The mouse and human genes encoding the recognition component of the N-end rule pathway

(ubiquitin/proteolysis/E3/N-recognin/Ubr1)

YONG TAE KWON*, YUVAL REISS†, VICTOR A. FRIED‡, AVRAM HERSHKO§, JEONG KYO YOON*, DAVID K. GONDA§, PITCHAI SANGAN¶, NEAL G. COPELAND¶, NANCY A. JENKINS‡, and ALEXANDER VARSHAVSKY*,**

*Division of Biology, California Institute of Technology, Pasadena, CA 91125; †Department of Biochemistry, Tel Aviv University, Tel Aviv 69978, Israel; ‡Department of Cell Biology and Anatomy, New York Medical College, Valhalla, NY 10595; §Unit of Biochemistry, Faculty of Medicine, Technion, Haifa 31096, Israel; ¶Department of Molecular Biophysics and Biochemistry, Yale University School of Medicine, New Haven, CT 06520-8024; and †Mammalian Genetics Laboratory, Advanced Bioscience Laboratories-Basic Research Program, National Cancer Institute-Frederick Cancer Research and Development Center, Frederick, MD 21702

Contributed by Alexander Varshavsky, May 5, 1998

ABSTRACT  The N-end rule relates the in vivo half-life of a protein to the identity of its N-terminal residue. The N-end rule pathway is one proteolytic pathway of the ubiquitin system. The recognition component of this pathway, called N-recognin or E3, binds to a destabilizing N-terminal residue of a substrate protein and participates in the formation of a substrate-linked multibuquitin chain. We report the cloning of the mouse and human Ubr1 CDNAS and genes that encode a mammalian N-recognin called E3α. Mouse UBR1p (E3α) is a 1,757-residue (200-kDa) protein that contains regions of sequence similarity to the 225-kDa Ubr1p of the yeast Saccharomyces cerevisiae. Mouse and human UBR1p have apparent homologs in other eukaryotes as well, thus defining a distinct family of proteins, the UBR family. The residues essential for substrate recognition by the yeast Ubr1p are conserved in the mouse UBR1p. The regions of similarity among the UBR family members include a putative zinc finger and RING-H2 finger, another zinc-binding domain. Ubr1 is located in the middle of mouse chromosome 2 and in the syntenic 15q15-q21.1 region of human chromosome 15. Mouse Ubr1 spans ~120 kilobases of genomic DNA and contains ~50 exons. Ubr1 is ubiquitously expressed in adults, with skeletal muscle and heart being the sites of highest expression. In mouse embryos, the UBR1 expression is highest in the branchial arches and in the tail and limb buds. The cloning of Ubr1 makes possible the construction of Ubr1-lacking mouse strains, a prerequisite for the functional understanding of the mammalian N-end rule pathway.

A number of regulatory circuits involve metabolically unstable proteins. Short in vivo half-lives are also characteristic of damaged or otherwise abnormal proteins (1–4). Features of proteins that confer metabolic instability are called degradation signals, or degrons. The essential component of one degradation signal, called the N-degron, is a destabilizing N-terminal residue of a protein (5, 6). The set of amino acid residues that are destabilizing in a given cell type yields a rule, called the N-end rule, which relates the in vivo half-life of a protein to the identity of its N-terminal residue. Similar, but distinct, versions of the N-end rule pathway are present in all organisms examined, from mammals to fungi and bacteria (6–8).

In eukaryotes, the N-degron comprises two determinants: a destabilizing N-terminal residue and an internal lysine or lysines (8). The Lys residue is the site of formation of a multibuquitin chain (9). The N-end rule pathway is thus one pathway of the ubiquitin (Ub) system. Ub is a 76-residue protein whose covalent conjugation to other proteins plays a role in a multitude of processes, including cell growth, division, differentiation, and responses to stress (1, 3, 4, 10). In most of these processes, Ub acts through routes that involve the degradation of Ub-protein conjugates by the 26S proteasome, an ATP-dependent multisubunit protease (11).

The N-end rule is organized hierarchically. In the yeast Saccharomyces cerevisiae, Asn and Gln are tertiary destabilizing N-terminal residues in that they function through their enzymatic deamidation into the secondary destabilizing N-terminal residues Asp and Glu (12). The destabilizing activity of N-terminal Asp and Glu requires their enzymatic conjugation to Arg, one of the primary destabilizing residues (6). The primary destabilizing N-terminal residues are bound directly by the UBR1-encoded N-recognin (also called E3), the recognition component of the N-end rule pathway (13). In S. cerevisiae, N-recognin is a 225-kDa protein that binds to potential N-end rule substrates through their primary destabilizing N-terminal residues—Phe, Leu, Trp, Tyr, Ile, Arg, Lys, and His. N-recognin has at least two substrate-binding sites. The type 1 site is specific for the basic N-terminal residues Arg, Lys, and His. The type 2 site is specific for the bulky hydrophobic N-terminal residues Phe, Leu, Trp, Tyr, and Ile (6).

The known functions of the N-end rule pathway include the control of peptide import in S. cerevisiae (through degradation of Cup9p, a transcriptional repressor of the peptide transporter Ptr2p); a role in controlling the Snf1p-dependent phosphorylation cascade that mediates osmoregulation in S. cerevisiae; the degradation of Gpa1p, a Gα protein of S. cerevisiae; and the degradation of alphaviral RNA polymerases in virus-infected metazoan cells (6, 14).

The mammalian counterpart of the yeast UBR1-encoded N-recognin (E3) was characterized biochemically in extracts from rabbit reticulocytes (15–17). Rabbit E3α was shown to be spe-

Abbreviations: Ub, ubiquitin; kb, kilobase; id., identity; si., similarity; BAC, bacterial artificial chromosome; FISH, fluorescence in situ hybridization; em, embryonic day.

Data deposition: Nucleotide sequences reported in this work have been deposited in the GenBank database [accession nos. AF061555 (mouse Ubr1 cDNA) and AF061556 (human Ubr1 cDNA)].

††To whom reprint requests should be addressed at: Division of Biology, 147-75, Caltech, 1200 East California Boulevard, Pasadena, CA 91125. e-mail: avarsh@cco.caltech.edu.

††The names of mouse genes are in italics, with the first letter uppercase. The names of human and S. cerevisiae genes are also in italics, all uppercase. If human and mouse genes are named in the same sentence, the mouse gene notation is used. The names of S. cerevisiae genes are Roman, with the first letter uppercase and an extra lowercase “p” at the end. The names of the corresponding mouse and human proteins are the same, except that all letters but the last “p” are uppercase. The latter usage is a modification of the existing convention (33), to facilitate simultaneous discussions of yeast, mouse, and human proteins. In some citations, the abbreviated name of a species precedes the gene’s name.
the rabbit Ubr1 cDNA was used to isolate, using PCR and a λgt11 mouse liver cDNA library, the corresponding 392-bp mouse Ubr1 cDNA fragment. This fragment then was used to screen the same cDNA library, yielding a 2.4-kb fragment of the mouse Ubr1 cDNA that encoded several of the peptide-derived sequences of the rabbit Ubr1. The encoded sequence was also significantly similar to that of the N-terminal region of S. cerevisiae Ubr1p (13) and contained the putative start (ATG) codon of the mouse Ubr1 ORF. To isolate the rest of the 5′ region of the Ubr1 cDNA, a′-rapid amplification of cDNA ends (RACE)—PCR (20) was performed with poly(A)′ RNA from mouse L cells and a primer from the 2.4-kb DNA fragment. 3′-RACE—PCR (20) was used to amplify a downstream region of Ubr1 cDNA. The resulting DNA fragment (nucleotides 2,470–3,467) then was used to screen a λgt10 mouse cDNA library from MEL-C19 cells. Five overlapping cDNA isolates (MR16, MR17, MR19, MR20, and MR23) that together spanned the entire Ubr1 cDNA were mapped and subcloned into Bluescript II SK+ (Stratagene), yielding the plasmid MR26, which contained the entire ORF of Ubr1. The ORF region of Ubr1 cDNA was sequenced on both strands at least twice, using independently derived cDNA clones.

MATERIALS AND METHODS

Isolation and Partial Sequencing of Mammalian E3α (UBR1p). Rabbit E3α was purified from reticulocyte extracts by using affinity chromatography with immobilized protein substrates of UBR1p and elution with dipeptides bearing destabilizing N-terminal residues (16). The resulting preparation was fractionated by SDS/PAGE. The band of ~180-kDa E3α was excised and subjected to digestion with trypsin. Amino acid sequences were determined for 14 peptides of rabbit UBR1p (Fig. 1A) by using standard methods (20).

Isolation of the Full-Length Mouse Ubr1 cDNA. A strategy that included the intrapeptide-interpeptide PCR (21) was used (see the legend to Fig. 1).

Isolation of a Partial Human UBR1 cDNA. Poly(A)′ RNA from human 293 cells was subjected to reverse transcription–PCR, using sets of primers corresponding to sequences of the mouse Ubr1 cDNA. One of the reactions yielded a 1.0-kilobase (kb) fragment that encompassed a region of the human UBR1 cDNA (Fig. 2).

Mouse and Human Genomic Ubr1 Fragments. A library of mouse genomic DNA fragments (strain SvJ) in bacterial artificial chromosome (BAC) (22) vector (Genome Systems, St. Louis) was used, as described in the legend to Fig. 2.

Northern, Southern, and Whole-Mount in Situ Hybridizations. Mouse and human multiple-tissue Northern blots
and either mouse or human Ubr1 cDNA fragments labeled with $^{32}$P were used (20). Southern hybridizations were carried out by using standard techniques (20). Mouse embryos were staged, fixed, and processed for in situ hybridization as described (23). For sectioning, the stained embryos were embedded in OCT medium (Sakura Finetek, Torrance, CA). A 1.2-kb Ubr1 cDNA fragment (nucleotides 3,150–3,355) was used as a template for synthesizing anti-sense- or sense-strand RNA probes labeled with digoxigenin (23).

**Chromosome Mapping of the Mouse and Human Ubr1.** The mapping of mouse Ubr1 was carried out by using the interspecific backcross analysis (24), essentially as described (18). Human UBR1 was mapped by using fluorescence in situ hybridization (CLONTECH), and either mouse or human Ubr1 cDNA fragments labeled with $^{32}$P were used (20). Southern hybridizations were carried out by using standard techniques (20). Mouse embryos were staged, fixed, and processed for in situ hybridization as described (23). For sectioning, the stained embryos were embedded in OCT medium (Sakura Finetek, Torrance, CA). A 1.2-kb Ubr1 cDNA fragment (nucleotides 3,150–3,355) was used as a template for synthesizing anti-sense- or sense-strand RNA probes labeled with digoxigenin (23).
RESULTS AND DISCUSSION

Isolation of the Mouse Ubr1 cDNA. Tryptic peptides of the purified rabbit E3a (UBR1p) (16) were isolated and sequenced (see Materials and Methods), yielding 14 short regions of E3a (Fig. 1A). These regions yielded significant similarities to the deduced sequence of S. cerevisiae Ubr1p (13). We used intrapeptide PCR (21) to identify a unique (nondegenerate) sequence of the rabbit Ubr1 cDNA. This method allowed amplification of a short unique DNA sequence by using two degenerate PCR primers (derived by reverse translation) that flank this sequence and correspond to the outermost regions of a single peptide (Fig. 1B). Several intrapeptide nucleotide sequences were obtained this way (Fig. 1). These sequences, together with those of the original degenerate primers, then were used to amplify a 392-bp fragment of the rabbit Ubr1 cDNA, using intrapeptide PCR (Fig. 1). This fragment encoded peptides T120 and T134 at either end and peptide T100 in the middle (Fig. 1B). A homologous 392-bp fragment of the mouse Ubr1 cDNA then was amplified by using the same method (Fig. 1B). The rabbit and mouse 392-bp Ubr1 cDNA fragments were 88% and 89% identical at the nucleotide and amino acid sequence levels, respectively, but lacked significant similarities to S. cerevisiae UBR1 (data not shown).

The 392-bp mouse Ubr1 cDNA fragment then was used, in conjunction with standard cDNA library screening and rapid amplification of cDNA ends–PCR (20), to isolate multiple Ubr1 cDNA fragments, and to assemble them into a 5,271-bp ORF amplification of cDNA ends–PCR (20), to isolate multiple cDNA fragments, and to assemble them into a 5,271-bp ORF. We determined the exon and restriction map of the 120-kb region of the mouse Ubr1 cDNA (Fig. 2), were, respectively, 91% and 94% identical to the corresponding mouse Ubr1 cDNA and UBR1p sequences. Overlapping genomic DNA fragments of human UBR1 that, together, encompassed a ~21-kb region of the human UBR1 gene and corresponded to the 1.0-kb fragment of the human UBR1 cDNA (Fig. 2), were isolated from human DNA by using cDNA-derived primers and PCR. Partial sequencing showed that this ~21-kb region of human UBR1 contained 11 exons whose length ranged from 49 to 155 bp, a distribution of exon lengths similar to that in a different region of mouse Ubr1 (Fig. 2). All of the sequenced exon/intron junctions (~23 exons), which encompassed a ~52-kb region of the mouse and human Ubr1, contained the consensus GT and AG dinucleotides characteristic of the mammalian nuclear pre-mRNA splice sites (data not shown) (20). Extrapolating from these data on the mouse and human Ubr1 genes and the corresponding regions of their cDNAs (Fig. 2), a mammalian Ubr1 gene is expected to be ~120 kb long and to contain ~50 exons.

The Mouse UBR1p Protein and its Homologs. The low overall sequence similarity of mouse UBR1p (E3a) to Ubr1p of either S. cerevisiae [22% identity (id.), 48% similarity (si.)] or another budding yeast, Kluyveromyces lactis (21% id., 48% si.), belied the presence of five regions, denoted I–V, which were significantly similar between the mouse and yeast versions of UBR1p (Figs. 3 and 4). By contrast, the Nmn1-encoded asparagine-specific N-terminal amidase, the most upstream component of the mouse N-end rule pathway, lacks sequence similarities to its S. cerevisiae counterpart Nta1p (12, 18). Databases of ~21-kb mouse regions identified other likely homologs of mouse UBR1p. In particular, the 1.927-residue protein of the nematode Caenorhabditis elegans (GenBank accession no. U88308) (32% id., 53% si.; termed Ce-Ubr1); the 1.872-residue S. cerevisiae protein (GenBank accession no. Z73196) (21% id., 47% si.; termed Sc-UBR2; ref. 4); the 2.168-residue C. elegans protein (GenBank accession no. U40029) (21% id., 45% si.; termed Ce-Ubr2); and the 7.94-residue CER3p protein of the plant Arabidopsis thaliana (GenBank accession no. X95962) (26% id., 49% si.) CER3p is involved in wax biosynthesis in A. thaliana (27). In addition, a 147-residue sequence of the yeast Candida albicans (http://alces.med.umich.edu/bin/genelist?LUBR1) was similar to the N-terminal region of mouse UBR1p (Fig. 4).

The presence of high-similarity regions I–V among these deduced sequences (Figs. 3 and 4) suggested the existence of a distinct protein family, termed UBR. The 66-residue region I, near the N terminus of UBR1p, is a particularly clear UBR family-identifying region (e.g., 61% id., 75% si. between mouse and C. elegans UBR1p) (Figs. 3 and 4).

Recent genetic analyses of S. cerevisiae Ubr1p (N-recogin) have shown that the regions I–III contain residues essential for the recognition of N-end rule substrates by Ubr1p. In particular, Cys-145, Val-146, Gly-173, and Asp-176 of region I were identified as essential residues of the type 1 binding site of S. cerevisiae Ubr1p (A. Webster, M. Ghislanz, and A.V., unpublished data). All four of these residues were conserved between the yeast, mouse, and C. elegans UBR1p (Figs. 3 and 4). Region I is present in all of the known UBR family members except CER3p of A. thaliana, which contains only regions IV and V (Fig. 4). Region I encompasses a Cys/His-rich domain, Cys-X12-Cys-X2-Cys-X5-...
Cys-X2-Cys-X2-Cys-X1-Cys-X11-Cys (Figs. 3 and 4), which is distinct from the known consensus sequences of zinc fingers and other Cys/His-motifs. Residues Asp-318, His-321, and Glu-560 of UBR1p, which have been identified as essential for the type 2 binding site of this N-recognin (A. Webster, M. Ghislain, and A.V., unpublished data), were found to be retained in region II (Aas-318 and His-321) and region III (Glru-560) of the mouse and C. elegans UBR1p.

Region IV contains another Cys/His-rich domain of UBR1p, Cys-32-X2-Cys-X2-loop-1-Cys-X1-His-X2-His-X2-Cys-X2-Cys (Figs. 3 and 4), which is present in all of the UBR family members, and fits the consensus sequence of the RING-H2 finger, a subfamily of the previously defined RING motif (28). At least some of the RING-H2 sequences are sites of specific protein–protein interactions (28). Apc11p, a subunit of the Ub-protein conjugating complex, also contains a RING-H2 finger (29).

Another area of similarity (24–50% id., 46–70% si.) among the UBR family members is region V (Fig. 3 and data not shown). This region, 115 residues long in mouse UBR1p, near the protein’s C terminus, is particularly similar between mouse and C. elegans UBR1p (50% id., 70% si.) (Fig. 3). Region V is located 4–14 residues from the UBR proteins’ C termini, the exceptions being the S. cerevisiae and K. lactis UBR1p, which bear, respectively, 132- and 159-residue tails of unknown function that are rich in the acidic Asp/Glu residues (36% and 33%) (Fig. 3). No significant similarities could be detected between mammalian UBR1p and other E3s (recognins) of the metazoan Ub system, including E6AP (30) and subunits of the cyclosome/anaphase promoting complex, except for the presence of a RING-H2 finger domain in the latter (29). [Different E3 proteins of the Ub system recognize different degrons in protein substrates, thereby defining distinct Ub-dependent proteolytic pathways (1, 4).]

**Expression of Mouse and Human Ubr1.** The 5’- and 3’-proximal mouse cDNA probes yielded similar results, detecting a single ~8-kb transcript in several tissues (Fig. 5Aa and Ab). In the testes, however, the ~8-kb species of Ubr1 mRNA was a minor one, the major species being ~6 kb (Fig. 5Aa). The levels of either mouse or human Ubr1 mRNA were highest in skeletal muscle and heart (Fig. 5Aa). The expression of mRNA encoding E214K, one of the mouse Ub-conjugating (E2) enzymes and a likely component of the mouse N-end rule pathway (6), was also highest in skeletal muscle and heart (18).

The distinct Ubr1 mRNA pattern in the testis (Fig. 5Aa) was reminiscent of the analogous expression pattern of Ntan1 mRNA, which encodes the Asn-specific N-terminal amidase, another component of the mammalian N-end rule pathway. Specifically, the size of the major species of Ntan1 mRNA was ~1.4 kb in all of the examined mouse tissues except testis, where the major species was ~1.1 kb (18). The ~1.1-kb Ntan1 transcript recently was found to hybridize only to the 3’-half (exons 6–10 but not exons 1–5) of the Ntan1 ORF (Y.T.K. and A.V., unpublished data). The functional significance of the testis-specific Ubr1 and Ntan1 expression patterns remains to be understood.

We used whole-mount in situ hybridization to examine the expression of Ubr1 during embryogenesis. In e9.5 (9.5 days old)
Fig. 6. Chromosomal locations of the mouse and human Ubr1 genes. (A) Mouse Ubr1 was mapped to the middle of mouse chromosome 2 by using interspecific (M. musculus-M. spretus) backcross analysis (18, 24). Shown are the segregation patterns of mouse Ubr1 and the flanking genes in 66 backcross animals that were typed for all loci. For individual pairs of loci, more than 66 animals were typed. Each column represents the chromosome identified in the backcross progeny that was inherited from the [M. musculus C57BL/6J × M. spretus] F1 parent. Filled and empty squares represent, respectively, C57BL/6J and M. spretus alleles. The numbers of offspring that inherited each type of chromosome 2 are listed below the columns. (B) A partial mouse chromosome 2 linkage map (MMU2), showing Ubr1 in relation to the linked genes Thbs1, Ephb2, and B2m, and also, on the left, the distance and recombination distances between these loci, in centimorgans, and in the map locations, in parentheses. (C) A partial human chromosome 15 linkage map (HSA15). Each dot on the right, in the 15q15-q21.1 region, corresponds to the actually observed UBR1-specific double-dot FISH signal detected on human chromosome 15 (see also Fig. 5D).

mouse embryos, the expression of Ubr1 was highest in the branchial arches and in the buds of forelimbs and the tail (data not shown). In e10.5 embryos, the expression of Ubr1 became high in the hindlimb buds as well (Fig. 5B). This pattern was maintained in the limb buds of e15.5 embryos (Fig. 5B). High expression of Ubr1 in the limb buds was confined predominantly to the surface ectoderm (Fig. 5C). This pattern of Ubr1 expression in embryos (Fig. 5B and C) is similar, if not identical, to that of Nan1, which encodes asparagine-specific N-terminal amidase (18) (Y.T.K. and A.V., unpublished data), consistent with mouse Ubr1p and NTAN1p being components of the same pathway.

The enhanced expression of Ubr1 in the embryonic limb buds (Fig. 5B and C) is interesting in view of the conjecture that the N-end rule pathway might be required for limb regeneration in amphibians (31). The injection of dipeptides bearing destabilizing N-terminal residues into the stumps of amputated forelimbs of newts was observed to delay limb regeneration, whereas the injection of dipeptides bearing stabilizing N-terminal residues had no effect (31). Rigorous tests of this and other suggested functions in embryos (Fig. 5B) and data not shown).

The chromosomal location of human UBR1 was determined by using FISH (25), with human UBR1 genomic DNA fragments as probes (Figs. 5D and 6C). This mapping placed UBR1 at the 15q15–15q21.1 region of the human chromosome 15, an area syntenic with the independently mapped position of mouse Ubr1 (Fig. 6). Ubr1 is located in the regions of human chromosome 15 and mouse chromosome 2 that appear to be devoid of the previously mapped but uncloned mutations.

Mutations in the human gene CANP3, which encodes a subunit of calpain and is located very close, if not adjacent, to UBR1, have been shown to cause a myopathy called the limb-girdle muscular dystrophy (32).

Concluding Remarks. Isolation of the mouse and human Ubr1 cDNAs and genes (Figs. 2–6) should enable functional understanding of the mammalian N-end rule pathway, in part through the construction and analysis of mouse strains that lack Ubr1. Recent searches in GenBank identified several mouse and human sequences in expressed sequence tag databases that exhibited significant similarity to the C-terminal region of mouse UBR1p. The cloning and characterization of the corresponding cDNAs have shown that there exist at least two distinct mouse (and human) genes, termed Ubr2 and Ubr3, which encode proteins that are significantly similar to mouse UBR1p (Y.T.K. and A.V., unpublished data). Molecular and functional analyses of these Ubr1 homologs are under way.

We are grateful to A. Webster and M. Ghislain for permission to cite their unpublished data. We thank members of the Varshavsky lab, especially I. V. Davydov, for helpful discussions, and L. Peck, G. Turner, H. Rao, A. Kashina, and F. Du for comments on the manuscript. Y.T.K. thanks B. Yu for sharing his Northern hybridization data on human β-actin mRNA. We gratefully acknowledge the sequencing of K. lactis UBR1 by P. Waller. N.G.C. and N.A.J. thank D. J. Gilbert and D. B. Householder for excellent technical assistance. D.K.G. was a Scholar of the Leukemia Society of America. This study was supported by National Institutes of Health grants to A.V. (DK35920 and GM31530), V.A.F. (NS42954), and D.K.G. (GM54534), and by a grant to N.G.C. from the National Cancer Institute.