Structure and Expression of the TREX1 and TREX2 3′→5′ Exonuclease Genes

Dan J. Mazur and Fred W. Perrino‡
Wake Forest University School of Medicine, Department of Biochemistry, Winston-Salem, North Carolina 27157

The TREX1 and TREX2 genes encode mammalian 3′→5′ exonucleases. Expression of the TREX genes in human cells was investigated using a reverse transcription-polymerase chain reaction strategy. Our results show that TREX1 and TREX2 are expressed in all tissues tested, providing direct evidence for the expression of these genes in human cells. Potential transcription start sites are identified for the TREX genes using rapid amplification of cDNA ends to recover the 5′-flanking regions of the TREX transcripts. The 5′-flanking sequences indicate transcription initiation from consensus putative promoters identified −140 and −650 base pairs upstream of the TREX1 open reading frame (ORF) and −623 and −753 base pairs upstream of the TREX2 ORF. Novel TREX1 and TREX2 cDNAs are identified that contain protein-coding sequences generated from exons positioned in genomic DNA up to 18 kilobases 5′ to the TREX1 ORF and up to 25 kilobases 5′ to the TREX2 ORF. These novel cDNAs and sequences in the GenBank® data base indicate that transcripts containing the TREX1 and TREX2 ORFs are produced using a variety of mechanisms that include alternate promoter usage, alternative splicing, and varied sites for 3′ cleavage and polyadenylation. These initial studies have revealed previously unrecognized complexities in the structure and expression of the TREX1 and TREX2 genes.

The multistep processes of DNA replication, repair, and genetic recombination often require the excision of 3′ nucleotides to generate DNA 3′ termini suitable for subsequent metabolic steps. The apparent diversity of proteins containing 3′→5′ exonuclease activity likely reflects the different requirements for these enzymes in the maintenance of the human genome. In some cases these exonuclease are found in large proteins that contain multiple catalytic and functional properties. The 3′→5′ proofreading exonucleases are functional domains in the mammalian DNA polymerases δ (1), ε (2), and γ (3). These proofreading enzymes remove incorrectly polymerized nucleotides during DNA synthesis and minimize the incorporation of mismatches into the genome. The Werner syndrome protein (WRN) contains a 3′→5′ exonuclease activity in one functional domain and a 3′→5′ DNA helicase activity in another (4, 5). Deficiencies in the WRN protein increase genomic instability (6). The multifunctional p53 protein contains a 3′→5′ exonuclease localized to the central core domain (7). This core region in p53 also contains the sequence-specific DNA binding domain that functions in cell-cycle checkpoint control in mammalian cells (8). The hRAD1 (Ustilago maydis REC1) and hRAD9 are human homologues of yeast DNA damage checkpoint response proteins (9). These proteins also contain 3′→5′ exonuclease activities (10–12). The yeast mre11 mutant is defective in recombinational DNA repair (13). The purified MRE11 protein (14, 15) and a protein complex containing MRE11 contain 3′→5′ exonuclease activities (16). The TREX1 and TREX2 proteins are relatively small dimeric proteins that contain potent 3′→5′ exonucleases (17). The presence of 3′ excision activities in this apparently diverse collection of proteins, and likely others, probably reflects the multiple pathways present in human cells requiring the modification of DNA 3′ termini. However, insufficient information is currently available to understand the molecular pathways in which the different 3′→5′ exonuclease function.

Some insights into the catalytic requirements for 3′→5′ exonucleases have been gleaned from protein structure and mutagenesis studies and from protein sequence analysis. The proofreading exonuclease domains of the Escherichia coli DNA polymerase I large fragment and the bacteriophage T4 DNA polymerase have nearly identical folding patterns despite minimal overall sequence identity (18–20). Mutagenesis studies identify critical amino acids in three conserved motifs, Exo2 I, Exo2 II, and Exo2 III, that are positioned to coordinate two metal ions at the active site (21–23). The proofreading exonuclease of the mammalian DNA polymerases also contain these three Exo motifs (24). Statistical modeling strategies have been developed to identify additional proteins that might contain 3′ excision activity (25, 26). This methodology revealed the conserved exonuclease motifs in the WRN protein (27, 28), and biochemical analysis confirmed the 3′→5′ exonuclease activity in this protein (4, 29). The TREX sequences contain the Exo I and Exo II motifs and a variation in the Exo III motif, renamed Exo IIIe. The Exo IIIe motif is characterized by the presence of the sequence HXAXDĐ rather than HXXXDĐ (30–32) and is detected in the RNase T subfamily of exonuclease (25, 28, 33). The Exo IIIe motif in the TREX proteins suggests that these mammalian exonuclease most closely relate to the bacterial epsilon subunit of DNA polymerase III, exonuclease I, and the recently described exonuclease X (34).

The increasing number of 3′→5′ exonuclease activities detected in proteins from human genes indicates that a variety of structural folds distinct from the proofreading exonuclease are likely. The multifunctional Escherichia coli exonuclease III

1 Mazur, D. J., and Perrino, F. W. (2001) J. Biol. Chem., papers in press 10.1074/jbc.M010623200.
2 The abbreviations used are: Exo, exonuclease; no., GenBank® accession no.; RT, reverse transcription; PCR, polymerase chain reaction; ORF, open reading frame; EST, expressed sequence tag; bp, base pair(s); kb, kilobase(s); contig, group of overlapping clones; AML, acute myeloblastic leukemia; NNPP, Neural Network Promoter Prediction; GSS, genome survey sequence.
has a potent 3′ excision activity, and the structure of this protein is similar to APEX, the major human apurinic/apyrimidinic endonuclease (35, 36). Conserved residues in these enzymes indicate a common catalytic mechanism involving a single metal ion. The 3′ excision activity of the APEX protein is relatively weak and appears to be influenced by substrate and reaction conditions (37) as well as by the structure of the 3′ terminal nucleotide (38, 39). The structure of the p53 protein is similar to the exonuclease III and APEX proteins (40), and p53 protein is reported to contain 3′ → 5′ exonuclease activity (7). The hRAD1(REC1) and hRAD9 recombinant proteins contain 3′ → 5′ exonuclease activities, but extensive sequence and modeling analyses have not provided insights into the catalytic mechanisms of these proteins (41). Additional studies will be necessary to identify the complete repertoire of human genes encoding 3′ → 5′ exonucleases.

The gene for TREX1 encodes the major 3′ exonuclease activity measured in extracts prepared from mammalian cells. A 3′ exonuclease activity was first detected in biochemical assays and named DNase III by T. Lindahl et al. (42). Recently, the human and mouse cDNAs encoding 3′ → 5′ exonucleases were identified by sequencing peptides generated from the purified bovine (17) and rabbit (43) enzymes. A second closely related mouse cDNA, named Trex2,3 was discovered in data base searches using the TREX1 cDNA as a query sequence (17). We have measured expression from both TREX genes using a RT-PCR strategy and investigated in detail the 5′-flanking regions of these genes. The human and mouse TREX1 proteins are 314 amino acids in length and not 304 as previously reported (17, 43). Our analysis confirms expression of TREX1 and provides the first evidence for the expression of TREX2 in human cells. Novel cDNAs containing the TREX1 and TREX2 ORFs have been identified that contain exons spanning 18 kb for TREX1 and 25 kb for TREX2. The salient features of the TREX genes are presented in this report.

MATERIALS AND METHODS

DNAs—Oligonucleotide primers were synthesized in the DNA laboratory of the Wake Forest University Comprehensive Cancer Center and are listed in Table I. The bovine genomic DNA was from Sigma-Aldrich Co. (D1501). The mouse genomic DNA (strain 129 SV) was a generous gift from P. Dawson (Wake Forest University School of Medicine). The human genomic DNA was the BAC clone RP11–24C3 (no. AC021328) purchased from Research Genetics.

PCR of the TREX1 ORF from Genomic DNA—For amplification of TREX1 from genomic DNA, the PCRs (100 μl) contained 10 mM Tris-HCl, pH 9.0, 50 mM KCl, 0.1% Triton X-100, 200 μM dNTPs, 1.5 mM MgCl2, 50 ng of genomic DNA, and 1 μl each of the forward and reverse primers (Table I). The TREX2 PCRs also contained 5% Me2SO. Reactions were heated to 95 °C for 5 min prior to addition of Taq DNA polymerase (2.5 units, Promega Corp.) at 80 °C. The reactions were performed for 35 cycles at 95 °C for 1 min, 60 °C for 1 min, and 72 °C for 2 min. The products were resolved by agarose gel electrophoresis, recovered from the ethidium bromide-stained gels using spin columns (Qiagen), and sequenced using a PerkinElmer Life Sciences ABI Prism 377 automated DNA sequencer. The two-round PCRs were performed according to the manufacturer’s specifications using the nested Marathon Adapter primer pair and the TREX1- and TREX2-specific primer pairs indicated in Figs. 2, 4, 6, and 7. The TREX2 PCRs also contained 5% Me2SO. The first-round PCR products were fractionated on agarose gels and recovered from the gels using spin columns (Qiagen) in three separate size-selected pools. Samples of the size-selected products were used as templates in the second round PCR. Distinct product bands were recovered from gels, cloned into the pGEM®-T Easy vector (Promega Corp.), and sequenced.

Expression Analysis by PCR of the TREX1 and TREX2 Transcripts—Total RNA was recovered from blast cells of a patient diagnosed with acute myeloblastic leukemia (AML) by guanidine isothiocyanate extraction and cesium centrifugation (44). The AML RNA was treated with RNase-free DNase I (Promega Corp.) and further purified using a RNaseasy column (Qiagen). The AML RNA or tissue specific RNA (CLONTECH Laboratories, Inc.) (5 μg) was hybridized to 0.5 μg of oligoT12, primer (Promega Corp.) for 5 min at 95 °C and then 10 min at 70 °C. The RNA was reverse transcribed with SuperScript II (Life Technologies, Inc.) for 2 h at 42 °C to generate cDNA. The PCR conditions using cDNA and the tissue-specific cDNA were as described above for genomic DNA. The nested primer pairs for the two-round PCR of AML cDNA are described in Figs. 3 and 5. The template for the first round PCR was 200 ng of reverse transcribed AML RNA, and the template for the second round was a sample (1 μl) of the first-round PCR products. The products from the second round were resolved by agarose gel electrophoresis, recovered from the gel, and sequenced. A single-round PCR was performed for TREX1 and TREX2 expression analysis using the 13 human tissue-specific cDNAs and the primers indicated in the text.

Identification of Novel TREX1 and TREX2 Transcripts—Potential exons encoded in the genomic DNA of the 5′-flanking regions of the TREX1 and TREX2 ORFs were identified using the gene-finding algorithm GENSCAN (47). The PCR conditions for amplification of novel TREX1 and TREX2 ORFs were as described above for genomic DNA. The specific primer pairs used in the two round amplification reactions are indicated in the text and in Figs. 6 and 7. The products from the second round were resolved by agarose gel electrophoresis, recovered from the gel, and sequenced.

RESULTS AND DISCUSSION

The peptide sequences generated from a purified mammalian 3′ → 5′ exonuclease identified the human TREX1 cDNA from EST W24304 in the GenBank® data base in two independent studies (17, 43). More recently, we used the TREX1 sequence (no. AF151105) in a BLAST search of the GenBank® database to identify additional TREX1 ESTs (i.e. no. BE616406, AV774291, R23917, AA279657) and the human BAC clone RP11–24C3 (no. AC021328). Sequence alignments of these TREX1 ESTs indicated variations in the 5′-flanking regions (data not shown). These sequence variations prompted the systematic analysis presented in this work of the TREX1 cDNAs and the TREX1 genomic sequence in the human BAC clone RP11–24C3.

The TREX1 ORF—The genomic DNA sequences flanking the mouse, bovine, and human TREX1 genes were examined to confirm the single ORF structure of this gene. Initial studies of human and mouse TREX1 cDNA sequences identified a common ATG codon positioned near the 5′ end of the TREX1 ORFs (17). The recombinant proteins produced from the human and mouse TREX1 cDNAs using this ATG as a start codon generated active 3′ → 5′ exonucleases. However, mouse TREX1 ESTs (i.e. no. AI162180, BF577448, AA197643) contain a second in-frame ATG codon 30 nucleotides upstream raising the possibility that the initiating methionine in the TREX1 ORF had not been identified. To identify the initiating Met for TREX1, genomic DNA sequences positioned at the 5′ end of the TREX1 ORFs were sequenced. The mouse TREX1 ORFs were recovered using PCR, and the nucleotide sequences of these PCR products were determined. Primer pairs used in the two reactions were designed such that the TREX1 ESTs indicated in Fig. 1 and the available sequences in the GenBank® dbEST data base (Fig. 1, Table I). The lengths of the genomic DNA fragments recovered from the PCRs were 1114 bp (mouse), 507 bp (bovine), and 532 bp (human). Alignments of the ESTs with the recovered genomic sequences identified consensus intron donor and acceptor sequences and indicated that an RNA splicing process modified...
the 5'-flanking regions of the TREX1 transcripts (data not shown). The deduced amino acid sequences were determined from the genomic DNA sequences and aligned using ClustalW to determine the relative identity at the 5' ends of the TREX1 ORFs (Fig. 1). The alignment shows that the Met labeled 1 is the only Met conserved in all three mammalian sequences, indicating that this is the initiating Met of the TREX1 ORF. No sequence identity is detected prior to the proposed initiating Met. Additionally, the translated genomic sequences for mouse and bovine genomic DNA contain in-frame stop codons at positions 228 and 240, providing further support for the assignment of the initiating Met for TREX1 (Fig. 1). The translated human genomic sequence has two additional in-frame Met at positions 238 and 255. The significance of these potential Met codons is currently unknown. The human and bovine TREX1 sequences at the initiating ATG are identical to the Kozak consensus sequence (45), and the mouse sequence differs at a single position (Fig. 1). Additional PCRs of genomic DNA have confirmed the single ORF structures of TREX1 in the mouse, bovine, and human genomes (data not shown). The human and mouse TREX1 ORFs indicate a coding region of 314 amino acids, and the bovine TREX1 ORF is 315 amino acids in length.

**Fig. 1. Identification of the TREX1 initiating Met.** The indicated mouse (Ms), bovine (Bv), and human (Hu) TREX1 ESTs are aligned with the corresponding TREX1-containing genomic DNA sequences recovered by PCR. The TREX1 ORFs and the 5'-flanking exons (filled boxes) are connected by the intron sequences (solid lines). The alignments identify sequences present in the ESTs and genomic DNA (dotted lines) and sequences removed from ESTs by RNA splicing (solid, bent lines). The arrows indicate the positions of the PCR primers used to recover genomic DNA. The deduced amino acid sequences from mouse, bovine, and human TREX1 ORFs in the genomic sequences were aligned using ClustalW. The positions of identity in all three sequences (*) and two of three sequences (:) are indicated. The TREX1 protein sequences are boxed, and the deduced amino acids residues at positions prior to the proposed initiating methionine 1 are assigned negative values. The positions of in-frame stop codons (X) and residues prior to the stop codons (−) are indicated. The putative mouse, bovine, and human Kozak and consensus Kozak sequences are shown.

The 5'-flanking region of TREX1 cDNAs was examined using a 5'-RACE procedure. A two-round PCR was designed using spleen cDNA with the nested TREX1-specific reverse primers (T1rv1 and T1rv2) and the cDNA adapter primers. Seven independent clones ranging from 133 to 612 bp in length were recovered, and these sequences were aligned with the TREX1 genomic sequence (Fig. 2). To identify genomic sequences in the 5'-flanking region of TREX1 that might serve as transcription initiation sites, the sequence analysis Neural Network Promoter Prediction (NNPP) program was used. Two potential promoters positioned −28 and −40, providing further support for the assignment of the initiating Met for TREX1 (Fig. 1). The translated human genomic sequence has two additional in-frame Met at positions −38 and −55. The significance of these potential Met codons is currently unknown. The human and bovine TREX1 sequences at the initiating ATG are identical to the Kozak consensus sequence (45), and the mouse sequence differs at a single position (Fig. 1). Additional PCRs of genomic DNA have confirmed the single ORF structures of TREX1 in the mouse, bovine, and human genomes (data not shown). The human and mouse TREX1 ORFs indicate a coding region of 314 amino acids, and the bovine TREX1 ORF is 315 amino acids in length. A Drosophila TREX homolog (no. AE003581) encodes a protein of 351 amino acids and contains two exons. The homologous relationship between the mammalian and Drosophila genes is apparent by computational analysis using the COGNITOR program (46). The products of these genes fit into the same cluster of orthologous groups of proteins represented by the E. coli DNA polymerase III-ε subunit. Although the biochemical relationship between these proteins is very likely to catalyze the removal of nucleotides from DNA 3' termini, the evolutionary relationship between these genes, the cellular functions, and the three-dimensional structures are not known.

The 5'-flanking Region of TREX1 Transcripts—The 5'-flanking region of TREX1 cDNAs was examined using a 5'-RACE procedure. A two-round PCR was designed using spleen cDNA with the nested TREX1-specific reverse primers (T1rv1 and T1rv2) and the cDNA adapter primers. Seven independent clones ranging from 133 to 612 bp in length were recovered, and these sequences were aligned with the TREX1 genomic sequence (Fig. 2). To identify genomic sequences in the 5'-flanking region of TREX1 that might serve as transcription initiation sites, the sequence analysis Neural Network Promoter Prediction (NNPP) program was used. Two potential promoters positioned −140 and −650 bp from the TREX1 ORF are identified (Fig. 2). The 5' ends of four TREX1 cDNAs (Fig. 2, labeled 3–6) align with the genomic sequence at positions, indicating transcription initiation at the −650 consensus putative promoter sequence. Two of the cDNAs (Fig. 2, labeled 1 and 2) align at positions indicating transcription initiation at the −140 or the −650 putative promoter sequences. In addition, the 5' end of another cDNA (Fig. 2, labeled 7) was positioned 5' to both predicted promoter sequences, indicating additional or alternative promoters are present in the 5'-flanking region of TREX1. Two of the cDNAs (Fig. 2, labeled 6 and 7) were spliced at consensus intron donor and acceptor sequences.
that had been previously identified in human ESTs, providing further support for a RNA splicing modification of the 5'-flanking region of TREX1 transcripts.

Splicing of the TREX1 Transcripts and Expression in Human Tissues—The TREX1 cDNA sequences recovered in the 5'-RACE analysis were compared with the 5'-flanking regions of TREX1 ESTs available in the GenBank data base. These sequences indicated the presence of one intron donor sequence and two acceptor sequences (Fig. 3). Thus, in addition to the unprocessed TREX1 transcript, two splicing pathways were possible for processing of the TREX1 transcripts. It was predicted that splicing from the donor site to acceptor site A would generate a transcript encoding the complete TREX1 ORF, whereas splicing to acceptor site B would generate a transcript that lacks necessary TREX1 sequence to encode an active TREX1 protein (Fig. 3A). It is possible that these alternatively spliced TREX1 transcripts reveal a pathway for regulation of TREX1 by alteration of the mRNA stability or translation efficiency. A two-round PCR was designed to estimate the relative abundance of the three possible TREX1 transcripts using reverse transcribed RNA from AML cells. The TREX1 cDNAs were amplified using the nested TREX1 ORF-specific primers (T1rv1 and T1rv2) and the nested 5'-flanking region primers (T1fr1 and T1fr2). The three possible TREX1 transcripts were detected by agarose gel electrophoresis of the PCR products (Fig. 3B). Sequencing of the cloned products confirmed the identity of the least abundant 532-bp band as the product of the unspliced TREX1 transcript. The 212- and 102-bp bands resulted from amplification of the two spliced TREX1 transcripts. The most abundant band is the 212-bp product generated by splicing from the donor site to acceptor site A positioned 26 base pairs 5' to the predicted initiating methionine codon. These data indicate that the most abundant TREX1 transcripts in AML cells, initiating at the 2650 putative promoter, are processed by an RNA splicing mechanism that removes a 320-bp intron from the 5'-flanking region of the TREX1 transcripts. Analysis of mouse and bovine TREX1 ESTs in the data base reveal a similar RNA splicing pathway to conserved acceptor sites positioned at 27 bp in mouse and 221 bp in bovine prior to the initiating ATG codons, suggesting conservation of this mechanism between mammalian species.

The donor site to acceptor site A pathway is the predominant pathway for processing of TREX1 transcripts in various human tissues. To determine the pattern of TREX1 gene expression in human cells, a single-round RT-PCR experiment was performed using the T1rv2 and T1fr1 primers and reverse transcribed total RNA from 13 different tissues (Fig. 3). The three PCR products generated from TREX1 transcripts are detected

| Name              | Sequence                                      |
|-------------------|-----------------------------------------------|
| TREX1 reverse primers |                                              |
| Msrv              | GGCTCGGAGCAGAGATGG                            |
| Bvrv              | GAAACAGGAGAAGAGAGT                             |
| T1rv1             | GGCTCTCAGGGCCACAT                              |
| T1rv2             | ACAGCTGCTGCTGAGCTGTGCTGCTGCTG                |
| T1rv5             | GGGAAACAGGAGGAGCACAGAG                    |
| T1rv6             | GAGGCGCGGCGGACAT                              |
| TREX1 forward primers |                                             |
| Msfr              | GGACAGGGGAGACGAGA                             |
| Bvfr              | GAAACAGGAGAAGAGAGT                             |
| T1fr1             | GTGYTAAAGCCAGGAGACTG                          |
| T1fr2             | GAGAGTGTCGAGCCCA                              |
| T1fr3             | GCTGCTGCTAAGCTGTGCTGCTGCTG                |
| T1fr4             | GCCCAAATCGAGGGCGGCGGCGGCTG                  |
| T1fr5             | TGAGCTGCTGAGCTGACAG                          |
| T1fr6             | GCTTCTCGAGGCAGACGAC                         |
| T1fr7             | AGTACATGCTGCTGAGATGG                        |
| T1fr8             | CCAGGAGAAAAGAGGACGACGCTCG                   |
| T1fr9             | ACGGGTGACACTGCGA                             |
| T1fr10            | GTCCTCGAGAGGCTTCA                            |
| T1fr11            | GTGCCAGAGGCCAGAG                             |
| T1fr12            | GGAACCGCGTCAGTCAGAC                        |
| T1fr13            | AGCATCTAGCAGAGAG                              |
| T1fr14            | CGTGCTCAGAGCGCGC                            |

**FIG. 2.** A 5’-RACE analysis of the TREX1 ORF. The TREX1 cDNAs (1–7) were recovered by 5’-RACE, and the sequences were aligned with the TREX1 genomic sequence (filled box and solid line). The alignment identifies sequences present in the cDNAs and genomic DNA (dotted lines) and sequences removed from cDNAs by RNA splicing (solid, bent lines). The 5’ end positions (arrows), the NNPP-predicted promoters (2650 and –140), and the TREX1-specific primers (T1rv1 and T1rv2) are indicated. Hu, human.
in all tissues tested (Fig. 3C). In addition, the relative ratios of the three PCR products are similar in each tissue, with the 212-bp product being the most abundant in all tissues. Some quantitative differences in the staining intensities of the 212-bp products are apparent, suggesting variability in expression between tissue samples. These data suggest that TREX1 expression in spleen, prostate, thymus, and AML cells is higher than that in heart, skeletal muscle, and bone marrow. A more quantitative analysis will be necessary to substantiate these variations in expression, but ubiquitous expression of TREX1 is clear from these results.

The 5'-Flanking Region of TREX2 Transcripts—In previous work from this laboratory, an active 3' → 5' exonuclease was generated from a single mouse Trex2 EST identified in the GenBank database (17). To date only Trex2 ESTs from mouse have been deposited in the GenBank database. A PCR strategy was developed to identify human TREX2 cDNAs and to investigate the 5'-flanking region of these cDNAs using a
Fig. 4. A 5′-RACE analysis of the TREX2 ORF. The TREX2 cDNAs (1–6) were recovered by 5′-RACE, and the sequences were aligned with the TREX2 genomic sequence (filled box and solid line). The alignment indicates that all of the recovered cDNA sequences are identical to the genomic DNA sequence (dotted lines). The 5′ end positions (arrows), the NNPP-predicted promoters (−753 and −623), and the TREX2-specific primers (T2 rv1 and T2 rv2) are indicated. Hu, human.

Fig. 5. Splicing of TREX2 transcripts and expression in human tissues. The RNA splicing pathways (solid, bent lines) from donor site to acceptor site A or acceptor site B and the three possible TREX2 transcripts (1–3) are indicated (A). The PCR products (643, 130, and 99 bp) are predicted using the indicated nested primer pairs (T2 rv1, T2 rv2 and T2 rv3). Agarose gel electrophoresis of the PCR products generated using reverse transcribed AML RNA (B, lane 2) indicates the presence of all three TREX2 transcripts. Lane 1 contains DNA size standards. Total RNA from various human tissues was subjected to RT-PCR using the TREX2-specific T2 rv1 and T2 rv2 primers. Agarose gel electrophoresis of the PCR products (C) indicates the presence of only the 643-bp unspliced TREX2 transcript in all tissues. Hu, human.
5’-RACE procedure. A two-round PCR was designed using spleen cDNA with nested TREX2-specific reverse primers (T2rv1 and T2rv2) and the cDNA adapter primers. Six independent clones ranging in length from 95 to 959 bp were recovered, sequenced, and aligned with the TREX2 genomic sequence (Fig. 4). The genomic sequence in the 5’-flanking region of TREX2 was examined using the NNPP program to identify potential transcription initiation sites. Two potential promoters positioned −2623 and −2753 bp from the TREX2 ORF were identified (Fig. 4). The 5’ ends of five TREX2 cDNAs (Fig. 4, labeled 1–5) align with the genomic sequence at positions, indicating transcription initiation at one of these consensus putative promoter sequences. The 5’ end of another cDNA (Fig. 4, labeled 6) was positioned 5’ to both predicted promoters, indicating additional or alternative promoters are present upstream in the 5’-flanking region of TREX2.

Splicing of the TREX2 Transcripts and Expression in Human Tissues—The genomic DNA sequence in the 5’-flanking region of the human TREX2 was examined using the NNPP program to identify potential transcription initiation sites. Two potential donor sites were identified, suggesting the possibility for a processing pathway for TREX2 transcripts similar to that for TREX1 transcripts (Fig. 5A). Thus, like the TREX1 transcripts, two splicing pathways are possible for processing of the TREX2 transcripts. However, unlike processing of a TREX1 transcript, splicing from the donor site to acceptor site A or to acceptor site B generates a TREX2 transcript encoding the complete ORF. A two-round PCR was performed using the nested TREX2 ORF-specific primers (T2rv1 and T2rv2) and the nested 5’-flanking region primers (T2fr1 and T2fr2) with AML cDNA to recover TREX2 cDNAs. The three possible TREX2 transcripts are detected upon agarose gel electrophoresis of the PCR products (Fig. 5B). Sequencing of the cloned products confirmed the identity of the most abundant 643-bp band as the product of the unspliced TREX2 transcript. The 130- and 99-bp bands resulted from amplification of the two spliced TREX2 transcripts. The relative intensity of the bands likely reflects the relative abundance of the TREX2 transcripts in AML cells with the unspliced TREX2 transcript being the predominant transcript. These results provide the first evidence for expression of the TREX2 gene in human cells and identify an RNA splicing process that removes a 513- or a 544-bp intron from the 5’-flanking region of the TREX2 transcripts. The sequence of the mouse Trex2 EST (no. AA060540) indicates splicing in the 5’-flanking region from the donor site to acceptor site B to remove a 623-bp intron, indicating conservation of this RNA splicing mechanism between mammalian species (data not shown).

An experiment was designed to measure the gene expression pattern for TREX2 in human cells. The pattern of TREX2 gene expression was determined in a series of RT-PCRs using the T2fr2 and T2rv2 primers, and reverse transcribed total RNA from 13 different tissues (Fig. 5). The PCR product generated from the unspliced TREX2 transcript was detected in all tis-
sues tested (Fig. 5C). The spliced TREX2 transcripts were not detected in this single-round PCR (data not shown), indicating that the unspliced form of the TREX2 transcript is the most abundant in all tissues tested. The staining intensities of the 643-bp products generated from thymus and spleen cDNA were greater than those from heart, skeletal muscle, and testis, suggesting some variations in expression levels of TREX2 in human cells. The recovery of TREX1 and TREX2 transcripts from all human tissues tested implicates these proteins in housekeeping functions such as the DNA repair processes necessary to maintain the integrity of the human genome.

Identification of Novel TREX1 Transcripts—Previous mapping experiments located the human TREX1 ORF to chromosome 3p21.3–21.2 (43). However, limited genomic sequence in this region precluded a detailed analysis of the sequence surrounding the TREX1 ORF. More recently, we used the TREX1 cDNA (no. AF151105) as the query sequence in a BLAST search of the GenBank® dbEST data base to identify the human BAC clone RP11–24C3 (no. AC021328). The presence of TREX1 on this genomic DNA fragment confirms the mapped position of TREX1 to 3p21.3–21.2 on the current Human Genome Map. The RP11–24C3 DNA clone is a “working draft” sequence consisting of 20 unordered contigs with gaps of ~100 bp. The correct order of the five contigs containing the TREX1 ORF was determined using human, mouse, and pig ESTs in the GenBank® dbEST data base and a genomic sequence in the GenBank® GSS data base (Fig. 6). A pig EST AW346748 identifies the two contigs labeled II and III immediately 5’ to the TREX1 ORF-containing contig labeled I. The translated protein sequence of the first 66 bp of the pig EST identifies the human BAC clone A035823 in the GSS data base. A BLAST search of the dbEST data base using the A035823 clone as the query sequence identifies several ESTs (i.e. no. AI031903) that indicate the position of the contig labeled V. Thus, five DNA fragments containing 25 kb of genomic DNA on RP11–24C3 have been ordered to reveal the genomic sequence positioned 5’ to the TREX1 ORF. We cannot exclude the possibility that additional contigs might be intervening but were not detected in our analysis.

A PCR strategy was developed to identify novel TREX1 cDNAs with possible transcription initiation sites positioned 18 kb 5’ to the TREX1 ORF. The potential exons contained within this 18-kb region were identified using the gene-finding algorithm GENSCAN (47). A hypothetical transcript encoding a protein of 887 amino acids containing 13 exons with the TREX1 ORF as the most 3’ exon is predicted (Fig. 6). Transcripts with this exact sequence have not been identified. However, human ESTs containing exons 1–4 (no. BE871415), exons 2–11 (no. AK022405), and exons 12 and 13 (no. BE615019) have been identified (Fig. 6). A PCR strategy was developed to identify novel TREX1 cDNAs that contain the exon sequences encoded in the 18-kb 5’-flanking region (Fig. 6). First, a two-round PCR was performed using the nested TREX1 ORF-specific primers...
TREX Gene Structure and Expression

(T1rv2 and T1rv1) and the nested 5′-flanking region primers (T1fr3 and T1fr4) with AML cDNA to amplify cDNAs containing exons 10–13. Two products were recovered from agarose gels and sequenced. The sequences confirm the presence of exons 10–13 in these TREX1 cDNAs (Fig. 6, labeled 1 and 2). Furthermore, the sequences indicate the PCR products were recovered from transcripts spliced from exon 10 to exon 11 precisely as predicted using the GENESCAN program, but neither cDNA was spliced from exon 11 to exon 12. Additional PCRs were performed using the nested TREX1 ORF-specific primers (T1rv2 and T1rv1) with the nested 5′-flanking region primers (T1fr5 and T1fr6) to amplify cDNAs containing exons 4–13 or with the nested 5′-flanking region primers (T1fr7 and T1fr8) to amplify cDNAs containing exons 1–13. Two additional TREX1 ORF-containing cDNAs were identified from these reactions (Fig. 6, labeled 3 and 4). The cDNA labeled 3 contains the GENESCAN predicted exons 4–13 and two additional exons labeled 5A and 6AB. The cDNA labeled 4 contains the predicted exons 1–13 and four additional exons: 5A, 6A, 6B, and 6C. The TREX1 ORF-containing cDNAs identified in these PCRs span the polyadenylation signal (Fig. 6, labeled poly(A) #1) identified between exons 11 and 12. To generate these transcripts mRNA synthesis must proceed past the nonconsensus poly(A) 1 signal (AUUAAA) and through the TREX1 ORF to the consensus poly(A) 2 signal (AATAAA) (Fig. 6). These data indicate a complex pattern of transcription initiation and RNA processing for the mammalian TREX1 gene and suggest an association between the TREX1 ORF and exons identified in the 5′-flanking region.

A 5′-RACE analysis supports transcription initiation 5′ to the GENESCAN-predicted exon 1. A two-round PCR was performed using spleen cDNA with nested primers designed from exon 10 sequences (T2rv3 and T2rv4) and the cDNA adapter primers. Two clones were recovered, and these sequences were aligned with the TREX2 genomic sequence (data not shown). The 5′ ends of these cDNAs are located near the predicted initiation ATG in exon 1. Analysis of this sequence using the NNPP program identifies a potential transcription initiation site positioned upstream from the predicted initiation ATG in exon 1 (Fig. 7). This transcription initiation site is more than 25 kb 5′ to the TREX2 ORF. The detection of novel TREX1 and TREX2 transcripts containing a dicistronic structure indicates a complex pattern of expression for these genes. The potential relationships between the ORFs encoded in the 5′-flanking regions and the TREX1 and TREX2 ORFs are not apparent. Furthermore, the ability to translate the TREX ORFs within the context of the dicistronic transcript has not been tested.

In conclusion, we have demonstrated that the TREX1 and TREX2 genes encode mammalian 3′ → 5′ exonucleases that are expressed in all human tissues examined. There are a number of similarities in the structures and the expression patterns of the TREX genes. The genomic sequence encoding the 314-amino acid TREX1 protein is contained in a single ORF. The 236-amino acid TREX2 protein is also encoded in a single ORF. For both TREX1 and TREX2, transcripts are initiated within 1 kb of the exonuclease ORFs, and intronic sequences are removed from the 5′-untranslated region by two possible RNA splicing pathways. Additional sites of transcription initiation are identified at positions 18 kb 5′ to the TREX1 ORF and 25 kb 5′ to the TREX2 ORF generating transcripts that contain the TREX ORF and a second upstream ORF of unknown function. The detection of TREX1 and TREX2 transcripts in all human cells indicates the ubiquitous expression of these genes and supports a requirement for these 3′ exonucleases in DNA repair pathways in human cells.

Acknowledgments—We thank Scott Harvey for excellent technical assistance throughout this work and Sallyanne Fossey and Don Bowen for hybrid scan analysis.

REFERENCES
1. Chung, D. W., Zhang, J., Tan, C.-K., Davie, E. W., Su, A., and Downey, K. M. (1991) Proc. Natl. Acad. Sci. U. S. A. 88, 11197–11201
2. Kesti, T., Prantti, H., and Syväoj, J. E. (1993) J. Biol. Chem. 268, 10298–10305
3. Ropp, P. A., and Copeland, W. C. (1996) Genomics 36, 449–458
4. Shen, J. C., Gray, M. D., Oshima, J., Ashwinl, S. K., Fry, M., and Loeb, L. A. (1998) J. Biol. Chem. 273, 34139–34144
5. Yu, C. T., Oshima, J., Fu, Y. H., Wijmaa, E. M., Hisama, F., Alisch, R., Mathews, S., Nakura, J., Miki, O., Ooias, M., Mulligaa, J., and Schellenbe, G. D. (1996) Science 272, 258–262
6. Shen, J. C., and Loeb, L. A. (2000) Trends in Genetics 16, 213–220
7. Mummenbrauer, T., Janus, F., Muller, B., Wissmuller, N., and Deppert, W. (1999) Oncogene 18, 7068–7177
8. Udell, C. M., Lee, S. K., and Davey, S. (1998) Nucleic Acids Res. 26, 3971–3979
9. Thelen, M. P., Onel, K., and Holoman, W. K. (1994) J. Biol. Chem. 269, 747–754
11. Parker, A. E., Van de Weyer, I., Laus, M. C., Oostveen, I., Yon, J., Verhasselt, P., and Luyster, W. H. (1998) J. Biol. Chem. 273, 18332–18339
12. Bessho, T., and Sancar, A. (2000) J. Biol. Chem. 275, 7451–7454
13. Ajimura, M., Leem, S. H., and Ogawa, H. (1995) Genetics 133, 51–66
14. Petritni, J. H., Walsh, M. E., DiMare, C., Shen, X. N., Kurenberg, J. R., and Weaver, D. T. (1995) Genomics 29, 80–96
15. Paull, T. T., and Gellert, M. (1998) Mol. Cell 1, 969–979
16. Trujillo, K. M., Yuan, S. S., Lee, E. Y., and Sung, P. (1998) J. Biol. Chem. 273, 21447–21450
17. Mazur, D. J., and Perrino, F. W. (1999) J. Biol. Chem. 274, 19655–19660
18. Freemont, P. S., Friedman, J. M., Beese, L. S., Sanderson, M. R., and Steitz, T. A. (1988) Proc. Natl. Acad. Sci. U. S. A. 85, 8924–8928
19. Beese, L. S., and Steitz, T. A. (1991) EMBO J. 10, 25–33
20. Wang, J., Yu, P., Lin, T. C., Konigsberg, W. H., and Steitz, T. A. (1996) Biochemistry 35, 8110–8119
21. Derbyshire, V., Grindley, N. D. F., and Joyce, C. M. (1991) EMBO J. 10, 17–24
22. Reha-Krantz, L. J., Stocki, S., Nonay, R. L., Dimayuga, E., Goodrich, L. D., Koningsberg, W. H., and Spicer, E. K. (1991) Proc. Natl. Acad. Sci. U. S. A. 88, 2417–2421
23. Reha-Krantz, L. J., and Nonay, R. L. (1993) J. Biol. Chem. 268, 27100–27108
24. Bernad, A., Blanco, L., Lazaro, J. M., Martin, G., and Salas, M. (1989) Cell 59, 219–228
25. Koonin, E. V., and Deutscher, M. P. (1993) Nucleic Acids Res. 21, 2521–2522
26. Mian, I. S. (1997) Nucleic Acids Res. 25, 3187–3195
27. Mushegian, A. R., Bastett, D. E., Jr., Boguski, M. S., Bork, P., and Koonin, E. V. (1997) Proc. Natl. Acad. Sci. U. S. A. 94, 5831–5836
28. Mosser, M. J., Holley, W. R., Chattarjee, A., and Mian, I. S. (1997) Nucleic Acids Res. 25, 5101–5118
29. Huang, S., Li, B., Gray, M. D., Oshima, J., Mian, I. S., and Campisi, J. (1998) Nat. Genet. 20, 114–116
30. Barnes, M. H., Spacciapoli, P., Li, D. H., and Brown, N. C. (1995) Gene (Amst.) 165, 45–50
31. Straus, B. S., Sagher, D., and Acharya, S. (1997) Nucleic Acids Res. 25, 806–813
32. Taha-Ben, S. A., and Schaefer, R. M. (1998) Nucleic Acids Res. 26, 4005–4011
33. Ito, J., and Brathwaite, D. K. (1998) Mol. Microbiol. 27, 235–236
34. Viananathan, M., and Lovett, S. T. (1999) J. Biol. Chem. 274, 30924–30929
35. Mol, C. D., Kuo, C. F., Thayer, M. M., Cunningham, R. P., and Tainer, J. A. (1995) Nature 374, 361–366
36. Gorman, M. A., Morera, S., Rothwell, D. G., de La Fortelle, E., Mol, C. D., Tainer, J. A., Hickson, I. D., and Freemont, P. S. (1997) EMBO J. 16, 6548–6558
37. Wilson, D. M., III, Takeshita, M., Grossman, A. P., and Demple, B. (1995) J. Biol. Chem. 270, 16002–16007
38. Chaudhry, M. A., Dedon, P. C., Wilson, D. M., III, Demple, B., and Weinfield, M. (1999) Biochem. Pharmacol. 57, 531–538
39. Chou, K. M., Rokhmanova, M., and Cheng, Y. C. (2000) J. Biol. Chem. 275, 31099–31005
40. Cho, Y., Gorina, S., Jeffrey, P. D., and Pavletich, N. P. (1994) Science 265, 346–355
41. Venclovas, C., and Thelen, M. P. (2000) Nucleic Acids Res. 28, 2481–2493
42. Lindahl, T., Gallj, J. A., and Edelman, G. M. (1969) J. Biol. Chem. 244, 5014–5019
43. Hoss, M., Robin, P., Naven, T. J., Pappin, D. J., Sguoros, J., and Lindahl, T. (1999) EMBO J. 18, 3866–75
44. Lizzadi, P. M. (1983) Meth. Enzymol. 96, 24–38
45. Kozak, M. (1987) Nucleic Acids Res. 15, 8125–8148
46. Tatusov, R. L., Galperin, M. Y., Natale, D. A., and Koonin, E. V. (2000) Nucleic Acids Res. 28, 33–36
47. Burge, C., and Karlin, S. (1997) J. Mol. Biol. 268, 78–94
Structure and Expression of the TREX1 and TREX2 3′→5′ Exonuclease Genes
Dan J. Mazur and Fred W. Perrino

J. Biol. Chem. 2001, 276:14718-14727.
doi: 10.1074/jbc.M010051200 originally published online January 29, 2001

Access the most updated version of this article at doi: 10.1074/jbc.M010051200

Alerts:
- When this article is cited
- When a correction for this article is posted

Click here to choose from all of JBC's e-mail alerts

This article cites 47 references, 21 of which can be accessed free at http://www.jbc.org/content/276/18/14718.full.html#ref-list-1