Cloning and Expression of the Human Gene for Transaldolase

A NOVEL HIGHLY REPEETITIVE ELEMENT CONSTITUTES AN INTEGRAL PART OF THE CODING SEQUENCE*

(Received for publication, June 8, 1993, and in revised form, October 1, 1993)

Katalin Banki, David Halladay, and Andras Perl†

From the Departments of Medicine, Microbiology and Immunology, and Pathology, State University of New York Health Science Center, College of Medicine, Syracuse, New York 13210

A novel highly repetitive retrotransposable element was cloned based on a limited sequence homology to the human T-cell leukemia virus and a related endogenous retroviral sequence, HRES-1. This repetitive element was found to constitute an integral part of the coding sequence of the human gene for transaldolase. In comparison with the intronless yeast gene, structural analysis of the human transaldolase genomic locus revealed that the human gene is comprised of five exons, second and third of which uniquely developed by insertion of a retrotransposable element. The 1329-base pair full-length cDNA, clone 42/4-1, contains an open reading frame coding for a protein of 336 amino acids with a predicted molecular mass of 38 kDa. This protein shows a 58% overall sequence homology with the 37-kDa yeast transaldolase. Antibodies raised against a 22-kDa recombinant polypeptide expressed from a 474-base pair 5′ fragment of clone 42/4-1, containing repetitive exons 2 and 3, cross-reacted with yeast transaldolase and recognized the 38-kDa native human protein. Detection of a retrotransposon in the coding sequence of the human transaldolase gene demonstrates the importance of these repetitive elements in evolution of the eukaryotic genome.

Retrotransposable elements (RTE)* are highly repetitive and they make up as much as 10% of the eukaryotic genome (1). These elements multiply through RNA intermediates by reverse transcription from RNA to DNA. The normal human genome contains a complex variety of RTE. The viral superfamily of RTE comprises a number of different endogenous retroviral sequences (ERS) which are related to known animal or human retroviruses. While some ERSs are represented in a single copy per haploid genome, others are highly repetitive and occur at a frequency of up to 1000 copies per haploid genome similar to the larger family of nonviral RTEs (2). Two members of the nonviral superfamily, the short interspersed elements (90–400 bp in size) and the long interspersed elements (up to 7000 bp in size) are present in the genome in copy numbers in excess of 100,000 (3). Retrotransposition of these highly repetitive elements is considered a major factor in the shaping and reorganization of the genome (1). The present study was initiated to identify potentially novel transcriptionally active RTEs. The data document cloning and sequencing of a so far uncharacterized highly repetitive RTE based on a limited sequence homology to the human T-cell leukemia virus (4) and a related ERS, HRES-1 (5). This novel RTE, which constitutes an integral part of the coding sequence of the human gene for transaldolase, is termed as TARE, transaldolase-associated repetitive element. The transaldolase enzyme, which catalyzes the transfer of a C3 fragment corresponding to dihydroxyacetone in the pentose phosphate pathway, was originally described in yeast (6). This metabolic pathway provides ribose 5-phosphate for the synthesis of nucleic acids and NADPH as a reducing agent (7). The present study describes cloning, sequencing, and expression of the human transaldolase gene, the 5′ region of which is encoded by a highly repetitive element.

MATERIALS AND METHODS

Screening of cDNA and Genomic Libraries—A Agt10 cDNA library of HL-60 human myelomonocytic leukemia cells, which were found to express HTLV-I and HRES-1-related transcripts, was screened under low stringency conditions (2 x SSC, 55 °C) with pMA1, a long terminal repeat, and a gag region-containing HTLV-I probe. Positive clones were identified by hybridization with a 32P-labeled pMA1 probe. Hybridization was carried out in 6 x SSC (1 x SSC is 0.15 M NaCl plus 0.015 M sodium citrate), 1 x Denhardt's solution (8), 1 x Denhardt's solution (8), 1 x Denhardt's solution (8), and 0.5% SDS. After overnight hybridizations, the filters were washed under low stringency conditions in 2 x SSC and 0.1% SDS at 55 °C for 2 h. Such screening would identify sequences containing at least 12 identical nucleotides in contiguity, assuming a 50% GC content (5, 9). A human lymphocyte genomic DNA library was prepared in λ DASH phage (Stratagene, La Jolla, CA) and screened with 42/4 and 42/1 cDNA clones under high stringency conditions as earlier described (5).

Southern Blot Hybridizations—High molecular weight genomic DNA was isolated from peripheral blood lymphocytes, digested to completion with restriction endonucleases (Life Technologies, Inc.), separated by electrophoresis in 0.7% agarose gels, denatured, and transferred to nylon membranes by Southern blotting as described earlier (5). The blots were hybridized as indicated above except for the addition of 100 μg/ml boiled salmon sperm DNA. Genomic blots were washed under high stringency conditions (9) in 0.1 x SSC, 0.1% SDS for 90 min at 65 °C.

Northern Blot Analysis—Total RNA was extracted by the RNAzol method (10). Poly(A)+ RNA was isolated by binding to poly(U)-Sepharose column (Life Technologies, Inc.), fractionated in 1% glyoxal gels, and transferred to nylon membranes (8). Hybridization and washing were done under high stringency conditions.

Hybridization Probes—The HTLV-I-specific pMA1 probe used in these experiments is a 1.5-kb EcoRI-ClaI fragment which contains the 5′ long terminal repeat and the entire gag gene of HTLV-I (4). As an HRES-1-specific probe, a 5.5-kb HindIII fragment of HRES-1/1 or its HTLV-related 430-bp EcoRI-Smal subfragment was used (5). TALH-
specific cDNA probes include the full-length 1329-nucleotide-long 4/2-4/1 clone and its 5' 474-bp EcoRI fragment, termed 4/2, and 3' 827-bp EcoRI fragment, termed 4/1. All cDNA clones were propagated in the Bluescript KS+ vector (Stratagene, La Jolla, CA). A human β-actin cDNA probe, pFHS (1.5-kb XhoI fragment) was used as a transcriptional control (11). All DNA probes were purified from vector in preparative agarose gels and labeled with [32P]dCTP by random oligomer priming (12).

**DNA Sequencing**—Double-stranded plasmid DNA was denatured by NaOH and sequenced in both strands by the chain termination method using the Sequenase System (United States Biochemical Corp.). The obtained sequence was analyzed with the University of Wisconsin Genetics Computer Group (UWGCG) software and submitted to GenBank (accession number: L19437).

**Prokaryotic Expression of Recombinant Protein**—A 474-bp EcoRI fragment, that is, the 4/2 section of the 4/2-4/1 cDNA, which contains an uninterrupted open reading frame, was ligated into the pEV plasmid vector and expressed in Escherichia coli (pRK2484cEts) (13). Construction of the vector is such that an ATG codon is placed before the codon corresponding to the first amino acid of the mature gene product. Bacterial cultures were grown at 30 °C in M9 medium with 0.5% glucose, 10 mM MgSO4, 10 μg/ml thiamine, 10 μg/ml thymine, 10 μg/ml proline, 1 μM CaCl2, 0.5% casamino acids. Expression of the recombinant protein was induced by growing the bacteria at 42 °C. Bacterial lysates were resuspended in 1/10 volume of Laemmli buffer and analyzed by SDS-PAGE (14).

**Antibodies**—New Zealand White rabbits were immunized on two separate occasions 3 weeks apart with 500 μg of gel-purified 22-kDa recombinant protein encoded by cDNA clone 4/2. Specific reactivity of immune sera 169 and 170 to the 22-kDa protein was evaluated by Western blot analysis using preimmune rabbit sera as an negative control.

**Western Blot Analysis**—Protein lysates from cells and tissues were resuspended in Laemmli buffer (14) at a total protein concentration of 4 mg/ml as determined by the Lowry method (15) using bovine serum albumin as standard. 40 μg of total protein in 10 μl/well was separated by SDS-PAGE and electrophoretically transferred to nitrocellulose (16). Nitrocellulose blots were hybridized to HRES-1 and 412 cDNA probes. The 474-bp HRES-l/HTLV-I-related EcoRI fragment, that is, the 4/2 section of the 4/2-4/1 cDNA, which contains an open reading frame capable of encoding a 336-amino acid-long protein, was sequenced in both strands by the chain termination method (18). The 474-bp HRES-l/HTLV-I-related EcoRI fragment was found to comprise the 5' end of the full-length 1329-bp cDNA clone, designated as 4/2-4/1. The 4/2-4/1 clone was sequenced in both strands by the chain termination method (2). DNA sequence homologies with HRES-1 and HTLV-I are confined to GC-rich sections within the repetitive region (Fig. 3) explaining the cross-hybridizations between these sequences. Separate high stringency Southern blot hybridizations of the 5' 474-bp EcoRI fragment and 3' 474-bp EcoRI fragment regions of the full-length 4/2-4/1 cDNA to human genomic DNA demonstrated that the 5' region is highly repetitive, whereas the 3' region is represented in a single copy per haploid genome (Fig. 4). The 1.3-kb 4/2-4/1 cDNA Encodes the Human Gene for Transaldolase—The 4/2-4/1 cDNA clone contains an open reading frame capable of encoding a 336-amino acid-long protein (Fig. 5). Computer search of GenBank, NBRF, and EMBL revealed that the 4/2-4/1 cDNA is different from any known human or viral sequence. However, a 52% DNA sequence homology was noted between clone 4/2-4/1 and the yeast transaldolase gene. The translated amino acid sequence of cDNA 4/2-4/1 showed a 58% overall homology with the transaldolase protein (Fig. 5).

**RESULTS**

Cloning and Sequencing of HTLV-I and HRES-1-related 4/2-4/1 cDNA—HRES-1 is a single-copy ERS which is transcribed into a 6-kb mRNA (5). However, lowering the stringency of hybridization of HRES-1 to human genomic DNA and total cellular RNA, a series of additional highly abundant DNA fragments and transcripts have been detected by Southern and Northern blot analysis, respectively (data not shown). HRES-1 exhibits no significant sequence homology with the previously characterized highly repetitive short interspersed elements and long interspersed elements. In order to identify potentially novel transcriptionally active RTEs low stringency screening of a human myelomonocytic cell line (HL-60) cDNA library with HRES-1 (HRES-1/1) and HTLV-I-specific probes (pMAI) was undertaken. The level of stringency applied was expected to detect sequences containing at least 12 complementary nucleotides in contiguity, assuming a 50% GC content (5, 9). Positive cDNA clones were digested with EcoRI and further examined by Southern blot hybridization. Under the stringency of hybridization applied, all cDNA clones contained an EcoRI insert of identical size (Fig. 1). Intensive cross-hybridization was noted between the uniform cDNA sequences, termed 4/2, and HTLV-I. A relatively lower intensity cross-hybridization between HRES-1 and 4/2 was also detected. The 4/2 cDNA probe annealed to a 1.3-kb mRNA species in poly(A)+ RNA from normal lymphocytes (not shown). Subsequently, a 1.3-kb full-length cDNA was cloned by re-screening the HL-60 cDNA library with the 4/2 probe. The 474-bp HRES-l/HTLV-I-related EcoRI fragment was found to comprise the 5' end of the full-length 1329-bp cDNA clone, designated as 4/2-4/1. The 4/2-4/1 clone was sequenced in both strands by the chain termination method (2). DNA sequence homologies with HRES-1 and HTLV-I are confined to GC-rich sections within the repetitive region (Fig. 3) explaining the cross-hybridizations between these sequences. Separate high stringency Southern blot hybridizations of the 5' 474-bp EcoRI fragment and 3' 474-bp EcoRI fragment regions of the full-length 4/2-4/1 cDNA to human genomic DNA demonstrated that the 5' region is highly repetitive, whereas the 3' region is represented in a single copy per haploid genome (Fig. 4).
Retrotransposable Element in Human Transaldolase Gene

FIG. 2. DNA sequence of a 1329-bp full-length cDNA clone 4/2-4/1. EcoRI restriction sites are underlined. Translated amino acid sequence of a 1008-nucleotide-long open reading frame is indicated. The polyadenylation site is indicated by a double underline. Exon start sites are marked (\( \hat{D} \)) based on comparative analysis with corresponding genomic clones (not shown). Upstream from the first methionine codon the cDNA contains an uninterrupted open reading frame starting at position three (frame c), allowing expression of a 157 amino acid long protein from the 5' 474-bp EcoRI fragment (Fig. 6).

FIG. 3. Nucleotide sequence homologies between the repetitive section, 5' 474-bp EcoRI fragment, of cDNA 4/2-4/1 and HTLV-I (4) and HRES-1(5), respectively, using the BESTFIT program of UWGCG software.

expression vector (13) and expressed in E. coli. Bacterial cell lysates were analyzed by SDS-PAGE (Fig. 6). The 4/2 cDNA-encoded 22-kDa recombinant protein was gel-purified and used to generate antisera in two rabbits. Antisera from both rabbits, Abs 169 and 170, are specific for the 22-kDa recombinant protein. A native human protein in various cell lines and tissues was identified by Abs 169 and 170 as a 38-kDa doublet (Fig. 7).

Abs 169 and 170, Which Are Highly Specific for a 38-kDa Human Protein, Show Immunoreactivity with Dansaldolase of the Yeast—To evaluate whether cDNA 4/2-4/1 corresponds to the human transaldolase gene, reactivity of Abs 169 and 170 with purified yeast transaldolase was evaluated. As shown in Fig. 7, Ab 170 showed specific immunoreactivity to the 38-kDa transaldolase protein from yeast. Ab 170 did not react with contaminating proteins, and the preimmune rabbit serum displayed no binding to any protein in the yeast transaldolase extract.

Co-purification of Dansaldolase Activity and Immunoreactivity to Ab 170 from Human Lymphocytes—Transaldolase was purified from human peripheral blood lymphocytes by sequential precipitation with acetone (18). Enzyme activity was measured by the transfer of the dihydroxyacetone three-carbon unit from donor D-fructose 6-phosphate to the acceptor D-erythrose 4-phosphate (19). A 16-fold enhancement of transaldolase specific activity in fraction 3 correlated with enrichment of the 38-kDa protein species as detected by Ab 170 (Fig. 7).

Enxon-Intron Organization and Mapping of a Retrotransposable Element, TARE, in the TU-H Genomic Locus—A human lymphocyte genomic DNA library was screened with 4/2- and 4/1-specific probes. Although TARE-containing genomic clones were identified by presence of 4/2, but not 4/1-specific DNA and variable flanking sequences, TAL-H genomic clones were selected based on the presence of adjacent 4/2- and 4/1-specific fragments. Exon-intron boundaries of the TAL-H gene locus were determined by sequence analysis of overlapping lambda DASH clones. Comparative analysis of genomic and cDNA sequences revealed that unlike the intronless TAL-Y locus the TAL-H gene locus contains five exons which span a chromosomal region of approximately 50 kb (Fig. 8). The 4/2 fragment contains four of the TAL-H exons. Exon boundaries within the background.
Retrotransposable Element in Human Transaldolase Gene

Fig. 5. Sequence homologies between the translated amino acid sequence of cDNA 4/2-4/1 (TAL-H) and the yeast transaldolase protein (TAL-Y). Potential phosphorylation sites are indicated in TAL-H. Calmodulin-dependent protein kinase (Calbin) and CAMP-dependent protein kinase sites are marked with asterisks.

Fig. 6. Expression of a 22-kDa protein from the 5' 4/2 (474-bp EcoRI) fragment of the 4/2-4/1 cDNA cloned into the vector pEV-vrfl (13). Vectors pEV-vrfl, pEV-vrfl, and pEV-vrfl allow expression of three alternative reading frames. pRR245clia bacteria were transformed with plasmids. Production of a 22-kDa recombinant protein encoded by the human 4/2 cDNA insert was induced by growing the cells at 42 °C for 2 h (lane 5). Cells were lysed in SDS-PAGE sample buffer (14). The lysates were electrophoresed in a 12% SDS-polyacrylamide gel and stained with Coomassie Brilliant Blue. Negative lanes are: lane 1, pEV-vrfl with no insert grown at 42 °C; lane 2, pEV-vrfl with 4/2 cDNA grown at 40 °C; lane 3, pEV-vrfl with 4/2 cDNA grown at 42 °C; lane 4, pEV-vrfl with 4/2 cDNA grown at 30 °C. i.e. uninduced control for lane 5; lane 6, pEV-vrfl with 4/2 cDNA grown at 40 °C; lane 7, pEV-vrfl with 4/2 cDNA grown at 42 °C.

cDNA are indicated in Fig. 2, while location of each exon within the TAL-H locus is shown in Fig. 8. Although the transaldolase enzyme was originally described (6), its structural gene was recently cloned from the yeast (TAL-Y; Ref. 20). The human transaldolase cDNA clone 4/2-4/1 is 1329 nucleotides in length, which correlates well with detection of a 1.3-kb mRNA in poly(A+) RNA from a variety of cell lines and tissues. Downstream of the open reading frame, the cDNA contains a typical polyadenylation site, AATAAA (21). The 1008-bp open reading frame in the human cDNA encodes a protein of 336 amino acids with a predicted molecular mass of 37.6 kDa. This is similar to the 37-kDa size of the yeast TAL gene comprised of 335 amino acids. The overall homology of translated amino acid sequences of the human (TAL-H) and yeast transaldolase proteins (TAL-Y) is 58%. However, there are blocks of 11–15 amino acids in which the identity is 100%. Presumably, these regions are important for the structure and/or function of the enzyme. Homology between TAL-H and TAL-Y was further substantiated by reactivity of antibodies to the human recombinant protein with yeast transaldolase. Comparative amino acid sequence analysis also indicated that the calmodulin-dependent protein kinase and five out of the seven protein kinase C phosphorylation sites are missing in the yeast enzyme. This

DISCUSSION

Retrotransposons are considered a major factor in the shaping and reorganization of the genome (1). The present study documents cloning, sequencing, and expression of a novel human retrotransposable element, TARE, which encodes the second and third exons of the human transaldolase gene, TAL-H. The transaldolase enzyme was originally described (6), and its structural gene was recently cloned from the yeast (TAL-Y; Ref. 20). The human transaldolase cDNA clone 4/2-4/1 is 1329 nucleotides in length, which correlates well with detection of a 1.3-kb mRNA in poly(A+) RNA from a variety of cell lines and tissues. Downstream of the open reading frame, the cDNA contains a typical polyadenylation site, AATAAA (21). The 1008-bp open reading frame in the human cDNA encodes a protein of 336 amino acids with a predicted molecular mass of 37.6 kDa. This is similar to the 37-kDa size of the yeast TAL gene comprised of 335 amino acids. The overall homology of translated amino acid sequences of the human (TAL-H) and yeast transaldolase proteins (TAL-Y) is 58%. However, there are blocks of 11–15 amino acids in which the identity is 100%. Presumably, these regions are important for the structure and/or function of the enzyme. Homology between TAL-H and TAL-Y was further substantiated by reactivity of antibodies to the human recombinant protein with yeast transaldolase. Comparative amino acid sequence analysis also indicated that the calmodulin-dependent protein kinase and five out of the seven protein kinase C phosphorylation sites are missing in the yeast enzyme. This
though the transaldolase gene locus appears to be a single copy
transcripts could represent poly(A-) RNA intermediates of
the 3' 411 fragment hybridizes to various abundant 0.5-9-kb
which span a chromosomal region of approximately
TARE. Thus, TARE may belong to the group of class
TAL-H locus suggest that these two exons may have uniquely
fragments of the full-length 41241 cDNA anneal to a single
positive clones in two different genomic libraries the copy num-
erative South-
men lymphocyte genomic DNA libraries. Comparative analysis
of genomic and cDNA sequences revealed that unlike the in-
tronless TAL-Y locus the TAL-H gene locus contains five exons
which development of insertion of TARE. Seven independent TARE isolates show a 296% sequence identity with the corresponding
exons 2 and 3 in the haploid genome. Direct repeats flanking exons 2 and 3 in the
structural differences suggest that function of the two enzymes
may be differentially regulated. Western blot analysis of protein lysates from Chinese hamster ovary cells and murine lymphocytes also demonstrated a 38-kDa protein indistinguishable from human TAL-H (not shown). This phylogenetic conservation from yeast to man suggests that TAL-H plays an essential function in cell biology.

Genomic organization of the TAL-H gene was investigated by sequence analysis of TAL-H-specific clones isolated from human lymphocyte genomic DNA libraries. Comparative analysis of genomic and cDNA sequences revealed that unlike the intronless TAL-Y locus the TAL-H gene locus contains five exons which span a chromosomal region of approximately 50 kb. Although the transaldolase gene locus appears to be a single copy element per haploid genome, exons 2 and 3 in the 5' region of the cDNA are highly repetitive. Based on comparative Southern blot analyses and the frequency of TAL-H exons 2 and 3 positive clones in two different genomic libraries the copy number of the TARE is estimated between 1000 and 10,000 per haploid genome. Direct repeats flanking exons 2 and 3 in the TAL-H locus suggest that these two exons may have uniquely developed by insertion of TARE. Seven independent TARE isolates show a >96% sequence identity with the corresponding exons of the TAL-H gene. Although both the 5' 4/2 and 3' 4/1 fragments of the full-length 4/2-4/1 cDNA anneal to a single 1.3-kb message in poly(A+) RNA, the 5' 4/2 fragment but not the 3' 4/1 fragment hybridizes to various abundant 0.5-9-kb transcripts in total cellular RNA (data not shown). These latter transcripts could represent poly(A-) RNA intermediates of TARE. Thus, TARE may belong to the group of class I RTE comprising the copia and Ty elements and intracisternal A-type particles (22). Presence of TARE in the coding sequence of the human transaldolase gene exemplifies that RTEs may be a major force in shaping of the eukaryotic genome.

REFERENCES