Cloning, Genomic Organization, and Chromosomal Localization of Human Cathepsin L*

(Received for publication, June 22, 1992)

Shyam S. Chauhan‡, Nicholas C. Popescu§, David Ray‡, Robert Fleischmann‡, Michael M. Gottesman‡, and Bruce R. Troen¶

From the ‡Laboratory of Cell Biology and the §Laboratory of Biology, National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20892 and the ¶Department of Internal Medicine, Division of Geriatric Medicine, Institute of Gerontology, Geriatric Research, Education and Clinical Center, Veterans Administration Medical Center, University of Michigan, Ann Arbor, Michigan 48109

Cathepsin L is a lysosomal cysteine protease whose expression and secretion is induced by malignant transformation, growth factors, and tumor promoters. Many human tumors express high levels of cathepsin L, which is a broad spectrum protease with potent elastase and collagenase activities. Two published human cathepsin L cDNA sequences differ only in their 5′-untranslated regions. In this study, we demonstrate the concurrent expression of two distinct human cathepsin L mRNAs (hCATL-A and hCATL-B) in adenocarcinoma, hepatoma, and renal cancer cell lines. Cloning of the human cathepsin L gene by polymerase chain reaction amplification of genomic DNA and subsequent sequencing reveals that hCATL-A and hCATL-B mRNAs are encoded by a single gene. The 3′ end of the first intron contains the 5′ portion of hCATL-B and is contiguous to the second exon of the gene. These data suggest either the possibility of alternative splicing or the presence of a second promoter within the first intron of the hCATL gene. We mapped the hCATL gene to chromosome 9q21–22. Sequencing of both the mouse and human cathepsin L genes demonstrates almost complete conservation of exon and intron position, but significant divergence in intron structure, possibly reflecting differences in regulation of expression of the mouse and human cathepsin L genes.

Sl

EXPERIMENTAL PROCEDURES

Materials—The cloning vectors pGEM 3 and pGEM 4Z were purchased from Promega Corp. Escherichia coli strain HB 101 was used for subcloning and purification of plasmid DNA. All restriction endonucleases were purchased from Bethesda Research Laboratories and used according to the manufacturer’s guidelines. Oligonucleotide primers used for the polymerase chain reaction (PCR)1 or DNA sequencing were synthesized in the laboratory using an Applied Biosystems 380B DNA synthesizer.

Cell Culture—The human cell line KB-3-1 is a subclone of the KB cervical adenocarcinoma cell line obtained from the American Type Culture Collection (29). KB-3-1 cells were grown as monolayers in Dulbecco’s modified Eagle’s medium (Quality Biologicals, Inc.) containing penicillin (50 units/ml) and streptomycin (50 μg/ml), glutamine (2 mM), and 10% fetal calf serum. HTB 44 (human kidney adenocarcinoma) cells were obtained from ATCC and grown in Dulbecco’s modified Eagle’s medium containing 10% fetal bovine serum, 1 mM sodium pyruvate, glutamine, penicillin and streptomycin, and nonessential amino acids. BL 7404 cells (human hepatoma) were obtained from Dr. D. W. Shen, Laboratory of Cell Biology, NCI, NIH and grown as described earlier (27).

Ribonuclease Protection Assay—A 415-bp (SacI-BglII) fragment of hCATL-A cDNA was cloned into the Smal and BamHI sites of

1 The abbreviations used are: PCR, polymerase chain reaction; bp, base pair(s).

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

The GenBank®/EMBL Data Bank accession number(s) L06426.

1 To whom correspondence and reprint requests should be addressed: 5510 MSRB I, 1150 West Medical Center Dr., Ann Arbor, MI 48109-0680.
pGEM 2 after digestion of the 3'-overhanging SacI end with T4 DNA polymerase (utilizing the 3' to 5' exonuclease activity). This construct was digested with PsiI, and antisense radiolabeled ribo-probe was prepared by Bioserve Biotechnologies (College Park, MD) using T7 RNA polymerase. Inclusion of vector sequences in the translatable start codon (bold ATGs), the sequences are identical until 4 nucleotides downstream of the translational termination codon (bold TGAs). Regions not shown are identical. The published hCATL-B cDNA sequence contains 289 bp at the 5' end; however, these nucleotides are not shown here because they are part of the multiple cloning site of M13, the vector used for the subcloning and sequencing of hCATL-B.

pGEM 42 after digestion of the 3'-overhanging SacI end with T4 DNA polymerase (utilizing the 3' to 5' exonuclease activity). This construct was digested with PsiI, and antisense radiolabeled ribo-probe was prepared by Bioserve Biotechnologies (College Park, MD) using T7 RNA polymerase. Inclusion of vector sequences in the translatable start codon (bold ATGs), the sequences are identical until 4 nucleotides downstream of the translational termination codon (bold TGAs). Regions not shown are identical. The published hCATL-B cDNA sequence contains 289 bp at the 5' end; however, these nucleotides are not shown here because they are part of the multiple cloning site of M13, the vector used for the subcloning and sequencing of hCATL-B.

Expression of Two Forms of Cathepsin L mRNA—We employed a ribonuclease protection assay to identify the two species of hCAT-L mRNA in KB-3-1, human renal adenocarcinoma HTB 44 and human hepatoma BL 7404 cells (Fig. 2). The riboprobe was transcribed from an hCAT-L cDNA fragment and contained 415 bp identical with the region from bp 222 to 638 of the hCAT-L A mRNA, spanning from exon 1 to exon 4 of the hCAT-L gene (Fig. 2A). Since bp 222–277 of the hCAT-L cDNA are not found in the hCAT-L B cDNA (Fig. 1), the riboprobe contained only 362 bp identical with bp 182–543 of the hCAT-L B cDNA. In cell lysates from all three human cell lines, probe fragments of 415 bp (representative of hCAT-L A mRNA) and 362 bp (representative of hCAT-L B mRNA) were protected from the RNase digestion (Fig. 2B, lanes 5, 6, and 7). In contrast, no protected fragment.

**RESULTS**

Expression of Two Forms of Cathepsin L mRNA—We employed a ribonuclease protection assay to identify the two species of hCAT-L mRNA in KB-3-1, human renal adenocarcinoma HTB 44 and human hepatoma BL 7404 cells (Fig. 2). The riboprobe was transcribed from an hCAT-L cDNA fragment and contained 415 bp identical with the region from bp 222 to 638 of the hCAT-L A mRNA, spanning from exon 1 to exon 4 of the hCAT-L gene (Fig. 2A). Since bp 222–277 of the hCAT-L cDNA are not found in the hCAT-L B cDNA (Fig. 1), the riboprobe contained only 362 bp identical with bp 182–543 of the hCAT-L B cDNA. In cell lysates from all three human cell lines, probe fragments of 415 bp (representative of hCAT-L A mRNA) and 362 bp (representative of hCAT-L B mRNA) were protected from the RNase digestion (Fig. 2B, lanes 5, 6, and 7). In contrast, no protected fragment.
the expression of hCATL-B mRNA was consistently several-fold higher than hCATL-A mRNA. Similar results were observed in the simultaneous cloning of another fragment which exhibited nucleotide similarities of only 89% and 87%, and one region contained a trans-splicing event (Fig. 5). Therefore, pcosHMEP3 does not contain the functional cathepsin L gene.

PCR Amplification of the Human Cathepsin L Gene—Because repeated attempts to clone the human cathepsin L gene by screening genomic libraries were unsuccessful, we attempted to clone this gene by utilizing PCR technology. We initially amplified a 2000-bp fragment (HCL1) using an oligonucleotide primer pair from bp 18–37 and 426–443 of the hCATL-A cDNA sequence. The termination codon within the coding region, introns 2–5 interrupt the open reading frame between codons (type 0 intron), intron 6 interrupts the open reading frame between codons (type 1 intron), and was seen in the hybridization with murine NIH-3T3 fibroblast lysates (lane 3). Both A and B forms of cathepsin L mRNA are expressed concurrently in the cell lines studied. However, the expression of hCATL-B mRNA was consistently several-fold higher than hCATL-A mRNA. Similar results were obtained for RAJI and osteosarcoma cells (data not shown). Hybridization of the riboprobe with lysates from NIH-3T3 cells that were stably transfected with an expression vector containing hCATL-A cDNA resulted in only the expected 415-bp protected fragment (Fig. 2B, lane 4).

Isolation of a Gene Closely Related to Cathepsin L—In an attempt to isolate the human CATL gene or genes, we screened a cosmid library prepared from human KB-V1 genomic DNA and obtained a positive clone that we designated pcosHMEP3. We sequenced subclones of pcosHMEP3 using oligonucleotide primers from the hCATL-A cDNA in order to determine whether the clone was identical with the human cathepsin L gene. Two regions contained putative exons (pcosHMEP3a and pcosHMEP3b) and were homologous to the hCATL-A cDNA, but they exhibited nucleotide similarities of only 89% and 87%, and one region contained a translational termination codon (Fig. 5). Therefore, pcosHMEP3 does not contain the functional cathepsin L gene.

Genomic Organization of the Human and Mouse Cathepsin L Genes—We mapped the intron-exon boundaries of both species, although there is considerable variation in the size of the comparable introns. All of the introns follow the canonical GT/AG consensus rule (Table II) (35). Among the introns within the coding region, introns 2–5 are 5' interrupted the open reading frame between codons (type 0 intron), intron 6 interrupts the codon after the first nucleotide (type 1 intron), and
Human Cathepsin L Gene

TABLE I
Sequences of the oligomers used for the amplification of the human cathepsin L gene
Sequences are 5' to 3'. The first oligonucleotide of each pair is a sense primer and the second is an antisense primer. Numbers in parentheses refer to location within hCATL-A cDNA for exon primers and the number of nucleotides upstream of an exon for the intron primers.

<table>
<thead>
<tr>
<th>Fragment amplified</th>
<th>Amplimers used</th>
<th>Source (intron/exon)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HCL1</td>
<td>AACCTTGAGCGCATCCGTG (1837)</td>
<td>Exon 1</td>
</tr>
<tr>
<td>HCL2</td>
<td>ACAAGTGACATTTCCATCTTGTG (-170; -151)</td>
<td>Intron 2</td>
</tr>
<tr>
<td>HCL3</td>
<td>CAGATGTGTGAGCTGTTG (-203; -186)</td>
<td>Intron 4</td>
</tr>
<tr>
<td>HCL4</td>
<td>AGCTGCACACTGCTGAGTG (-95; -77)</td>
<td>Exon 6</td>
</tr>
<tr>
<td>HCL5</td>
<td>ACTCGAATCATTGAAGATCCG (1388; 1408)</td>
<td>Exon 8</td>
</tr>
</tbody>
</table>

TABLE II
Splice junctions of the mouse and human cathepsin L genes
The consensus sequence for splice donor and acceptor sites is shown on the top line. Py = pyrimidine, IVS = intron. The residues surrounding the sites for introns 2-7 are shown. Nonidentical hCATL amino acids are shown beneath the corresponding underlined mcatL amino acids. The splice phases for the introns interrupting the open reading frame are shown.

| Consensus | 5' donor AG* GTTNNN...'IVS......(Gly) 
| Splice Phase |
| IVS1 | mcatL | GCGCCTCGCAG* GTTGGTCCAGA...IVS1...TTTTCTCTCCCTAG* GTTGGTCCAGA |
| hCATL | GCCGCCTCGCAG* GTTGGTCCAGA...IVS1...TTTTCTCTCCCTAG* GTTGGTCCAGA |
| IVS2 | mcatL | GTATGGCACG* GTTGGTCCAGA...IVS2...TTTTCTCTCCCTAG* GTTGGTCCAGA |
| hCATL | GTATGGCACG* GTTGGTCCAGA...IVS2...TTTTCTCTCCCTAG* GTTGGTCCAGA |
| IVS3 | mcatL | TTGAAGAACCAG* GTTGGTCCAGA...IVS3...TTTTCTCTCCCTAG* GTTGGTCCAGA |
| hCATL | TTGAAGAACCAG* GTTGGTCCAGA...IVS3...TTTTCTCTCCCTAG* GTTGGTCCAGA |
| IVS4 | mcatL | GAAGAGCAACG* GTTGGTCCAGA...IVS4...TTTTCTCTCCCTAG* GTTGGTCCAGA |
| hCATL | GAAGAGCAACG* GTTGGTCCAGA...IVS4...TTTTCTCTCCCTAG* GTTGGTCCAGA |
| IVS5 | mcatL | GTGGGCCAAAG* GTTGGTCCAGA...IVS5...TTTTCTCTCCCTAG* GTTGGTCCAGA |
| hCATL | GTGGGCCAAAG* GTTGGTCCAGA...IVS5...TTTTCTCTCCCTAG* GTTGGTCCAGA |
| IVS6 | mcatL | TATAGTTCAG* GTTGGTCCAGA...IVS6...TTTTCTCTCCCTAG* GTTGGTCCAGA |
| hCATL | TATAGTTCAG* GTTGGTCCAGA...IVS6...TTTTCTCTCCCTAG* GTTGGTCCAGA |
| IVS7 | mcatL | TCAGAGACAG* GTTGGTCCAGA...IVS7...TTTTCTCTCCCTAG* GTTGGTCCAGA |
| hCATL | TCAGAGACAG* GTTGGTCCAGA...IVS7...TTTTCTCTCCCTAG* GTTGGTCCAGA |

FIG. 4. Human and mouse cathepsin L genomic structures. Cathepsin L gene structures were determined by a combination of sequencing and PCR analysis as described under "Experimental Procedures." Black rectangles with white numbers represent exons. Unshaded rectangles represent introns. Breaks are present in those introns not drawn to scale. The 181-bp hatched area represents the 3' end of the first human intron that is congruent with the 5' untranslated region of hCATL-B cDNA. The intron and exon sizes are shown above and below the unshaded and black rectangles, respectively. The fragments of the hCATL gene amplified from genomic DNA (HCL1-HCL5) are shown with the arrows representing the approximate locations of the primers used for the PCR.
intron 7 interrupts the codon after the second nucleotide (type 2 intron) (36).

We have shown that sequences downstream of the transcription initiation site regulate transcription of the catL gene (22). Therefore, we sequenced the entire first intron of the hCATL gene (Fig. 5). To our surprise, the 5′-untranslated region of hCATL-B cDNA is found at the 3′ end of the first intron and is contiguous with the second exon of hCATL-A (Figs. 4 and 5). This confirms that both species of hCATL mRNA are encoded by a single gene.

**Mapping of the Human Cathepsin L Gene to Chromosome 9**—The chromosomal mapping of the hCATL gene was performed by in situ hybridization using probes generated both from the full-length hCATL-A cDNA and from the 5′-untranslated region of hCATL-A cDNA. Similar results were obtained with both probes. When probe radiolabeled from full-length hCATL-A cDNA was used, 246 grains from 110 metaphase spreads were counted with clustering of grains at two sites (Fig. 7A). 54 (22%) of the total number of grains scored clustered at region 9q21–22. A second hybridization site was identified on chromosome 10q23–24, consisting of 18 grains. The grains observed at these sites jointly represented 29% of the total grains scored. The remaining grains were randomly distributed over the rest of the chromosomes. Representative metaphases exhibiting label on chromosomes 9 and 10 are shown in Fig. 6. Similarly, after hybridization with the probe made from the 5′-noncoding region of hCATL-A cDNA, 68 grains from a total of 267 counted on 110 chromosomes were localized at 9q21–22 (Fig. 7B), and 12 grains were identified at 10q23–24. The localization of the hCATL gene to chromosome 9 was confirmed by PCR amplification of human-hamster somatic cell hybrid DNA, using the primer pair that generated the 800-bp HCL2 (Fig. 8). Amplification of an appropriately sized fragment was achieved using pGHCL2, total human genomic DNA, and hybrid DNA containing human chromosomes 5 and 10, pcosHMEP3, or total Chinese hamster ovary genomic DNA (Fig. 8, lanes 5, 6, and 7). The absence of an amplified fragment from chromosome 10 and the clustering of grains at

**Fig. 5. Nucleotide sequence of the 5′ end of the human cathepsin L gene.** Sequence of first exon, entire first intron, and 20 bp of the second exon. **Underlined nucleotides** at the 3′ end of the intron are those found in hCATL-B cDNA. The translational start codon in exon 2 is highlighted by the bold ATG. The two overlined nucleotides in the intron differ from the sequence in the hCATL-B cDNA (CC instead of GG).

**Fig. 6. In situ hybridization of the human cathepsin L gene.** a–d, metaphase chromosomes from normal human lymphocytes after hybridization with a radiolabeled cathepsin L probe (a and c). Chromosomes exhibiting label are indicated by arrows and are identified by G-banding as chromosomes 9 (b) and 10 (d).
The expression of cathepsin L is elevated in a variety of human malignancies, and cathepsin L has also been established as a marker for malignant transformation in mouse fibroblast cell lines. There is only one known form of mouse cathepsin L mRNA. We and two other laboratories have cloned the same full-length mcatL cDNA (2, 24, 37). Furthermore, primer extension assays of RNA from cells transiently transfected with a portion of the mcatL gene containing the promoter and first three exons and introns, exhibit a single major transcription initiation site (22). However, two human cathepsin L cDNAs with heterogeneous 5' ends were cloned (24, 25), suggesting the presence of two distinct species of mRNA. Two species of mRNA differing only in their 5'-untranslated regions have been reported for other proteins (38-41). The results of the ribonuclease protection assay demonstrate that there are two steady state species of hCATL mRNA and that expression of hCATL-B mRNA is severalfold greater than hCATL-A mRNA in all the cell lines studied. The presence of a single 415-bp protected fragment in the murine cells stably transfected with hCATL-A cDNA makes it unlikely that the 362-bp protected fragment represents a degradation product of the hCATL-A mRNA. These two mRNAs may result from alternative splicing or the presence of two promoters in the 5' region of the hCATL gene.

Differences in the 5'-untranslated mRNA regions have been implicated in the regulation of gene expression via effects on mRNA stability and translational efficiency (42-44). We believe that the two species of hCATL mRNA, detected in various cell lines, may play a role in the regulation of cathepsin L protein synthesis. Also, since these two mRNA species are present in different levels and hCATL-B mRNA is not a direct product of hCATL-A mRNA, there is likely differential regulation of these two messages at either the level of transcription or stability. Furthermore, there may be differences in their translational efficiency since their sequences diverge 12 bases upstream from the initiating ATG codon.

Screening of human genomic libraries, using a radiolabeled probe made from the full-length hCATL-A cDNA, resulted in the isolation of a genomic cosmids clone. However, since two nucleotide sequences were only about 88% similar to hCATL cDNA and one putative exon contained a translational termination codon, we proceeded to clone the hCATL gene by PCR amplification of genomic DNA. After the initial amplification of pGHCL1 using two primers complementary to the hCATL-A cDNA, we chose subsequently to use intron-exon primer pairs to obtain specific amplification of the true hCATL gene and avoid amplification of fragments containing regions from possible homologous genes. It is interesting to note that, concurrent with the amplification of pGHCL5 using two primers complementary to the hCATL cDNA, we observed the amplification of another fragment of DNA (Ex8b) which had 91% nucleotide similarity to a region in exon 8 of the hCATL gene. It is unclear whether this third homologous, but nonidentical, region is related to pcosHMEP3.

Sequencing of the hCATL and mcatL genes revealed the presence of 8 exons and 7 introns with complete conservation of the intron-exon junctions within the coding portions of the genes. The rat cathepsin L (rcatL) gene also contains the same intron-exon splice sites as does mcatL and similar sized introns (45). The available sequence for the first intron of rcatL is 85% similar to the mcatL first intron. Although exons 2 through 7 are the same size in the human and mouse genes, there is considerable variation in intron composition between the two species. The overall similarity between the first introns of the mouse and human cathepsin L genes is 48%, with only two elements of 14 and 16 base pairs exhibiting 75% similarity. This may have important functional implications for differential regulation of expression of the human and mouse cathepsin L genes. Evolutionary divergence of intron sequences has been reported for a number of genes, including leukemia inhibitory factor and plasminogen activator inhibitor-1 (46, 47). However, the mouse and human interleukin 6
genes exhibit significant similarity between the first, third, and fourth introns (48), and the rat and human myoxygenal deaminase genes contain small highly conserved regions within some of the introns (49).

In this study, we demonstrate two distinct hybridization sites of radiolabeled hCATL cDNA probes on chromosomes 9q21–22 and 10q23–24 (Fig. 6). Because we were able to amplify a hCATL gene fragment from a chromosome 9 hybrid and obtained no amplified fragment from a chromosome 10 hybrid, using the same primer pair, the proper assignment of the hCATL gene appears to be at 9q21–22. An in situ hybridization site on chromosome 9q21–22 was also observed in another study using a probe generated from the translated region of hCATL-B cDNA (50). However, Southern analysis of hybrid cell DNA resulted in bands that co-segregated with chromosome 10. We believe that the second hybridization site on chromosome 10 is not an artifact and represents a region truly homologous to the hCATL gene. Indeed, a cathepsin L-like gene has recently been isolated from a human placental cosmid library, and in situ hybridization localized that clone (pWE15-HCLL1) to chromosome 10. However, despite a similar exon-intron structure, exon regions from pWE15-HCLL1 are not identical with either the hCATL gene, pcoSHMEP3, or Ex8b. Interestingly, chromosomes 9 and 10 contain a functional promoter. Such experiments, along with intron upstream of a reporter gene will determine whether it should help to characterize the differential expression of the hCATL gene appears to be at 9q21-22. An in situ hybridization site on chromosome 9q21–22 and 10q23–24 (Fig. 6).

In summary, two human cathepsin L mRNA species are expressed concurrently at a similar ratio in several different cell lines. These two mRNA species are encoded by a single gene located on chromosome 9q21–22. Our results also suggest the presence of other distinct sequences homologous to human cathepsin L. Lastly we describe a new strategy for cloning a gene in the presence of other closely related putative exon sequences. Future transfection studies with the hCATL first intron upstream of a reporter gene will determine whether it contains a functional promoter. Such experiments, along with the cloning of the 5′-flanking regions of the hCATL gene, should help to characterize the differential expression of the two mRNA species.

REFERENCES


3 S. D. Bryce, personal communication.