  R  T  A  S  F  Q  G  Q

1201 CCCTTCATCTGGTGCATGCTGCACAACTTTGGGGGAAACCATGGTCTTTTTGGAGCCCTAGAGGCTGTGAACGGAGGCCCAGAAGCTGCCCGCCTCTTCCCCAACTCCACCATGGTAGGC
401  P  F  I  W  C  M  L  H  N  F  G  G  N  H  G  L  F  G  A  L  E  A  V  N  G  G  P  E  A  A  R  L  F  P  N  S  T  M  V  G

1321 ACGGGCATGGCCCCCGAGGGCATCAGCCAGAACGAAGTGGTCTATTCCCTCATGGCTGAGCTGGGCTGGCGAAAGGACCCAGTGCCAGATTTGGCAGCCTGGGTGACCAGCTTTGCCGCC
441  T  G  M  A  P  E  G  I  S  Q  N  E  V  V  Y  S  L  M  A  E  L  G  W  R  K  D  P  V  P  D  L  A  A  W  V  T  S  F  A  A

1441 CGGCGGTATGGGGTCTCCCACCCGGACGCAGGGGCAGCGTGGAGGCTACTGCTCCGGAGTGTGTACAACTGCTCCGGGGAGGCCTGCAGGGGCCACAATCGTAGCCCGCTGGTCAGGCGG
481  R  R  Y  G  V  S  H  P  D  A  G  A  A  W  R  L  L  L  R  S  V  Y  N  C  S  G  E  A  C  R  G  H  N  R  S  P  L  V  R  R

1561 CCGTCCCTACAGATGAATACCAGCATCTGGTACAACCGATCTGATGTGTTTGAGGCCTGGCGGCTGCTGCTCACATCTGCTCCCTCCCTGGCCACCAGCCCCGCCTTCCGCTACGACCTG
521  P  S  L  Q  M  N  T  S  I  W  Y  N  R  S  D  V  F  E  A  W  R  L  L  L  T  S  A  P  S  L  A  T  S  P  A  F  R  Y  D  L

1681 CTGGACCTCACTCGGCAGGCAGTGCAGGAGCTGGTCAGCTTGTACTATGAGGAGGCAAGAAGCGCCTACCTGAGCAAGGAGCTGGCCTCCCTGTTGAGGGCTGGAGGCGTCCTGGCCTAT
561  L  D  L  T  R  Q  A  V  Q  E  L  V  S  L  Y  Y  E  E  A  R  S  A  Y  L  S  K  E  L  A  S  L  L  R  A  G  G  V  L  A  Y

1801 GAGCTGCTGCCGGCACTGGACGAGGTGCTGGCTAGTGACAGCCGCTTCTTGCTGGGCAGCTGGCTAGAGCAGGCCCGAGCAGCGGCAGTCAGTGAGGCCGAGGCCGATTTCTACGAGCAG
601  E  L  L  P  A  L  D  E  V  L  A  S  D  S  R  F  L  L  G  S  W  L  E  Q  A  R  A  A  A  V  S  E  A  E  A  D  F  Y  E  Q

1921 AACAGCCGCTACCAGCTGACCTTGTGGGGGCCAGAAGGCAACATCCTGGACTATGCCAACAAGCAGCTGGCGGGGTTGGTGGCCAACTACTACACCCCTCGCTGGCGGCTTTTCCTGGAG
641  N  S  R  Y  Q  L  T  L  W  G  P  E  G  N  I  L  D  Y  A  N  K  Q  L  A  G  L  V  A  N  Y  Y  T  P  R  W  R  L  F  L  E

2041 GCGCTGGTTGACAGTGTGGCCCAGGGCATCCCTTTCCAACAGCACCAGTTTGACAAAAATGTCTTCCAACTGGAGCAGGCCTTCGTTCTCAGCAAGCAGAGGTACCCCAGCCAGCCGCGA
681  A  L  V  D  S  V  A  Q  G  I  P  F  Q  Q  H  Q  F  D  K  N  V  F  Q  L  E  Q  A  F  V  L  S  K  Q  R  Y  P  S  Q  P  R

2161 GGAGACACTGTGGACCTGGCCAAGAAGATCTTCCTCAAATATTACCCCGGCTGGGTGGCCGGCTCTTGGTGAtagattcgccaccactgggccttgttttccgctaattccagggcagat
721  G  D  T  V  D  L  A  K  K  I  F  L  K  Y  Y  P  G  W  V  A  G  S  W  *

2281 tccagggcccagagctggacagacatcacaggataacccaggcctgggaggaggccccacggcctgctggtggggtctgacctgggggattggagggaaatgacctgccctccaccacc
2401 acccaaagtgtgggattaaagtactgttttctttccacttaaa(a)15
```

FIG. 2. Nucleotide and deduced amino acid sequences of cDNA encoding human α-*N*-acetylglucosaminidase. An asterisk denotes the amino terminus of purified α-*N*-acetylglucosaminidase. Arrows indicate the position of introns. Potential NXS/T glycosylation sites are underlined with a heavy bar, an infrequently used NXSP site and a nonstandard NXC site are underlined with a wavy line. The polyadenylylation signal, ATTAAA, is underlined with a fine line. The following sequences correspond to the peptides from bovine α-*N*-acetylglucosaminidase (the numbers indicate first and last amino acid of the human sequence and degree of identity with the bovine sequence): 24–48, 80%; 110–126, 53%; 302–314, 77%; 367–377, 55%; 378–393, 94%; 432–461, 86%; 458–479, 73%; 662–675, 72%; 677–694, 67%; and 700–713, 64%.

and R674H, all occur at CpG sites, known to be mutagenic hotspots (reviewed in ref. 27). The 10-nt deletion occurs at a direct repeat of a tetranucleotide, GGAG, and may be the result of slipped mispairing during DNA replication (27).

The homozygous R674H mutation was shown not to be a polymorphism by allele-specific nucleotide analysis of genomic DNA from 47 individuals of the same ethnic group (Arab) and 53 individuals of other ethnic groups (data not shown); that mutation may therefore be presumed causal to the disease in patients A and H. Whether patients A and H were related to each other cannot be ascertained, because contact with the families has been lost. The consequences to enzyme activity of R643H and of several other missense mutations found in patients with Sanfilippo syndrome type B (A.S. and E.F.N., unpublished data) may need to be verified by expression of mutagenized cDNA. On the other hand, the homozygous deletion (503del10) with frameshift and premature termination, as well as the homozygous nonsense mutation (R626X), must be considered causal to the enzyme deficiency and the ensuing disease. As such, they provide confirmation for the identity of the *NAGLU* gene.

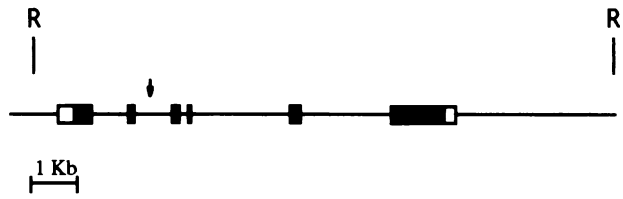

FIG. 3. Architecture of the human *NAGLU* gene. Boxes indicate exons, with the filled area indicating the coding sequence. The first and last nucleotide of each exon in this gene, GenBank accession no. U43572, are: I, 1085–1799; II, 2542–2689; III, 3482–3628; IV, 3813–3898; V, 6091–6347; and VI, 8167–9588. *Eco*RI sites are indicated with an R. The arrow indicates the start of the overlap with GenBank accession no. M84472.

This work was supported in part by a National Institutes of Health Grant NS22376 (E.F.N.) and by fellowships from the Wenner–Gren Center Foundation and the Hellmuth Hertz Foundation (A.S.). The University of California at Los Angeles Microsequencing Facility is supported in part by a Cancer Center Support Grant from the National

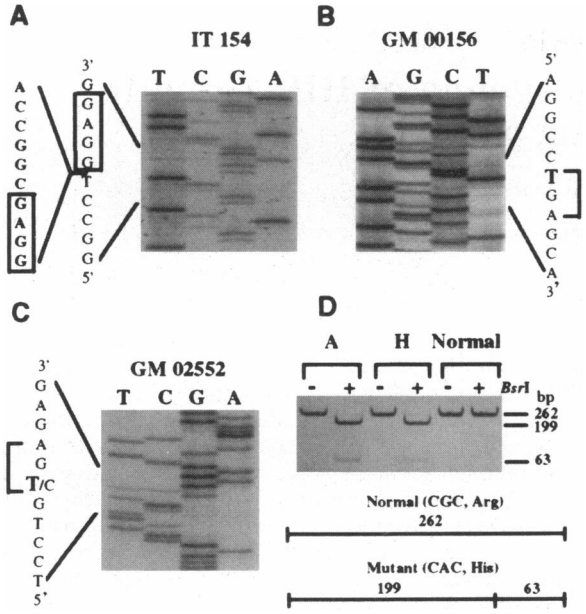

FIG. 4. Analysis of mutations in cell lines derived from patients with Sanfilippo syndrome type B. *A*, *B*, and *C* show the mutant sequence in genomic DNA of cell lines IT 154, GM 00156, and GM 02552, respectively, with the deviation from the normal shown in bold (deletion in *A* and substitution in *B* and *C*). Boxes in *A* indicate the tetranucleotide repeat. Brackets in *B* and *C* show the stop codon created by the mutation. (*D*) A homozygous gain of a *Bsr*I restriction site in the genomic DNA of cells from patients A and H.

Cancer Institute (CA 16042–20) to the Jonsson Comprehensive Cancer Center. The authors thank Rosa Lopez, Jennifer Rennecker, and Hui-Zhi Zhao for excellent assistance in various phases of this work, Dr. Shunji Tomatsu (Gifu University, Gifu, Japan) for providing the amino-terminal sequence of human liver α-*N*-acetylglucosaminidase, Dr. Paola Di Natale (University of Naples, Naples, Italy) for providing the cell line IT 154, and Drs. Mary-Claire King and Lori Friedman (University of Washington, Seattle) for a cosmid containing the *NAGLU* gene.

1. Neufeld, E. F. & Muenzer, J. (1995) in *The Metabolic and Molecular Bases of Inherited Disease*, eds. Scriver, C. R., Beaudet, A. L., Sly, W. S., & Valle, D. (McGraw–Hill, New York), pp. 2465–2494.
2. Murata, R., Nakajima, S., Tanaka, A., Miyagi, N., Matsuoka, O., Kogame, S. & Inoue, Y. (1989) *Am. J. Neuroradiol.* **10,** 1165–1170.
3. Van de Kamp, J. J. P., Neirmeijer, M. F., von Figura, K. & Geisberts M. A. H (1981) *Clin. Genet.* **20,** 152–160.
4. Di Natale, P. (1991) *J. Inherited Metab. Dis.* **14,** 23–28.
5. Robertson, D. A., Freeman, C., Morris C. P. & Hopwood, J. J. (1992) *Biochem. J.* **288,** 539–544.
6. Scott, H. S., Blanch L., Go, X.-H., Freeman, C., Orsborn, A., Baker, E., Sutherland, G., Morris, C. P. & Hopwood, J. J. (1995) *Nature Genet.* **11,** 465–467.
7. Thompson, J. N., Jones, M. Z., Dawson, G. & Huffman P. S. (1992) *J. Inherited Metab. Dis.* **15,** 560–578.
8. Zhao, H. G., Lopez, R., Rennecker, J. & Neufeld, E. F. (1994) *Am. J. Hum. Genet.* **55,** A252 (abstr.).
9. Zhao, H. G., Li, H. H., Schmidtchen, A., Bach. G. & Neufeld, E. F. (1995) *Am. J. Hum. Genet.* **57,** A185 (abstr.).
10. Marsh, J. & Fensom, A. H. (1985) *Clin. Genet.* **27,** 258–262.
11. Stoscheck, C. (1990) *Methods Enzymol.* **182,** 50–65.
12. Sasaki, T., Sukegawa, K., Masue, M., Fukuda, S., Tomatsu, S. & Orii, T. (1991) *J. Biochem.* **110,** 842–846.
13. Ferrara, P., Rosenfeld, J., Guillemot, J. C. & Capdevielle, J. (1993) *Tech. Protein Chem.* **4,** 379–389.
14. Chomczynski, P. & Sacchi, N. (1987) *Anal. Biochem.* **162,** 156–159.
15. Marchuk, D., Drumm, M., Saulino, A. & Collins, F. S. (1991) *Nucleic Acids Res.* **19,** 1154.
16. Frohman, M. A., Dush, M. K. & Martin, G. R. (1988) *Proc. Natl. Acad. Sci. USA* **85,** 8998–9002.
17. Paw, B. H., Wood, L. C. & Neufeld, E. F. (1991) *Am. J. Hum. Genet.* **48,** 1139–1146.
18. Orita, M., Suzuki Y, Sekiya T. & Hayashi, K. (1989) *Genomics* **5,** 874–879.
19. Kozak, M. (1986) *Cell* **44,** 283–292.
20. Von Heijne, G. (1986) *Nucleic Acids Res.* **14,** 4683–4690.
21. Neufeld, E. F. (1991) *Annu. Rev. Biochem.* **60,** 257–279.
22. Gavel, Y. & von Heijne, G.(1990) *Protein Eng.* **3,** 433–442.
23. Von Figura, K. (1977) *Eur. J. Biochem.* **80,** 525–533.
24. Di Natale, P., Salvatore, D., Daniele, A. & Bonatti, S. (1985) *Enzyme* **33,** 75–83.
25. Peltoketo, H., Isomaa, V. & Vihko, R.(1992) *Eur. J. Biochem.* **209,** 495–466.
26. Friedman, L. S., Lynch, E. D. & King, M. C. (1993) *Hum. Mol. Genet.* **2,** 821.
27. Cooper D. N., Krawczak, M. & Antonarakis, S. E. M. (1995) in *The Metabolic and Molecular Bases of Inherited Disease*, eds. Scriver, C. R., Beaudet, A. L., Sly, W. S. & Valle, D. (McGraw–Hill, New York), pp. 259–292.