Volume 15 Number 2 1987 Nucleic Acids Research

The human ubiquitin gene family: structure of a gene and pseudogenes from the Ub B subfamily

Rohan T.Baker and Philip G.Board

Department of Human Genetics, John Curtin School of Medical Research, Australian National University, PO Box 334, Canberra 2601, Australia

Received November 14, 1986; Accepted December 12, 1986

ABSTRACT

An ubiquitin cDNA clone was isolated from a human liver cDNA library. This clone contained two complete, and a portion of a third, ubiquitin coding sequences joined head to tail with no spacer peptides. Screening a human genomic library with a probe derived from the coding region of this cDNA identified a large number of cross-hybridising clones. Differential screening of these genomic clones with the 3' non-coding region of the cDNA identified three different 3'-positive clones. Sequence analysis of these three clones revealed: (i) a gene corresponding to the cDNA containing an intron in the 5' non-coding region and coding for three direct repeats of mature ubiquitin, and (ii) two related pseudogenes which appear to have arisen by reverse transcription and insertion into the genome. However, one pseudogene contains two repeats of the ubiquitin coding sequence, while the other contains only one. Hybridisation analysis of restricted human genomic DNA suggests the presence of one other closely related gene within the genome.

INTRODUCTION

Ubiquitin is a 76-amino acid (aa) protein that has been observed in all eukaryotic cells studied thus far (for a recent review see [1]). It exhibits remarkable evolutionary conservation, with identical aa sequence from insect to man, and only three substitutions within the plant and yeast sequences. Ubiquitin is involved in several distinct processes in the eukaryotic cell, all of which involve an unique covalent attachment of ubiquitin via its C-terminus by an isopeptide bond to free amino groups of other proteins. In the cytoplasm, ubiquitin is required for ATP-dependent, non-lysosomal proteolysis of both abnormal proteins, and normal proteins with rapid turnover (2,3). However, a recent report suggests that the rate of ubiquitination, and hence degradation, is a function of the N-terminal residue of the target protein: the "N-end rule" (4). Ubiquitin synthesis is increased during heat shock, presumably to cope with an elevated turnover of abnormal proteins during the stress response (5). In the nucleus, ubiquitin is conjugated to Lys 119 in histone H2A (6): this conjugate is not degraded, and has been implicated in the regulation of gene expression (7). Recently, ubiquitin has been observed

as a part of branched-chain cell surface receptor molecules, covalently attached to the core polypeptide of both the murine lymphocyte homing receptor (8,9) and platelet-derived growth factor receptor (10). Ubiquitin's role in the lymphocyte homing receptor (reviewed in [11]) is the most well characterised, with monoclonal antibody studies identifying ubiquitin at or near the adhesive domain of the receptor (8,11). There is also evidence that other cell surface molecules are ubiquitinated (8,11).

Recently, structural analyses of mRNAs and genes encoding ubiquitin have been reported in chicken, man, Xenopus, barley and yeast (5,12-16), and indirectly in Drosophila (17). In most cases, ubiquitin genes exhibit a novel structure, containing precise direct repeats of the 76-aa coding unit in a polyprotein format. In the two genes reported introns have been absent from the coding regions (12,16). There is considerable variation in the number of coding repeats: mRNAs isolated from chicken, Xenopus and barley (5,14,15) contain at least 3 repeats, while one yeast gene codes for a hexamer repeat (16), and one human gene contains nine direct repeats of the ubiquitin coding unit (12). One exception to this repeat structure is an incomplete human cDNA, encoding aa 5 through 76 of ubiquitin, followed directly by an 80-aa, non-ubiquitin, C-terminal extension (13). This extension contains a high proportion of basic aa (30%) and may play a role in the transport of ubiquitin into the nucleus (13). The C-terminal extension is conserved across human and rat mRNAs (13), while yeast also contains a fused ubiquitin gene (18).

Analysis of Xenopus poly(A)$^+$ RNAs reveals substantial population polymorphism (14), while barley (15) and man (12,13) show more stable patterns. Human poly(A)$^+$ RNAs contain three distinctly-sized ubiquitin gene transcripts of approximately 600, 1000 and 2450 nucleotides (nt) (12,13), which have been termed Ub A, Ub B and Ub C, respectively (12). The Ub C message is of the correct size to correspond to the nonomeric human gene (12), while the Ub A transcript corresponds to the single copy fused ubiquitin gene, based on the specific hybridisation of the C-terminal extension to the 600 nt mRNA (13).

In this paper we report the isolation and characterisation of a human cDNA and gene which correspond to the Ub B transcript. We also report the presence of two related ubiquitin processed pseudogenes in the human genome, which differ in the number of coding repeats they contain.

MATERIALS AND METHODS

Expression Screening of a Human Liver cDNA Library

A cDNA library made by blunt-end ligation of double stranded (ds) cDNA synthesised from human liver poly(A)$^+$ RNA into the PvuII site of the plasmid

pAT153/PvuII/8 (19) was generously provided by Dr D. Bentley (20). Inserts were excised from this library by BamHI and PstI digestion, and ligated into the expression plasmid pEX_2 (21) digested with BamHI and PstI. This library was used to transform E. coli K12 MC1061 (22), which contained the plasmid pcI_{857} (23) as the source of the temperature sensitive repressor. Expression screening was by the colony blot procedure of Stanley (24). The antiserum used to detect ubiquitin antigenic determinants was a gift from Dr T. Suzuki. Cross-reacting antigen was detected using a rabbit primary antiserum and a goat anti-rabbit IgG second antibody coupled to alkaline phosphatase (25).

Subcloning and Nucleotide Sequence Analysis

All restriction digests, DNA ligations and plasmid transformations were done by standard methods (26). The plasmid pUC18 (27), and M13 derivatives mp8, mp9 (28), mp18 and mp19 (27) were used as vectors for subcloning and/or sequencing. E. coli K12 JM103 (29) was used for transformations. Some restriction fragments for subcloning were isolated from low melting temperature agarose as described (30). Recombinant dsDNA was prepared by the alkaline lysis method (31) as modified by Hattori and Sakaki (32). Recombinant M13 phage single stranded DNA (ssDNA) was prepared as described (30). Nucleotide sequences were determined by the Sanger dideoxy chain termination method (30) as modified by Messing (33), using the 17mer sequencing primer (New England Biolabs). The sequence of some regions of one subclone was determined by the method of Lin et al (34) using DNaseI to generate deletion subclones.

Probe Preparation and Hybridisation

DNA fragments were transferred from agarose gels to nylon membranes as described by Reed and Mann (35). Hybridisation was done according to the membrane manufacturers instructions (New England Nuclear) in the presence of dextran sulfate. Probes used for genomic screening and characterisation were ssDNA molecules labelled with [α-^{32}P]dCTP (Amersham) and generated by primer extension employing the 17mer sequencing primer and M13 ssDNA sequencing subclones, as described by Burke (36).

Screening of a Human Genomic DNA Lambda Phage Library

A library constructed by partial Sau3AI digestion of human genomic DNA and ligation into BamHI digested phage EMBL3 (37) was generously provided by Dr D. Anson. E. coli K12 ED8655 (38) was used as the host for all manipulations with EMBL3 derivatives. The library was screened by the method of Benton and Davis (39) (see RESULTS for descriptions of probes used), and hybridising clones were isolated by two further rounds of plaque purification. DNA from positive phage clones was isolated by the small-scale rapid preparation

of Ozaki and Sharma (40), and by large scale preparations as described by Maniatis et al (26).

Restriction Endonuclease Mapping

Restriction maps of phage DNAs were produced by standard methods (single and double digestions followed by size fractionation on agarose gels [26]) and by a modification of the method of Rackwitz et al (41). Partially digested phage DNAs were electrophoresed under conditions for the separation of high molecular weight DNA (42), transferred to a nylon membrane and hybridised with a nick-translated probe originating from the right arm of EMBL3 (generously provided by Ms C. Merritt). Autoradiography of the hybridised membrane produced a ladder of fragments corresponding to the order and distance of a restriction enzyme site from the right arm of the phage.

Preparation of Human Genomic DNA

High molecular weight DNA was prepared from human blood by the method of Grunebaum et al (43). Aliquots (3μg) were digested with restriction endonucleases overnight and electrophoresed on agarose gels. DNAs were transferred to a nylon membrane and subsequently hybridised as described above.

RESULTS

Isolation and Sequence Analysis of a Human Liver Ubiquitin cDNA Clone

Approximately 40,000 colonies from the human liver cDNA expression library were screened and 15 immunopositive colonies were selected. Following two further rounds of screening, the resulting clone with the largest cDNA insert of 700 base pairs (bp) was selected and termed pRBL26. Restriction enzyme digestion of this insert with either BglII, SalI or PvuII produced fragments of approx. 230bp (not shown), the expected size for an ubiquitin coding unit. These restriction enzyme sites and others were utilised in determining the nucleotide of the cDNA insert of pRBL26 by the strategy shown in Figure 1A. The sequence of this cDNA is not presented in isolation, but is shown with the sequence of the subsequently isolated gene in Figure 3.

Analysis of the sequence revealed one long open reading frame of 504bp followed by a termination codon TAA, a 3' untranslated region of 142bp and a poly(A) tail of 50bp. The 3' untranslated region includes the polyadenylation signal AATAAA (44,45) 17bp upstream of the polyadenylation site. The coding region begins with 15 aa from the C-terminus of ubiquitin, followed by two direct repeats of the 76 aa ubiquitin sequence, and terminates with an extra, non-ubiquitin, cysteine residue. Thus pRBL26 contains a partial cDNA: presum-

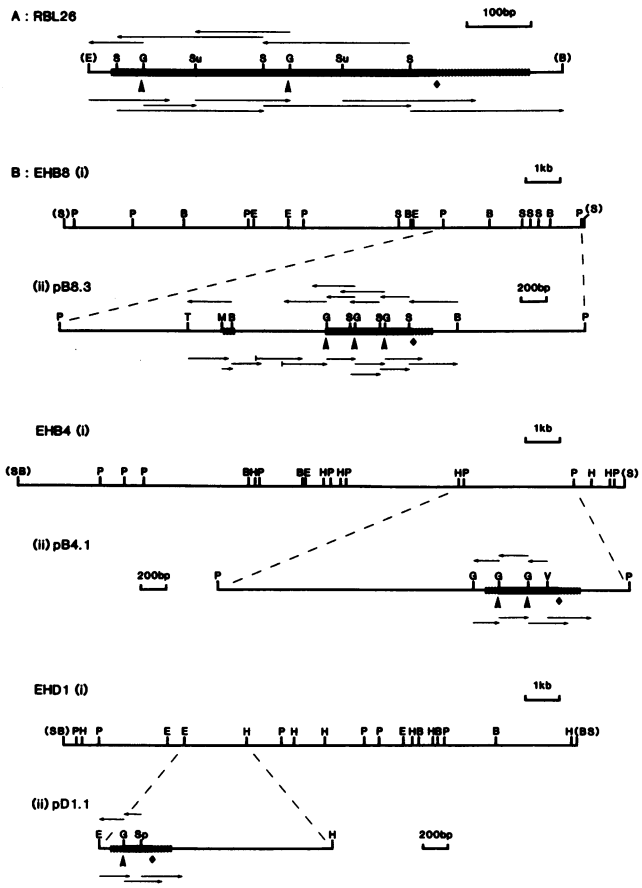

Figure 1. (A): Restriction map and sequencing strategy of the cDNA insert of pRBL26. The heavy black line represents the coding region, the hatched zone represents the 3' non-coding region, and the fine lines represent the poly(A) tail (3' end) and vector sequence (5' end). Restriction sites are abbreviated S:SalI, B:BamHI, E:EcoRI, G:BglII, Su:Sau3AI. Sites in parentheses arise from the vector. Horizontal arrows indicate the direction and extent of sequence determinations. Arrowheads indicate positions of in-frame start codons, while a diamond indicates the stop codon.(B): Restriction maps, subcloning and sequencing strategies for EHB8, EHB4 and EHD1. Genomic clone inserts are shown in (i) and subcloned fragments shown in (ii). Inserts are oriented so that transcription occurs from left to right. (i) All genomic inserts were mapped for S, B, E, HindIII (H), and PstI (P). Sites shown in parentheses arise from the vector, or from fusion of the insert and vector. (ii) Subcloned fragments are shown enlarged and the sequencing strategy indicated. Additional restriction sites shown are G, TaqI (T), MspI (M), PvuII (V), and SphI (Sp). Only those T, M, V, and Sp sites used for sequencing have been shown. Heavy black lines represent coding regions, while hatched zones represent 5' and 3' non-coding regions. Arrows, arrowheads and diamonds are as described in (A). Sequence arrows originating from a vertical bar indicate DNaseI subclones (34).

ably a full-length transcript would contain at least 3 ubiquitin coding units and include a 5' untranslated region.

Comparison of the sequence of this cDNA with the published human nonameric gene sequence (12) reveals that pRBL26 represents the product of a different gene. While the coding regions are highly homologous, the 3' non-coding regions show little similarity, either in length or nucleotide sequence (not shown). This evidence suggests that this cDNA may correspond to the 1000 nt Ub B mRNA observed (12,13). We therefore proceeded to isolate ubiquitin genomic clones to identify such a gene.

Isolation and Characterisation of Human Ubiquitin Genomic Clones

An M13 subclone containing a 228bp SalI fragment spanning the 3' coding repeat of the pRBL26 cDNA (Figure 1A) was used to generate a ssDNA probe. This probe was used to screen approximately 500,000 phage from a human genomic library, from which 20 positively-hybridising clones were selected. These clones were then screened with a probe generated from a subclone containing a 235bp SalI/BamHI fragment spanning the 3' non-coding region and poly(A) tail of the cDNA. This 3'-specific probe hybridised to only 3 of the 20 ubiquitin coding-positive clones. These clones, termed EHB8, EHB4 and EHD1, contained different genomic fragments based on restriction enzyme digestion patterns. Restriction maps were determined for each clone, and hybridising fragments were subcloned into pUC18 for further characterisation (See Figure 1B). A 4.1 kilobase pair (kb) PstI fragment from EHB8 was subcloned into pUC18 to generate pB8.3. Similarly, pB4.1 contains a 3.2kb PstI fragment from EHB4, while pD1.1 contains a 1.8kb EcoRI/HindIII fragment from EHD1 (Figure 1B).

Nucleotide Sequence Analysis of Ubiquitin Genomic Clones

Nucleotide sequences of the pRBL26-like regions of each genomic subclone were determined by the strategies shown in Figure 1B. The derived sequences are described below.

EHB4/pB4.1: Restriction mapping of the genomic subclone pB4.1 revealed three BglII sites separated by approximately 200 and 230bp, but no SalI sites. The determined nucleotide sequence is given in Figure 2. Analysis of the sequence reveals that it is very similar, but not identical to, the pRBL26 cDNA clone. The most notable difference is that EHB4 contains only two copies of an ubiquitin-like coding unit. It also contains the extra cysteine codon preceeding the stop codon, and a 3' non-coding region that is 95% homologous to the corresponding region of pRBL26, up until the polyadenylation site (compared in Figure 4B). The flanking regions of EHB4 show features characteristic of processed pseudogenes. Firstly, a poly(A) tail-like region is encoded

Figure 2. Nucleotide sequences of ubiquitin pseudogenes EHB4 (upper) and EHD1 (lower). Sequences are numbered from the A of the first ubiquitin-like initiation codons. Negative numbers are used in the 5' non-coding regions. Ubiquitin-like start codons, extra C-terminal cysteine codons and AATAAA signals are underlined. Ubiquitin-like stop codons are asterisked. The positions corresponding to the polyadenylation site of the pRBL26 cDNA clone are shown by #. Encoded poly(A) tail-like regions are shown by a dashed underline. Direct repeats spanning the pseudogene regions are overlined. EHB4 also contains an in-frame stop codon (doubly underlined, 193-195) and an 11bp insert (343-353) typed in lower case.

at the position corresponding to the polyadenylation site of the cDNA clone (nt 616 to 635 Figure 2). Secondly, the 5' and 3' regions are flanked by the direct repeat AGAAAYAGYTCAGTG (-115/-101 and 631/645, Y is a pyrimidine), of which the first 5 residues of the 3' repeat overlap the encoded poly(A) tail. EHB4 also exhibits other pseudogene-like features. The 5'coding repeat contains an in-frame stop codon TAG at the 65th codon. In addition, there are 13 other codon changes resulting in aa changes; 10 in the 5' repeat and 3 in the 3' repeat. The aa sequence "encoded" by this pseudogene is shown in figure 5. The 3' repeat also contains an 11bp insertion between the 38th and 39th codons (343 to 353, Figure 2) which, by itself, would be sufficient to disrupt the reading frame. The sequence CCCCCAGCCCC has been inserted into the normal coding sequence CCCCCCAG between the A and G residues, and appears to have

arisen by duplication, possibly during the reverse transcription event which created the pseudogene. In spite of these changes the 5' and 3' "coding" repeats respectively show 92 and 96% homology when compared with the 5' complete repeat of the pRBL26 cDNA (ignoring the 11bp insert). The lack of a 5' non-coding region in the cDNA clone prevents comparison of the 5' non-coding region of EHB4. However, the location of the 5' direct repeat (generated by pseudogene insertion) from -115 to -101 suggests that the 100bp upstream of the start codon is representative of the 5' non-coding region.

EHD1/pD1.1: Restriction mapping indicated that EHD1 was the least likely genomic clone to correspond to the pRBL26 cDNA insert, with no SalI sites, and only one BglII site within the hybridising region subcloned (see Figure 1B). Sequence analysis (Figure 2) reveals that while it is similar to the cDNA clone, it shows marked differences. It also shows similarity to the two-repeat pseudogene EHB4. Its most important feature is the presence of only one ubiquitin-like coding unit. As with EHB4, EHD1 contains the extra C-terminal cysteine codon preceeding the stop codon, and a 3' non-coding region that is respectively 94% and 93% homologous to the corresponding regions of pRBL26 and EHB4 (compared in Figure 4B). EHD1 also appears to be a processed pseudogene. A short poly(A) tail-like region is encoded 6bp downstream of the position corresponding to the cDNA polyadenylation site (382 to 389, Figure 2). Also, the 5' and 3' regions are flanked by the direct repeat AA$\overset{T}{A}$ATTCTGGA (-108/-98 and 386/396), again with some overlap of the 3' repeat and the encoded poly(A) tail. The single coding unit contains 8 codon changes leading to aa changes, none of which are in-frame stop codons. The translation of this coding unit is shown in Figure 5. In addition, EHD1 contains an amber termination codon (TAG), compared to the TAA ochre codons of pRBL26 and EHB4. The coding unit is 93% homologous to the 5' complete repeat of the pRBL26 cDNA in spite of these changes. Most interestingly, the 98bp upstream of the ubiquitin-like initiation codons of the two pseudogenes are 88% homologous, and they diverge at the position of the 5' direct repeats (compared in Figure 4A). There is no homology between the two pseudogenes either upstream of the 5' direct repeat or downstream of the 3' direct repeat (Figure 4). This is in agreement with their processed pseudogene nature, in that they may be inserted at any location in the genome.

EHB8/pB8.3: Restriction mapping of the genomic subclone pB8.3 indicated that its BglII and SalI restriction patterns were the most similar to the pRBL26 cDNA insert. Sequence analysis revealed that EHB8 contains the gene most likely represented by the cDNA clone. The sequence is presented in Figure 3.

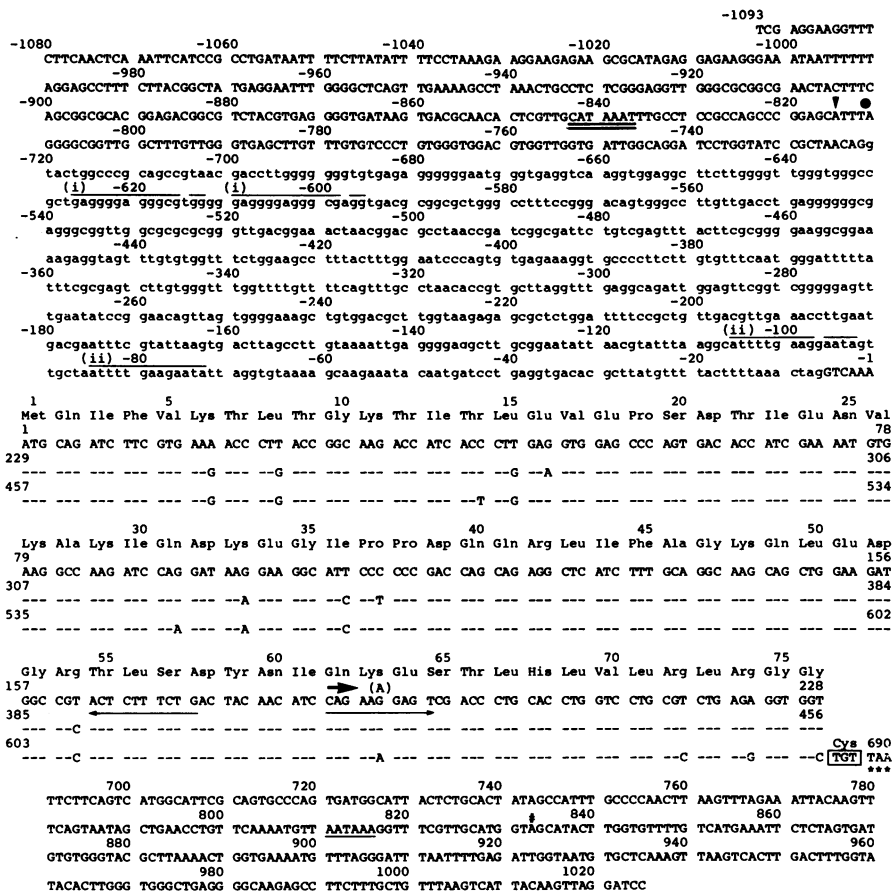

Figure 3. Nucleotide sequence of a human 3-repeat ubiquitin gene and the cDNA insert of pRBL26. The sequence is numbered as in Figure 2. The nucleotide sequence of the 5' ubiquitin coding direct repeat (1-228) is given in full; in the following repeats, nucleotide identity with the 5' repeat is indicated by a dash. The ubiquitin aa sequence is shown and numbered above the 5' repeat. The extra C-terminal cysteine codon is boxed, and the stop codon is asterisked. The AATAAA signal is underlined, and the polyadenylation site shown #. The intron within the 5' non-coding region is typed in lower case. The 5' limit of homology with the two pseudogenes EHB4 and EHD1 (see Figures 2 and 4) is shown by a filled circle. A possible CAP site 4bp 5' of this position is indicated by an arrowhead. A putative TATA box (-843 to -837) is doubly underlined. Two direct repeats within the intron ((i), (ii)) are overlined. The cDNA insert of pRBL26 begins within the 5' coding repeat at 184 (heavy arrow) and continues to the polyadenylation site (833), followed by a poly(A) tail of 50 nt. The cDNA differs in only one position from the gene sequence: an A in parentheses at 189. An inverted repeat within the coding region presumably responsible for the breakpoint in the cDNA is underlined with arrows.

This gene consists of three ubiquitin coding repeats, followed by an extra C-terminal cysteine codon, a stop codon TAA, and a 3' non-coding region identical to that of the cDNA clone. The coding repeats encode the correct ubiquitin protein sequence, are joined directly head to tail and neither the coding region nor the 3' non-coding region are interrupted by introns. The cDNA sequence begins with the 62nd codon of the 5' repeat, spans the middle and 3' repeats, and matches the 3' non-coding region until 17bp downstream of the AATAAA signal, where the cDNA has a poly(A) tail. As with many other genes, there is some ambiguity about the exact polyadenylation site: the first A of the poly(A) tail could be encoded by the gene (nt 833). The cDNA sequence differs in only one position from the corresponding region of the gene: the sixth nt of the cDNA is an A, while the gene has a G (nt 189). This difference is a silent change in the 3rd position of a codon; in fact the corresponding codon in the 3' repeat also has an A in this position. This single discrepancy is most likely an artefact arising during cDNA construction and cloning, or it may represent allelic variation.

Analysis of the 3' non-coding region around the polyadenylation signal and site reveals other sequences that have been implicated in 3' end formation. Mclauchlan et al (46) have identified a consensus sequence YGTGTTYY located approximately 30 nt downstream of the AATAAA signal, which has been shown to be required for efficient formation of mRNA 3' termini. EHB8 contains such a sequence with one mismatch GGTGTTTT 31 nt from the AATAAA signal (842 to 849). This consensus sequence is observed in 67% of mammalian genes examined (46). Another sequence, CAYTG, is possibly involved in the selection of a polyadenylation site in conjuction with the AATAAA signal via hybridisation with the small nuclear RNA U4 (47). A similar sequence, CATGG (827 to 831) occurrs in EHB8 within the polyadenylation region AATAAA$(N)_{10}$CATGGNA*, where N is any nt and A* is the first nt of the poly(A) tail. This sequence matches the consensus for a Class I U4 RNA hybrid (47), and suggests that U4-mediated polyadenylation may operate here. The presence or absence of these two sequences has been implicated in the differential usage of alternate polyadenylation sites in the human albumin gene (48). Another sequence observed in several eukaryotic genes is TTCAAA or close derivatives, which occurrs 32 to 62 nt downstream of AATAAA (48). EHB8 contains a recognised derivative, TTAAAA, 67 nt from AATAAA (883 to 888), but also contains the consensus TTCAAA beginning 11 nt upstream from AATAAA (800 to 805). Involvement of this sequence in mRNA 3' end formation has yet to be confirmed (48).

Initial sequence determination of EHB8 only continued for approximately

300bp upstream of the 5' coding repeat initiation codon. The cDNA insert of pRBL26 was incomplete and contained no 5' non-coding sequence for comparison with EHB8. However, comparison can be made with the approx. 100bp 5' flank of the two pseudogenes, which is presumably representative of the 5' non-coding region of the mRNA. This comparison produced an interesting result. The first 8 residues 5' of the start codons are identical in all 3 clones, but further upstream EHB8 shows no homology to either of the two pseudogene regions. The sequences diverge 5' of the sequence AGGT. This sequence matches the consensus resulting from the splicing of an intron, AGG^{T}_{G} (49). In addition, the EHB8 sequence 5' to this region is TTTAAACTAGGT (-16 to -5) which, except for the A triplet, matches the intron/exon junction consensus yyyyyynyagG^{T}_{G} (49) (intron in lower case). These observations indicate that an intron may exist in the 5' non-coding region of EHB8. This intron would not be present in the processed pseudogenes which explains the divergence of the gene and pseudogene sequences upstream of the splice site. We therefore proceeded to search further upstream for the presence of a 5' non-coding exon.

Characterisation of the 5' Non-Coding Region of EHB8

An M13 subclone containing a 195bp EcoRI/BglII fragment from the 5' flank of the genomic pseudogene subclone pD1.1 (Figure 1B) was used to generate a ssDNA probe. This subclone contained 10bp of ubiquitin coding sequence, 98bp of 5' flank common to both pseudogenes, and a further 90bp specific to EHD1 (-187 to 10 Figure 2). This probe hybridised to a 1.3kb PstI/BamHI fragment separated from the coding region by approx. 740bp, and more specifically to 350bp TaqI/BamHI and 80bp MspI/BamHI subfragments. The nucleotide sequence around this BamHI site was determined by the strategy shown in Figure 1B. Comparison of this sequence with the two pseudogene 5' flanks revealed that this BamHI site was within the 5' homologous region and had been lost from both pseudogenes as a result of nt substitutions. Homology begins at the sequence CAGGTA (-724 to -719), which matches the exon/intron junction consensus $^{A}_{C}$AGgtr (49) (intron in lower case, r is a purine). These results confirm the presence of an intron within the 5' non-coding region of this gene.

The sequence of the intron was determined by sequencing the 740bp BamHI/BglII fragment. Some regions of this fragment were sequenced using the DNaseI method (34) to generate deletion subclones. These subclones are indicated in Figure 1B as sequencing arrows originating from vertical bars. The intron was found to be 715bp in length, with the "coding-like" strand relatively G rich (35.4%) and C poor (15.8%). The G residues are distributed non-randomly: 56% are within the clusters $(G)_nXG$ or $GX(G)_n$ (n= 2 to 7, X= A,T or C) which occur

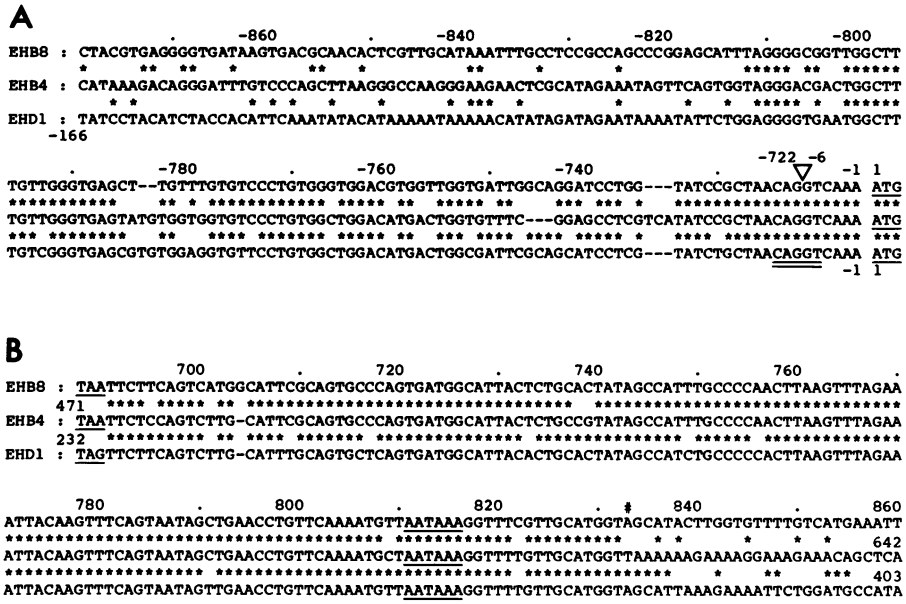

Figure 4. Comparison of non-coding regions. (A): Comparison of 5' non-coding regions. The regions 5' of the ubiquitin-like initiation codons of EHD1 (lower line) and EHB4 (centre line) were compared to the corresponding region of EHB8 (top line) after the deletion of the 715bp intron (shown by ▽). An asterisk above a nucleotide indicates homology between it and the EHB8 gene sequence. Numbering is as in Figures 2 and 3. Dashes indicate gaps introduced to maximise homology. A sequence corresponding to the post-splice consensus (ref. 49) is doubly underlined. (B): Comparison of 3' non-coding regions. The regions 3' of the ubiquitin-like stop codons of EHD1 (lower line) and EHB4 (centre line) were compared to the corresponding region of EHB8 (top line). Numbering, asterisks and dashes are as in (A). The AATAAA signals are underlined, and the polyadenylation site of EHB8 is shown by #.

37 times within the intron. There are several direct repeats within the intron. The two longest, 15 nt with one mismatch and 13 nt with one gap are indicated in Figure 3. The sequence AGGT at the splice junctions is redundant: the exon/intron and intron/exon boundaries in Figure 3 have been chosen to confer with the "GT-AG" rule (49).

The total EHB8 sequence thus determined is presented in Figure 3. The region from -811 to -1 contains a 96bp sequence homologous to the 5' flanks of the two pseudogenes, interrupted by a 715bp intron. The 5' flanks of all 3 genomic clones are compared in Figure 4A. For the purposes of this comparison the intron has been omitted from the EHB8 sequence as indicated in Figure 4A. A 2bp gap has been introduced into EHB8 to maximise homology, while a 3bp gap

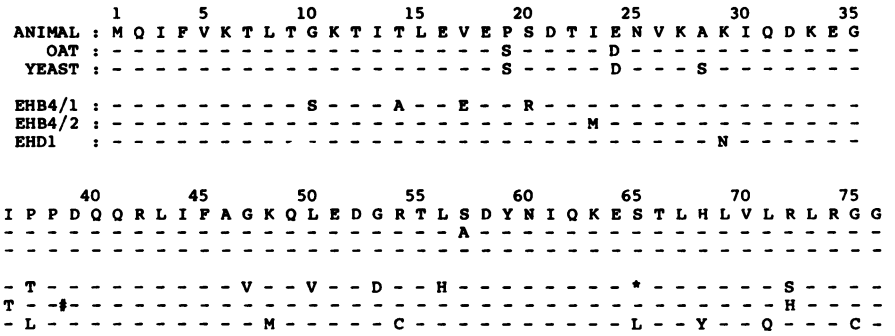

Figure 5. Amino acid sequences of ubiquitin and ubiquitin pseudogene-encoded proteins. Top 3 lines: comparison of determined and derived aa sequences of ubiquitin from 3 kingdoms. ANIMAL represents sequences in man (52,12,13,this paper), cow (53), mouse (9), chicken (5), trout (54), Xenopus (14) and fly (55); OAT represents sequences in oat (56) and barley (15); and YEAST represents the yeast protein (16,57). A dash indicates identity with the animal sequence. Lower 3 lines: aa sequences encoded by the 1st and 2nd repeats of EHB4 (EHB4/1 and EHB4/2) and EHD1. Dashes are as above. An asterisk represents a stop codon, and the 11bp insert in EHB4/2 between residues 38 and 39 is shown by #. Extra C-terminal residues have been omitted.

has been introduced into all three sequences to align them at the BamHI site (nt -743 to -738, EHB8). Ignoring the 2bp deletion, the 5' flanks of EHB4 and EHD1 respectively show 80.4% and 82.3% homology to the derived 5' flank of EHB8. Homology ceases at a position corresponding to the 3' end of the pseudogene direct repeats, with the EHB8 sequence at this position exhibiting no homology to either repeat.

The transcription initiation (CAP) site has not been determined for this gene. However, an analysis of the sequence upstream of the 5' limit of homology with the pseudogenes reveals a possible CAP site and TATA box. Directly upstream of the 5' limit is a sequence matching the consensus CAP site (49): the A residue within the sequence CATTT (-816 to -812). A possible TATA box begins 28bp upstream, with the sequence CATAAAT (-843 to -837). A transition from T to C is the most commonly observed deviation at the first position of the TATA box (49). The sequence "TATA" does not occurr until a further 200bp 5' of this sequence (-1046 to -1043). The upstream region does not contain a consensus CAT box. However, the sequence ATTTGG (-964 to -959) represents the complement of CCAAAT, a possible CAT box. The presence of an active CAT box on the non-coding strand has been shown for the herpes simplex virus thymidine kinase gene (50) and has been suggested for the human glucagon gene (51).

A transcript of this gene originating from the putative CAP site with

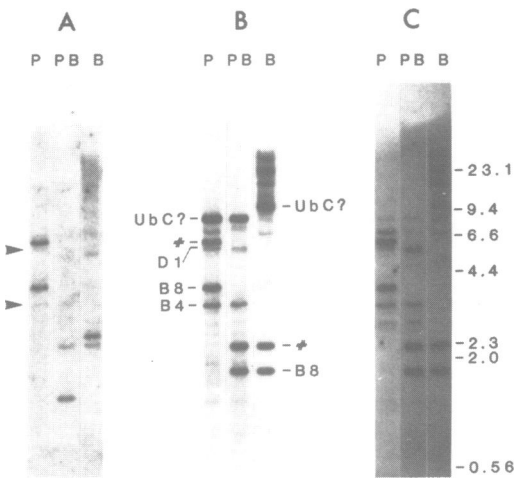

Figure 6. Hybridisation analysis of total human genomic DNA. Human genomic DNA was digested with PstI (P), PstI and BamHI (PB), or BamHI (B), electrophoresed through a 0.8% agarose gel, transferred to a nylon membrane and hybridised. Probes used were 5' non-coding specific (panel A); ubiquitin-coding specific (panel B); and 3' non-coding specific (panel C). The size standards on the right of panel C are from a HindIII digest of λ-DNA and are in kb. Arrowheads on panel A indicate weakly hybridising bands corresponding to the two pseudogenes. Details of probes and band assignments are given in the text.

the intron spliced would be 933 nt plus a poly(A) tail in length, which corresponds to the observed 1000 nt mRNA (12,13).

Hybridisation Analysis of Total Human Genomic DNA

Human chromosomal DNA was digested with PstI, PstI/BamHI, or BamHI, electrophoresed through an agarose gel, transferred to a nylon membrane, and hybridised with probes derived from different fragments of the ubiquitin gene (Figure 6). Neither enzyme cuts within known ubiquitin coding sequences. Panel A represents hybridisation with a probe generated from a 350bp TaqI/BamHI subclone from the 5' end of EHB8, spanning 70bp of 5' non-coding region and 280bp of further upstream sequence. Panel B represents hybridisation with a coding region probe as described earlier, while Panel C represents hybridisation with a 3' non-coding specific probe. This probe was generated from a 373bp SalI/BamHI subclone from the 3' end of EHB8, which spans 40bp of coding region, the stop codon, and 330bp downstream from the stop codon.

The coding region-specific probe (Panel B) hybridises to a large number of fragments, consistent with previous observations (12). The 5' and 3' non-coding probes (Panels A and C) produce fewer bands, consistent with their gene-specific nature. Comparison of the 3 panels with known restriction maps allows the assignment of some ubiquitin genes/pseudogenes to observed hybridising fragments (HFs). A 4.0kb PstI fragment which hybridises with all three probes corresponds to the 3-repeat gene EHB8 and is marked B8 (Panel B). This gene also produces 1.7kb BamHI HFs (Panels B and C), and 1.3kb PstI/BamHI and 2.5kb BamHI HFs (Panel A), consistent with its restriction map. Similarly,

EHB4 (B4) is assigned to a 3.2kb PstI HF with all 3 probes, while EHD1 (D1) is assigned to a 5.5kb PstI HF. The 5' non-coding probe hybridises weakly to the pseudogene HFs (arrows, Panel A) as the 350bp probe only contains 70bp of the 5' non-coding region present in the pseudogenes. Restriction maps of the EHB4 and EHD1 pseudogenes suggest that they would produce BamHI HFs of larger than 9.5 and 10.5kb respectively, and presumably correspond to some of the large HFs observed in the BamHI digests. The hybridisation signal strength with the coding probe (Panel B) is proportional to the number of coding units present in the HF. Thus D1 represents a one-repeat HF, B4 a two-repeat HF, and B8 a 3-repeat HF. A further HF identified by all 3 probes has been indicated # on Panel B. This HF exhibits approx. equal intensity with all 3 probes when compared to the EHB8 HF. This result suggests firstly a 3-repeat gene, and secondly that it has 5' and 3' flanks highly homologous to EHB8, not only in the transcribed regions, but also further up- and downstream. This HF does not represent restriction fragment length polymorphism, as an identical pattern was observed in 38 unrelated individuals (not shown). Thus another EHB8-like gene is present in the genome, which may have arisen through gene duplication, and subsequent loss or creation of restriction enzyme sites.

Another HF present only on Panel B may represent the 9-repeat Ub C gene reported (12). The 8.5kb PstI and 10kb BamHI HFs produce a much stronger signal than the 3-repeat gene, are consistent with the published restriction map (12), and do not hybridise with either of the EHB8 non-coding probes.

At least one other ubiquitin gene is present in the human genome: the single repeat fused gene corresponding to the Ub A transcript (13). This gene has not yet been characterised and its restriction map is unknown. The only unassigned PstI HF on Panel B is a one-repeat intensity, 6.7kb fragment which may correspond to the fused gene. Alternatively, the fused gene HF could be masked by another HF of similar length. Several other HFs of weak intensity are observed with the coding and 3' non-coding probes (Panels B and C). These most likely represent pseudogenes with sequences considerably diverged from the gene sequences.

DISCUSSION

Ubiquitin Gene Structure

The human 3-repeat ubiquitin gene described here is structurally similar to other ubiquitin mRNAs and genes reported (5,12-16). The most striking feature is the presence of directly repeated coding units observed in all cases except for the single repeat fused cDNA (13). Other common features include

the lack of introns within coding and 3' non-coding regions (12,16), and the presence of a non-ubiquitin, C-terminal extension preceeding the stop codon. The only exception is a Xenopus cDNA which had no extra residue (14). The C-terminal extension is a single, variable residue, with the exception of the fused ubiquitin-80 aa cDNA (13). Known C-terminal residues are: asparagine (yeast [16]), lysine (barley [15]), tyrosine (chicken [5]), valine (human 9-repeat gene [12]), and cysteine in EHB8. The function of these one-residue C-terminal extensions is not known, but it is notable that there is both inter- and intra-species diversity. Their function may be to block the reactive C-terminal glycine of the mature protein. The intraspecies variation in man may then imply selective activation of poly-precursors based upon differential cleavage of the C-terminal blocking residue.

Another consistent feature of reported ubiquitin cDNAs is their failure to extend to and beyond the initiating ATG codon (5,12-15, pRBL26). This phenomenon has prevented the analysis of the 5' non-coding region of the only otherwise fully characterised gene reported - the human 9-repeat gene (12). Comparison of cDNA sequences reveals that some are structurally very similar. The cDNA insert of pRBL26 and the Xenopus cDNA (14) are structural analogs, containing 2.2 repeats of the coding sequence and initiating at the same nt (184, Figure 3). The barley cDNA (15) contains 2.24 repeats and initiates 9bp upstream. Analysis of the sequence around these positions reveals a 10bp inverted repeat (IR), separated by 11bp (163/172 and 184/193, Figure 3), which is also present in the second and third coding repeats at the corresponding positions. The initiation points of pRBL26 and the Xenopus cDNA (14) are colinear with the IR, while the barley cDNA (15) begins within the 11bp separating the IRs. It is therefore highly likely that the IRs have formed a snap-back loop structure and self-primed during cDNA synthesis. These IRs may also be involved in forming other secondary structures as discussed later.

An Intron Within The 5' Non-Coding Region

The lack of cDNA 5' non-coding regions has prevented previous analysis of gene 5' flanks. However, the isolation of two processed pseudogenes of the human 3-repeat ubiquitin gene EHB8 has provided sequences representative of the mRNA 5' untranslated region. These sequences have allowed the identification of a 715bp intron within the 5' non-coding region of the gene, a feature previously reported absent from ubiquitin genomic sequences. The intron is relatively G-rich and contains short direct repeats, the significance of which is presently unknown.

A search of the EMBL and GenBank nucleic acid sequence databases with

the intron sequence revealed no close homologies, but distant homology to several reported sequences. These were most commonly the inverted terminal repeats of various adenovirus types; for example, type 12 (58). However, these and other distantly homologous sequences are G-rich/C-poor in base composition, and show only random homology when compared to the EHB8 intron by dot-matrices (not shown).

The presence of introns within gene 5' non-coding regions is a common event. However, the location of a gene's only intron in its 5' flank occurs less frequently. One such case is a rat preproinsulin gene (59). Most species possess only one preproinsulin gene, and it appears that the ancestral form of the gene had two introns (60). The rat has two active preproinsulin genes, one of which has arisen by gene duplication and subsequent loss of its coding region intron, leaving a single 5' non-coding intron. In contrast, introns are absent from known ubiquitin gene coding regions (12,16), indicating that the 3-repeat ubiquitin gene EHB8 has not arisen by a similar mechanism. Another group of 5' non-coding single intron genes is the avian feather keratin and associated keratinisation genes (61). A recent study has concluded that DNA sequences within the 5' intron affect the efficiency of the accurate initiation of transcription of chicken feather keratin genes in Xenopus oocytes (61). The significance of the 5' non-coding intron in EHB8 is presently unclear, and the presence or absence of introns in the 5' untranslated regions of other ubiquitin genes has yet to be determined (12,16).

Messenger RNA Structure and Pseudogene Creation

The 3-repeat ubiquitin gene EHB8 contains various sequence elements in its 3' non-coding flank which have been shown or postulated to be involved in 3' end formation: AATAAA (44,45), YGTGTTYY (46), CAYTG (47), and TTCAAA (48). The 5' non-coding region has been tentatively characterised by employing the processed pseudogenes, with a putative CAP site, and TATA and CAT boxes identified. It has been observed that processed pseudogenes often represent full-length dsDNA copies of mature mRNA (62): for example, two human apoferritin H processed pseudogenes recently described by Costanzo et al (63) are full-length reverse transcripts. The two ubiquitin processed pseudogenes described here appear to represent separate reverse transcription events, based on the different positions of the encoded poly(A) tails 6 nt apart (see Figure 4B). It is not known whether this represents normal or erroneous use of alternate polyadenylation sites, or a transcript of a different, but very similar gene. However, the fact that both processed pseudogenes have only one nt difference in the 5' non-coding flank length (Figure 4A) suggests that they may indeed

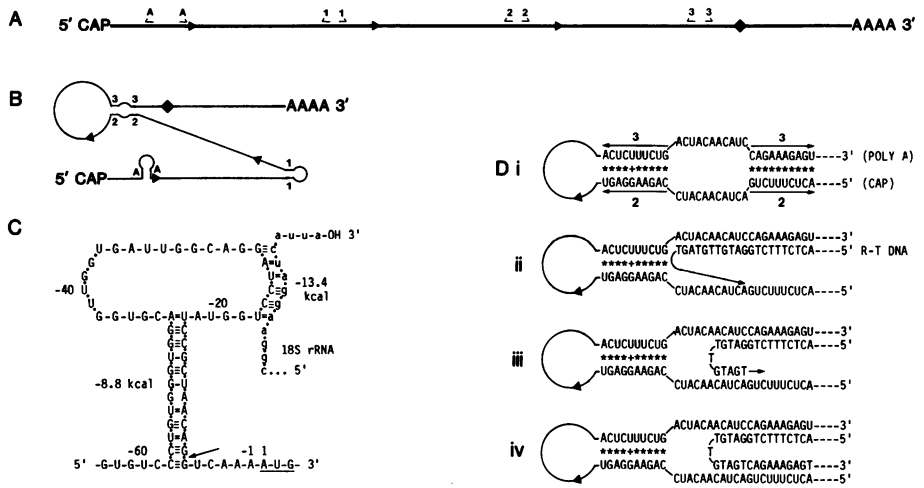

Figure 7. Ubiquitin mRNA secondary structures. (A) Mature transcript of the 3 repeat gene EHB8 in linear form. CAP = 5' capping structure; AAAA = poly(A) tail; ▶ = AUG start codons; and ◆ = the stop codon. Inverted repeats within the coding repeats are shown by small arrows above the mRNA and numbered. A different inverted repeat in the 5' non-coding region is similarly indicated and labelled "A". (B) A putative secondary structure of the mRNA, formed by base-pairing of repeats. Symbols are as above. (C) Possible 5' non-coding secondary structure and association with human 18S rRNA. Nucleotides are numbered from the A of the first start codon (underlined). Negative numbers are used in the 5' non-coding region. The stabilities of the stem-loop structure and association with 18S rRNA were calculated by the method of Tinoco et al (65). The splice point is arrowed. (D, i to iv) Enlargement of the loop-stem-bubble-stem structure from (B) and putative events during reverse transcription. Inverted repeats are indicated as in (A). R-T DNA = reverse transcribed DNA. Base-pairing is shown by *, with a G-U pair shown by +.

be full length reverse transcripts, and that the proposed CAP site and TATA box may be functional in vivo. As noted above, a mature transcript from this CAP site correlates with the observed Ub B mRNA length (12,13).

The 5' non-coding exon does not encode a leader or signal peptide, as there is an in-frame stop codon TAA 4 aa upstream from the start codon (-727 to -725, Figure 3). However, as was noted for the rat preproinsulin and other genes with 5' non-coding introns (59), the 5' exon has the capability to form a stable stem-loop structure (ΔG = -8.8 kcal) 6 nt before the AUG start codon (Figure 7C). In addition, the open loop can base pair with the 3' end of human 18S rRNA (64) forming another stable association (ΔG = -13.4 kcal). This structure may function in ribosome binding and subsequent translation of the mRNA (59), although whether this involves the 18S rRNA or not is presently

unclear. The splice junction lies at the base of the stem (Figure 7C): the splice event would bring this structure adjacent to the start codon, as is seen in other 5' non-coding intron-containing genes (59).

The two genomic clones EHB4 and EHD1 represent processed pseudogenes. EHB4 has a reading frame severely disrupted by a stop codon and a frame-shift insertion and would not be functional. The one-repeat pseudogene EHD1, however, has an uninterrupted reading frame and could potentially code for an ubiquitin-like protein (see Figure 5). However, heterogeneity in animal ubiquitin protein sequences has not been observed to date, and it is unlikely that this processed pseudogene is expressed.

The most striking feature of the pseudogenes is the precise deletion of one (EHB4) or two (EHD1) coding repeats when compared to EHB8. Most significantly, the 5' and 3' non-coding regions have not been affected by these deletions. Also, no "full-length" (ie 3-repeat) processed pseudogenes have yet been observed, suggesting that the deletion of coding repeats occurs as a consequence of the reverse transcription events leading to pseudogene creation. We have developed a model to explain this phenomenon, shown in Figure 7. Panel A shows the linear, mature transcript of EHB8, with the 10bp IRs in each coding repeat shown and numbered. The IR "A" in the 5' non-coding region is also shown: it plays no part in this model. Presumably the IRs within each coding repeat would pair to form stable stem-loop structures: ΔG = -8.4 kcal for IRs 2 and 3, and -11.8 for IR 3. However, it is possible that the IRs in one coding repeat could pair with the IRs in another repeat. Such an event is shown in Panel B, with IRs 2 paired with IRs 3 to form a loop-stem-bubble-stem structure. The loop represents almost one complete coding unit (197 nt) and the structure is quite stable (ΔG = -25.2 kcal). Panel D represents a putative sequence of events leading to the precise deletion of one coding repeat during reverse transcription. The loop-stem-bubble-stem structure from Panel B is shown in detail in (i). Panel (ii) shows a reverse transcription event which has transcribed from the mRNA 3' end through the first stem and into the "bubble". At this stage, if reverse transcription was interrupted, the nascent ssDNA has the capability to base-pair with the other strand of the bubble. This event is shown in Panel (iii). Continued reverse transcription (Panel iv) results in the precise deletion of one complete coding unit: the 228 nt from IR 2 to IR 3. A sequence comparison between the 3-repeat gene EHB8 and the 2-repeat pseudogene EHB4 indicates that EHB4 is most similar to the putative dsDNA product from Panel D: that is, the structure in Panel B is most likely responsible rather than a structure formed between IRs 1 and 2.

The one-repeat pseudogene EHD1 can also be explained by this model, by a pairing of IRs 1 and 3, and subsequent deletion of two coding repeats during reverse transcription. The feasibility of this model is currently under investigation.

ACKNOWLEDGMENTS

We thank Ms M. Coggan for technical assistance with restriction mapping, and Mrs C. Baker for typing the manuscript. R.B. acknowledges the receipt of a Commonwealth Postgraduate Research Award.

REFERENCES

1. Finley,D. and Varshavsky,A. (1985) Trends Biochem. Sci. 10, 343-347.
2. Hershko,A. and Ciechanover,A. (1982) Annu. Rev. Biochem. 51, 335-364.
3. Ciechanover,A., Finley,D. and Varshavsky,A. (1984) Cell 37, 57-66.
4. Bachmair,A., Finley,D. and Varshavsky,A. (1986) Science 234, 179-186.
5. Bond,U. and Schlesinger,M.J. (1985) Mol. Cell. Biol. 5, 949-956.
6. Busch,H. and Goldknopf,I.L. (1981) Mol. Cell. Biochem. 40, 173-187.
7. Levinger,L. and Varshavsky,A. (1982) Cell 28, 375-385.
8. Siegelman,M., Bond,M.W., Gallatin,W.M., St John,T., Smith,H.T., Fried,V. A. and Weissman,I.L. (1986) Science 231, 823-829.
9. St John,T., Gallatin,W.M., Siegelman,M., Smith,H.T., Fried,V.A. and Weissman,I.L. (1986) Science 231, 845-850.
10. Yarden,Y., Escobedo,J.A., Kuang,W-J., Yang-Feng,T.L., Daniel,T.O., Tremble,P.M., Chen,E.Y., Ando,M.E., Harkins,R.N., Francke,U., Fried,V. A., Ullrich,A. and Williams,L.T. (1986) Nature 323, 226-232.
11. Gallatin,M., St John,T.P., Siegelman,M., Reichert,R., Butcher,E.C. and Weissman,I.L. (1986) Cell 44, 673-680.
12. Wiborg,O., Pedersen,M.S., Wind,A., Berglund,L.E., Marcker,K.A. and Vuust,J. (1985) EMBO J. 4, 755-759.
13. Lund,P.K., Moats-Staats,B.M., Simmons,J.G., Hoyt,E. D'Ercole,A.J., Martin,F. and Van Wyk,J.J. (1985) J. Biol. Chem. 260, 7609-7613.
14. Dworkin-Rasti,E., Shrutkowski,A. and Dworkin,M.B. (1984) Cell 39, 321-325.
15. Gausing,K. and Barkardottir,R. (1986) Eur. J. Biochem. 158, 57-62.
16. Ozkaynak,E., Finley,D. and Varshavsky,A. (1984) Nature 312, 663-666.
17. Izquierdo,M., Arribas,C., Galceran,J., Burke,J. and Cabrera,V.M. (1984) Biochim. Biophys. Acta 783, 114-121.
18. Marx,J.L. (1986) Science 231, 796-797.
19. Giannelli,F., Choo,K.H., Rees,D.J.G., Boyd,Y., Rizza,C.R. and Brownlee, G.G. (1983) Nature 303, 181-182.
20. Reid,K.B.M., Bentley,D.R. and Wood,K.J. (1984) Phil. Trans. R. Soc. Lond. 306, 345-354.
21. Stanley,K.K. and Luzio,J.P. (1984) EMBO J. 3, 1429-1431.
22. Casadaban,M.J. and Cohen,S.N. (1980) J. Mol. Biol. 138, 179-207.
23. Remaut,E., Tsao,H. and Fiers,W. (1983) Gene 22, 103-113.
24. Stanley,K.K. (1983) Nucl. Acids Res. 11, 4077-4092.
25. Board,P.G. (1984) Ann. Hum. Genet. 48, 223-228.
26. Maniatis,T., Fritsch,E.F. and Sambrook,J. (1982) "Molecular Cloning: A Laboratory Manual." Cold Spring Harbour Laboratory, New York.
27. Norrander,J., Kempe,T. and Messing,J. (1983) Gene 26, 101-106.
28. Messing,J. and Vieira,J. (1982) Gene 19, 269-276.
29. Messing.J., Crea,R. and Seeburg,P.H. (1981) Nucl. Acids Res. 9, 309-321.

30. Sanger,F., Coulson,A.R., Barrell,B.G., Smith, A.J.H. and Roe,B.A. (1980) J. Mol. Biol. 143, 161-178.
31. Birnboim,H.C. and Doly,J. (1979) Nucl. Acids Res. 7, 1513-1523.
32. Hattori,M. and Sakaki,Y. (1986) Anal. Biochem. 152, 232-238.
33. Messing,J. (1983) in Grossman,L. and Moldave,K. (eds), Methods in Enzymology, Academic Press, New York, Vol. 101, pp.20-78.
34. Lin,H-C., Lei,S-P. and Wilcox,G. (1985) Anal. Biochem. 147, 114-119.
35. Reed,K.C. and Mann,D.A. (1985) Nucl. Acids Res. 13, 7207-7221.
36. Burke,J.F. (1984) Gene 30, 63-68.
37. Frischauf,A.M., Lehrach,H., Poustka,A.M. and Murray,N. (1983) J. Mol. Biol. 170, 827-842.
38. Murray,N.E., Brammar,W.J. and Murray,K. (1977) Mol. Gen. Genet. 150, 53-61.
39. Benton,W.D. and Davis,P.W. (1977) Science 196, 180-181.
40. Ozaki,L.S. and Sharma,S. (1984) In Morel,C.M. (ed), "Genes and Antigens of Parasites: A Laboratory Manual", Fundacao Oswaldo Cruz, Brazil, 2nd edn, pp. 172-173.
41. Rackwitz,H-R., Zehetner,G., Frischauf,A-M. and Lehrach,H. (1984) Gene 30, 195-200.
42. Uher,L. (Bethesda Research Laboratories) (1986) Focus 8:1, 10-11.
43. Grunebaum,L., Cazenave,J-P., Camerino,G., Kloepfer,C., Mandel,J-L., Tolstoshev,P., Jaye,M., De la Salle,H. and Lecocq,J-P. (1984) J. Clin. Invest. 73, 1491-1495.
44. Proudfoot,N.J. and Brownlee,G.G. (1976) Nature 263, 211-214.
45. Fitzgerald,M. and Shenk,T. (1981) Cell 24, 251-260.
46. McLauchlan,J., Gaffney,D., Witton,J.L. and Clements,J.B. (1985) Nucl. Acids Res. 13, 1347-1368.
47. Berget,S.M. (1984) Nature 309, 179-182.
48. Urano,Y., Watanabe,K., Sakai,M. and Tamaoki,T. (1986) J. Biol. Chem. 261, 3244-3251.
49. Breathnach,R. and Chambon,P. (1981) Annu. Rev. Biochem. 50, 349-383.
50. Graves,B.J., Johnson,P.F. and McKnight,S.L. (1986) Cell 44, 565-576.
51. White,J.W. and Saunders,G.F. (1986) Nucl. Acids Res. 14, 4719-4730.
52. Schlesinger,D.H. and Goldstein,G. (1975) Nature 255, 423-424.
53. Schlesinger,D.H., Goldstein,G. and Niall,H.D. (1875) Biochemistry 14, 2214-2218.
54. Watson,D.C., Levy,B.W. and Dixon,G.H. (1978) Nature 276, 196-198.
55. Gavilanes,J.G., de Buitrago,G.G., Perez-Castells,R. and Rodriques,R. (1982) J. Biol. Chem. 257, 10267-10270.
56. Vierstra,R.D., Langan,S.M. and Schaller,G.E. (1986) Biochemistry 25, 3105-3108.
57. Wilkinson,K.D., Cox,M.J., O'Connor,L.B. and Shapira,R. (1986) Biochemistry 25, 4999-5004.
58. Shinagawa,M. and Padmanabham,R. (1980) Proc. Natl. Acad. Sci. USA 77, 3831-3835.
59. Lomedico,P., Rosenthal,N., Efstratiadis,A., Gilbert,W., Kolodner,R. and Tizard,R. (1979) Cell 18, 545-558.
60. Bell,G.I., Pictet,R.L., Rutter,W.J., Cordell,B., Tischer,E. and Goodman, H.M. (1980) Nature 284, 26-32.
61. Koltunow,A.M., Gregg,K. and Rogers,G.E. (1986) Nucl. Acids Res. 14, 6375-6392.
62. Sharp,P.A. (1983) Nature 301, 471-472.
63. Costanzo,F., Colombo,M., Staempfli,S., Santoro,C., Marone,M., Frank,R., Delius,H. and Cortese,R. (1986) Nucl. Acids Res. 14, 721-736.
64. McCallum,F.S. and Maden,B.E.H. (1985) Biochem. J. 232, 725-733.
65. Tinoco,I., Borer,P., Dengler,B., Levine,M., Uhlenbeck,O., Crothers,D. and Gralla,J. (1973) Nature New Biol. 246, 40-41.